Machine learning in Communications

XILINX

Agenda

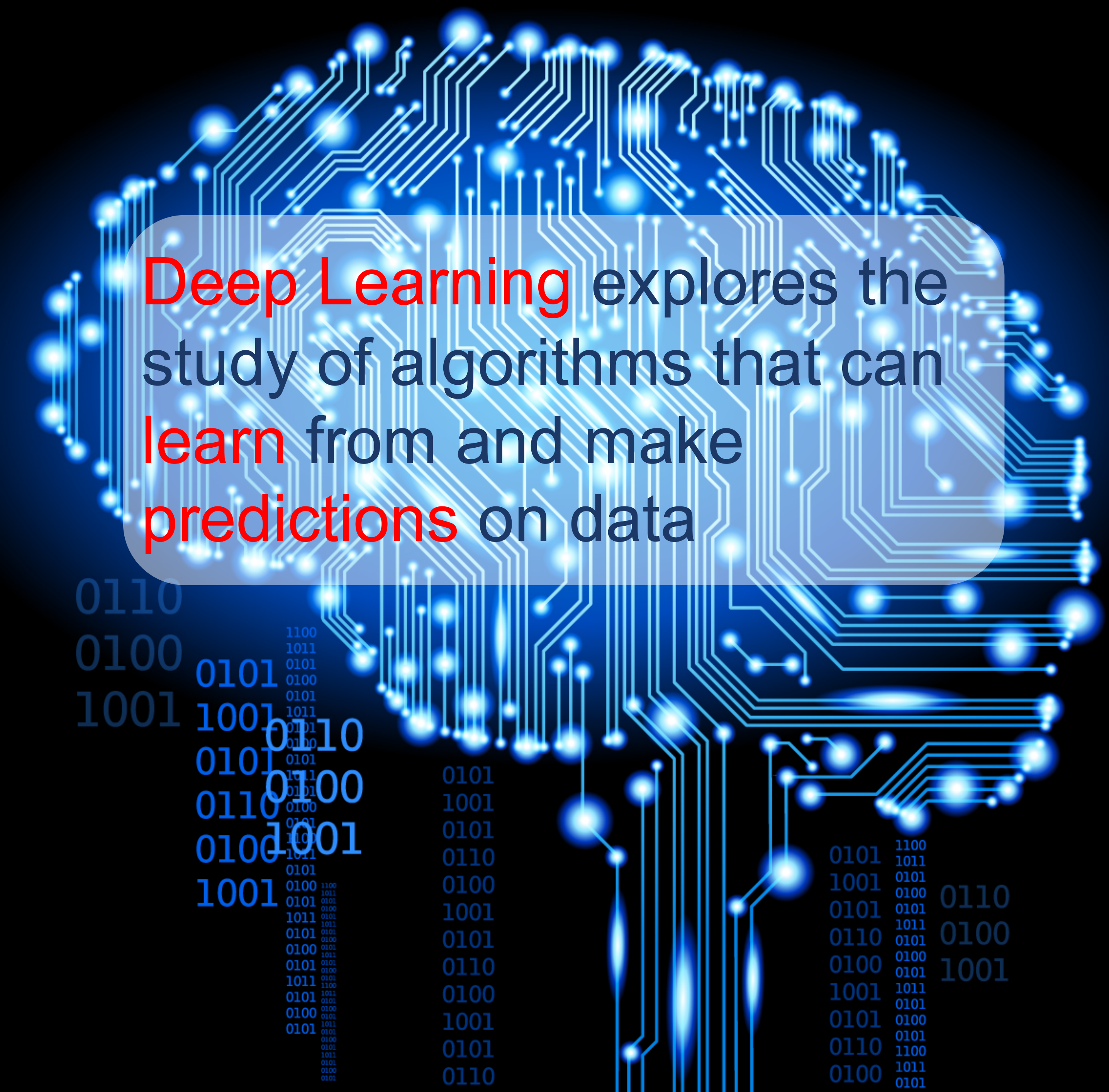
> Introduction

> Machine learning in Communications

> Machine learning use cases

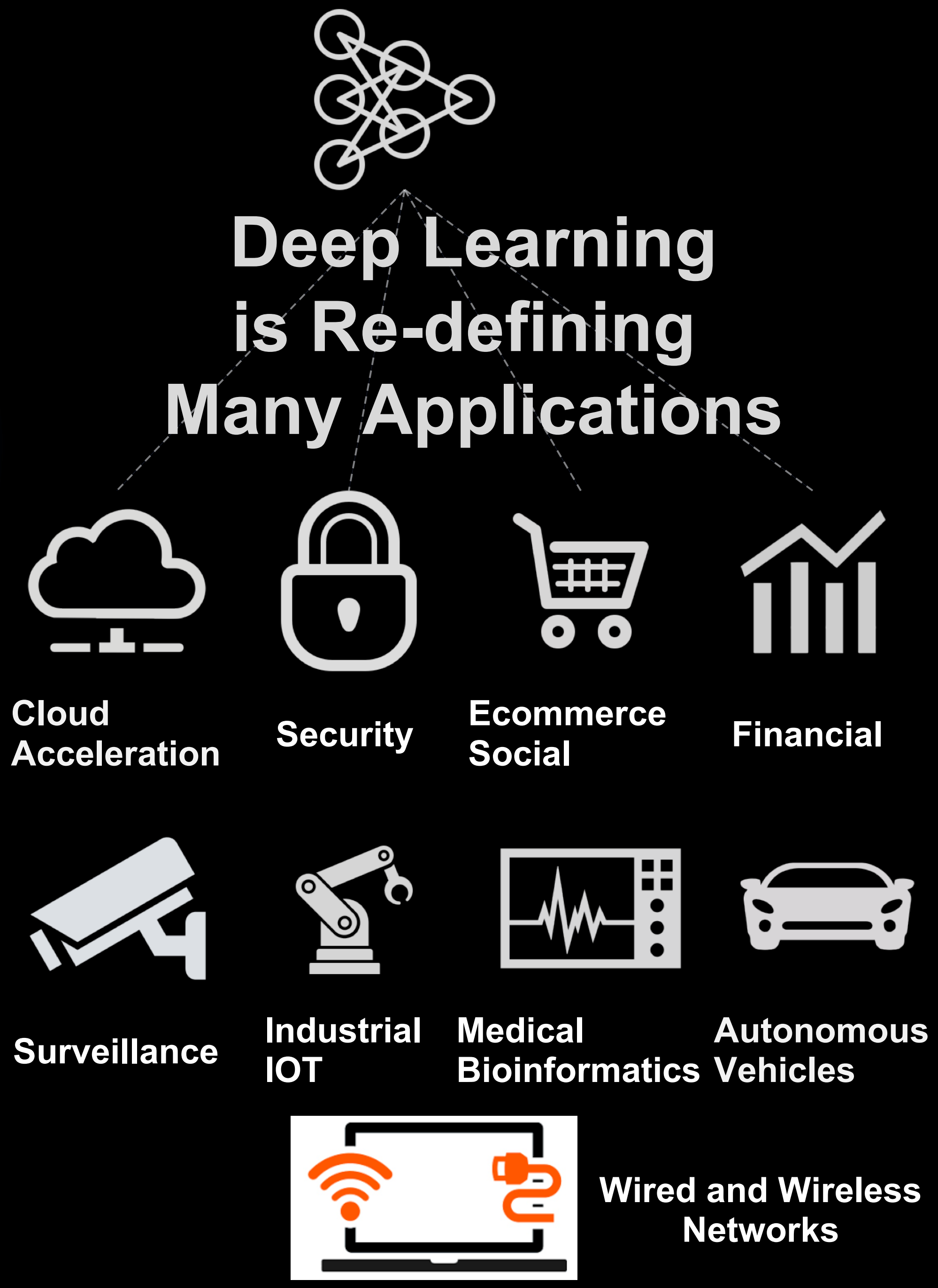
> ML Algorithm for Communication Networks

> Summary



Deep Learning explores the study of algorithms that can learn from and make predictions on data

Deep Learning is Re-defining Many Applications



Market Drivers

Network Traffic doubling every 15-18 months

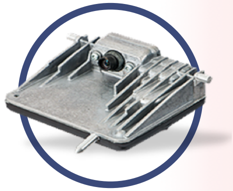
Market Drivers

Network complexity has increase with updated infrastructure

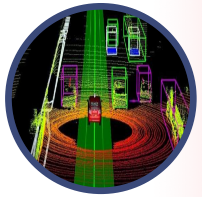
ML/AI will be key components for monetization in Next Gen Networks

Booming "Edge + Cloud" AI Demands

Autonomous driving as "well defined" use case



Front Camera



LiDAR

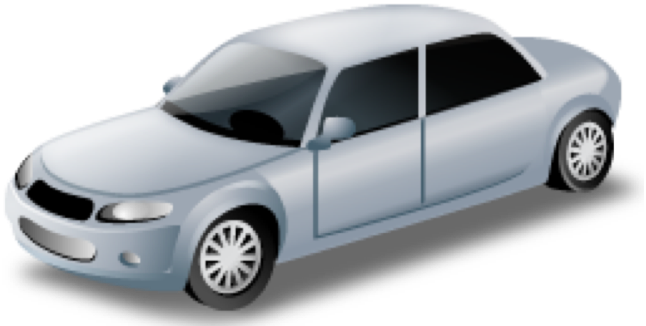
Surround View
Camera

Driver Monitoring

Smart Sensors

> Shorter latency
> Lower power consumption
> Lower cost

Scalable ADAS & AD Platforms From Sensor to Central ECU

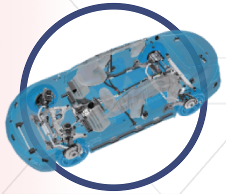


Networking – Which use case ?

Central AD ECU

> Supercomputer
> High throughput
> Security & privacy

Compute Acceleration



Data Aggregation
& Pre-processing

Sensor Fusion

Large Scale Simluation

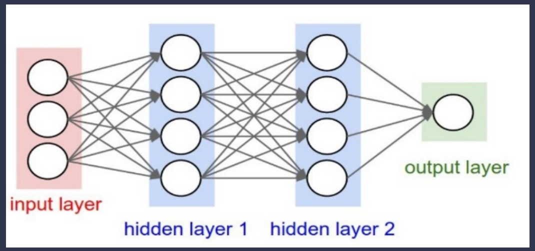
ML for networking - Why

- Millions/Billions of independent flows (Video, Social, Web)
- Complex networks with higher aggregation
- Manual policy assignment can not work
- Reactive control/action is slow
  - Predictive adjustment is needed for efficient networks

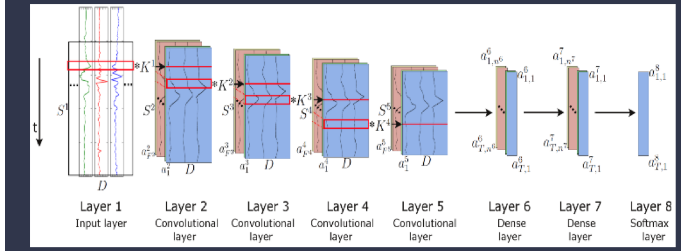
XILINX

Networking is diversified – Many ML use cases

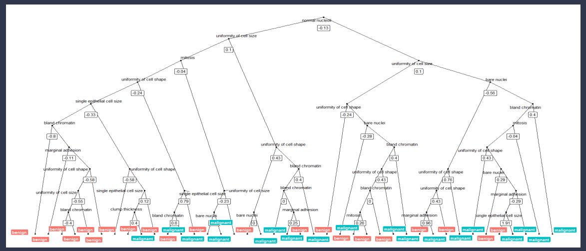
What



MLP

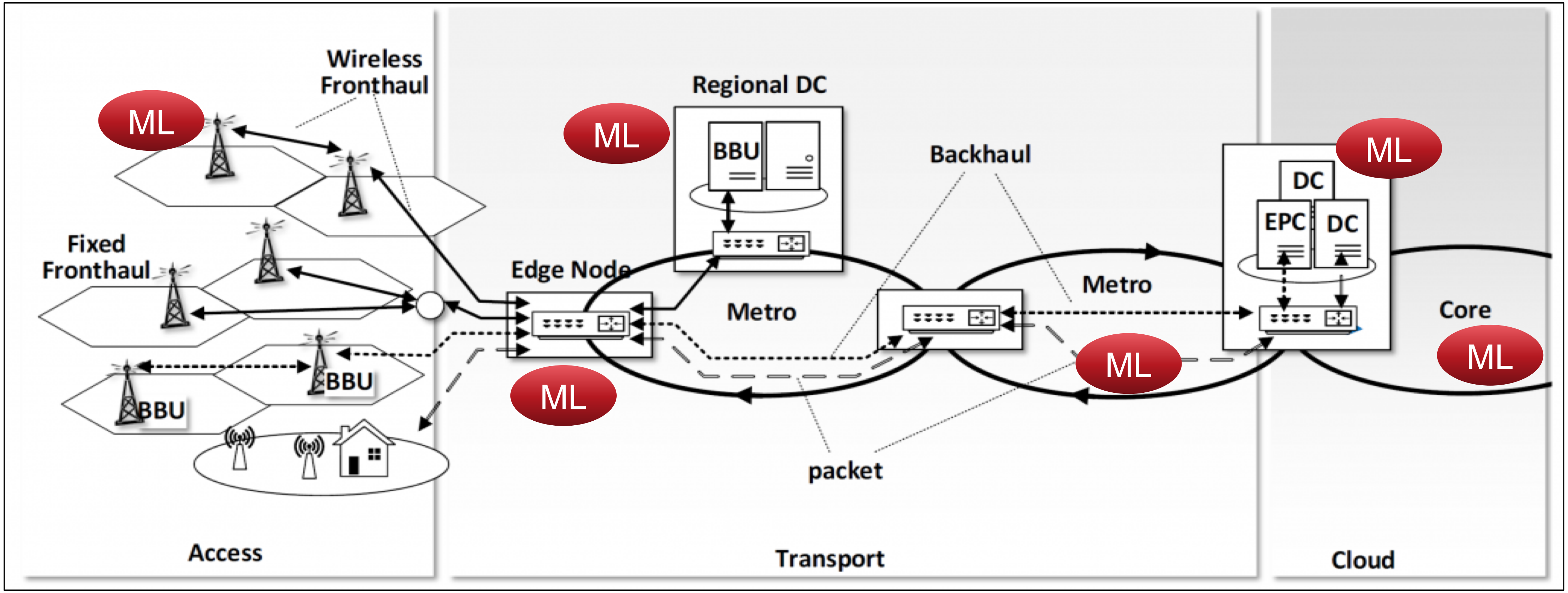


CNN



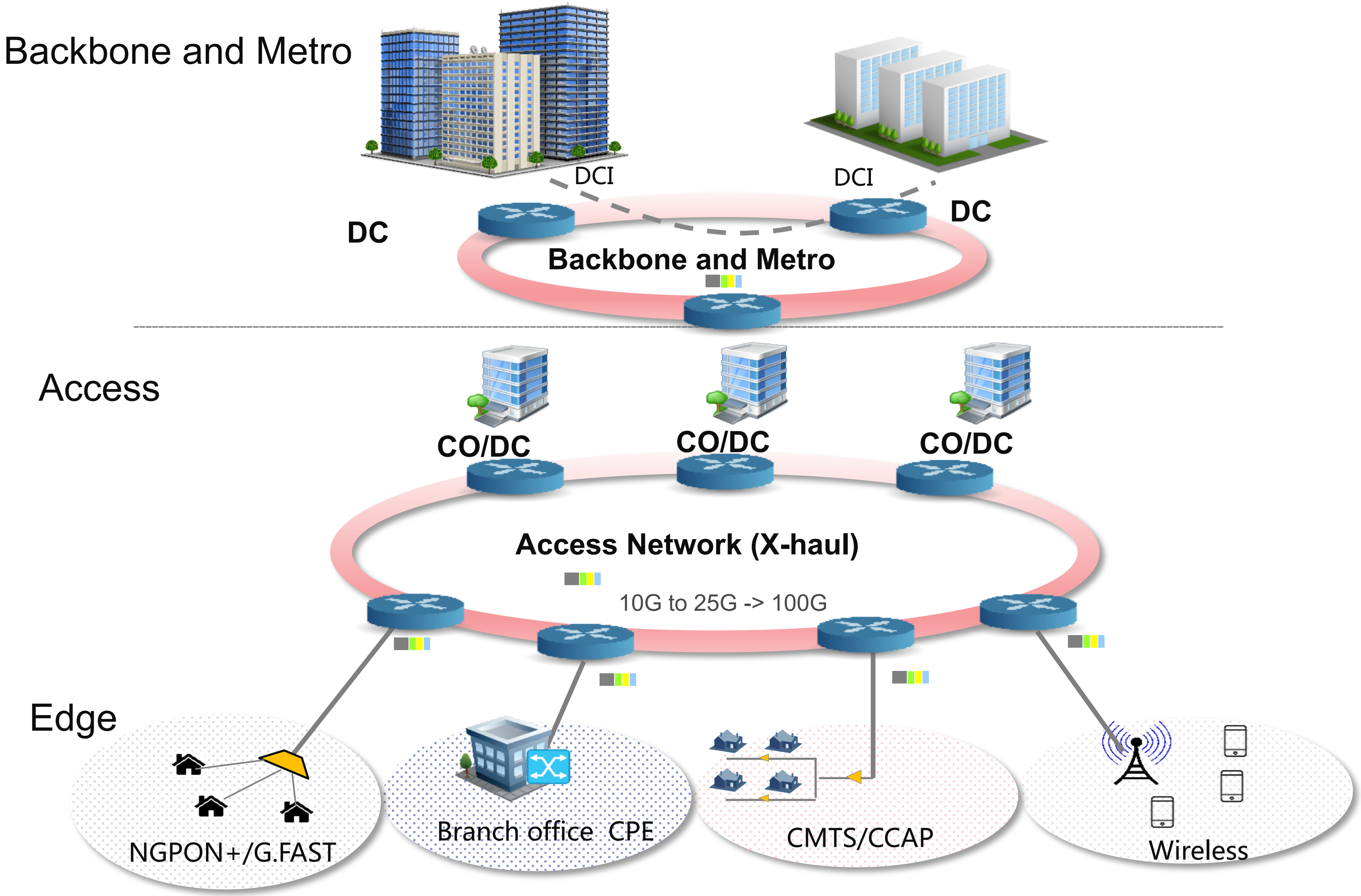
Decision Trees

Where



XILINX.

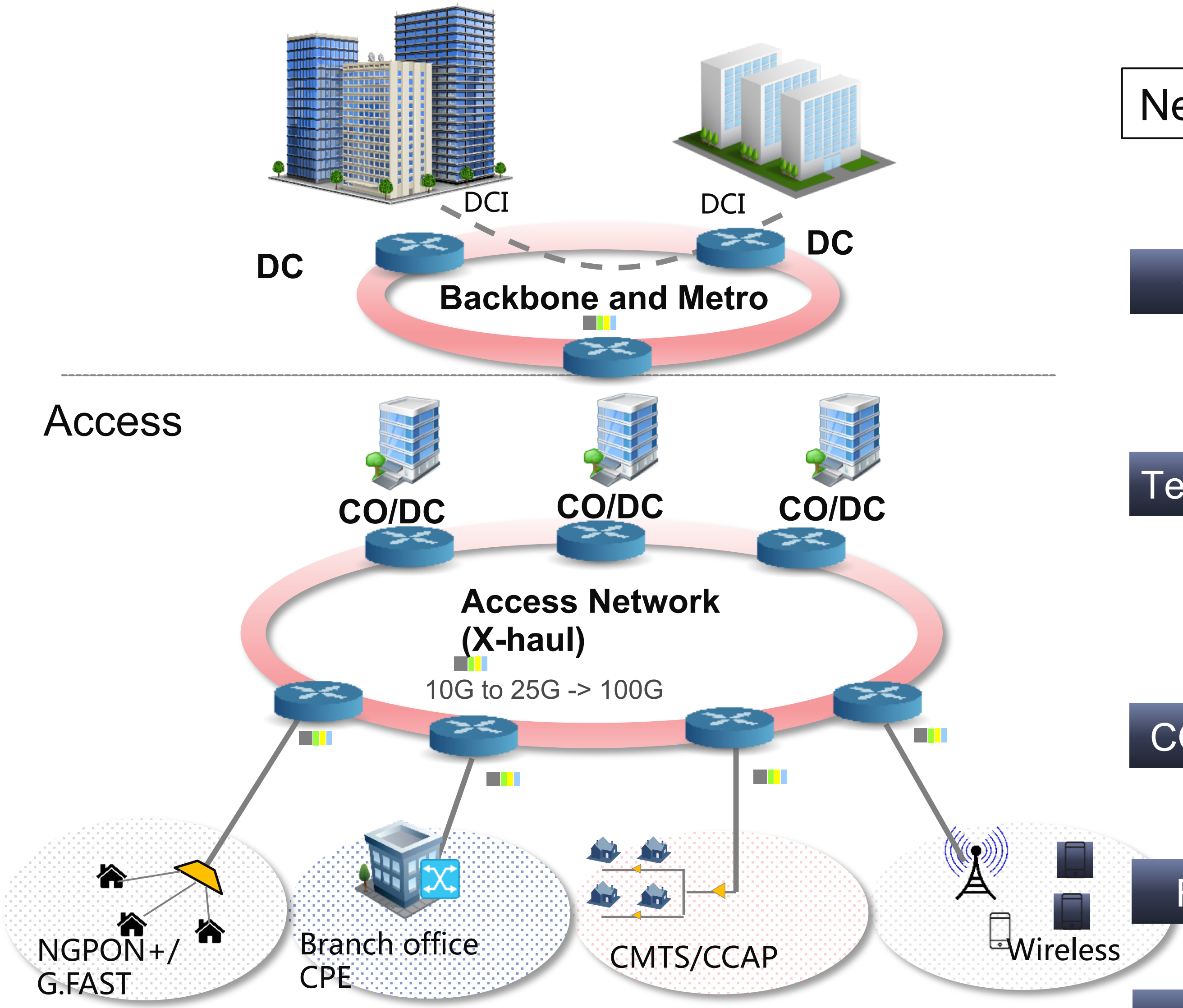
AI/ ML in Networks



XILINX

AI/ ML in Networks

Backbone and Metro

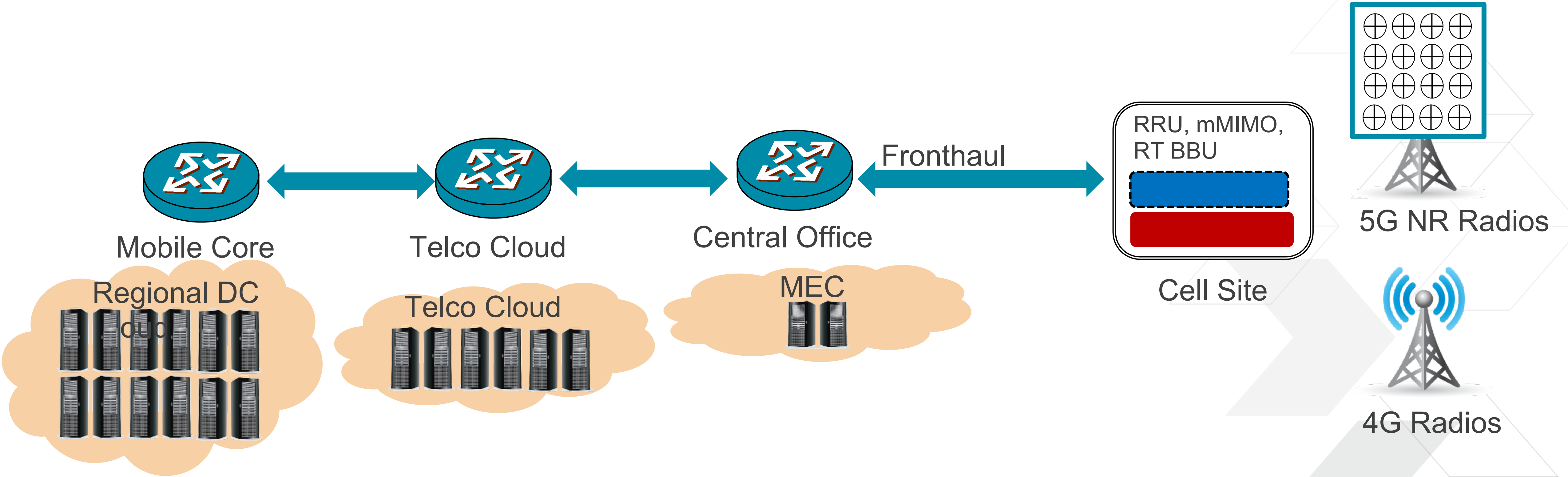


Machine learning Model

Network Segment	Parameters for quality/efficiency
OTN	Optical parameters
Telco DC	QoS ,Traffic policies,Security
CO/Edge Cloud	Telemetry and Security Policies
Radio Units	Resource allocation, Beam steering
IoT devices/Gateway	Security, Interconnectivity, OPs

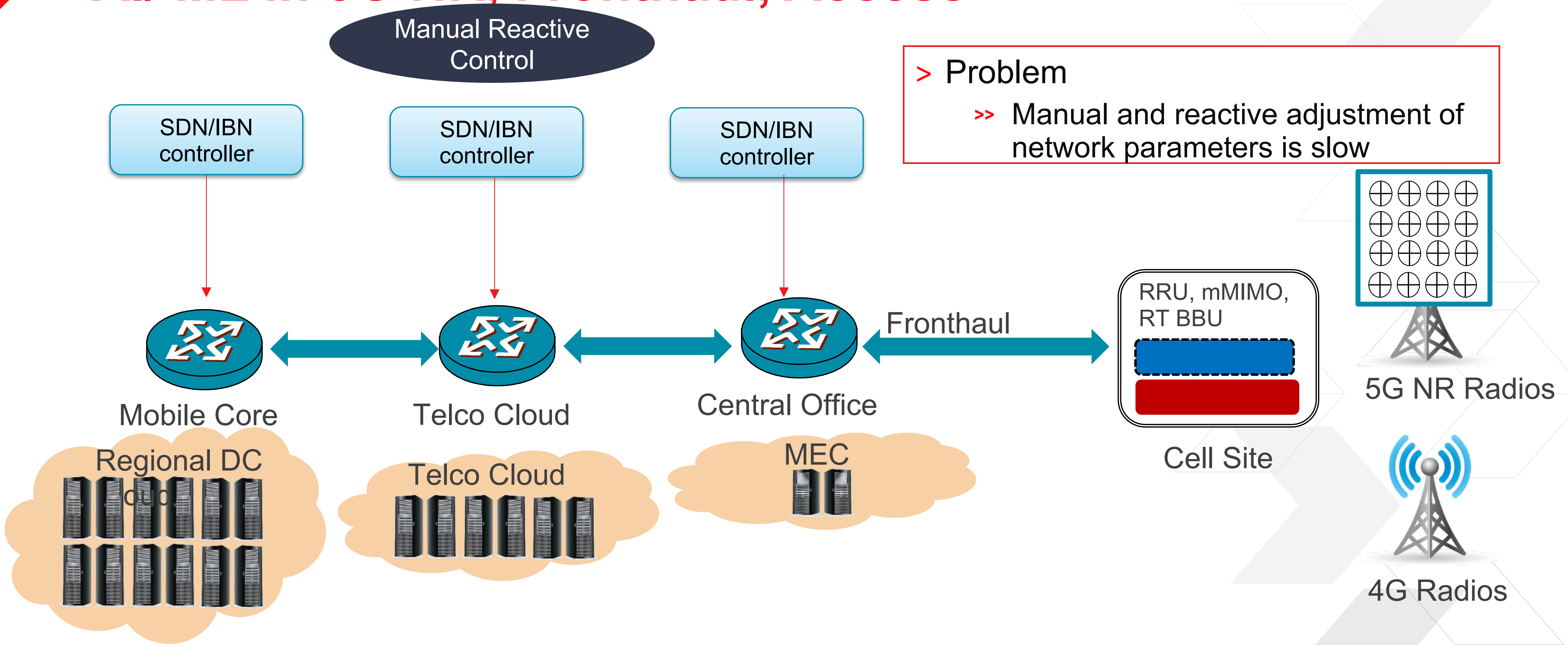
XILINX

AI/ ML in 5G NR, Fronthaul, Access



AI/ ML in 5G NR, Fronthaul, Access

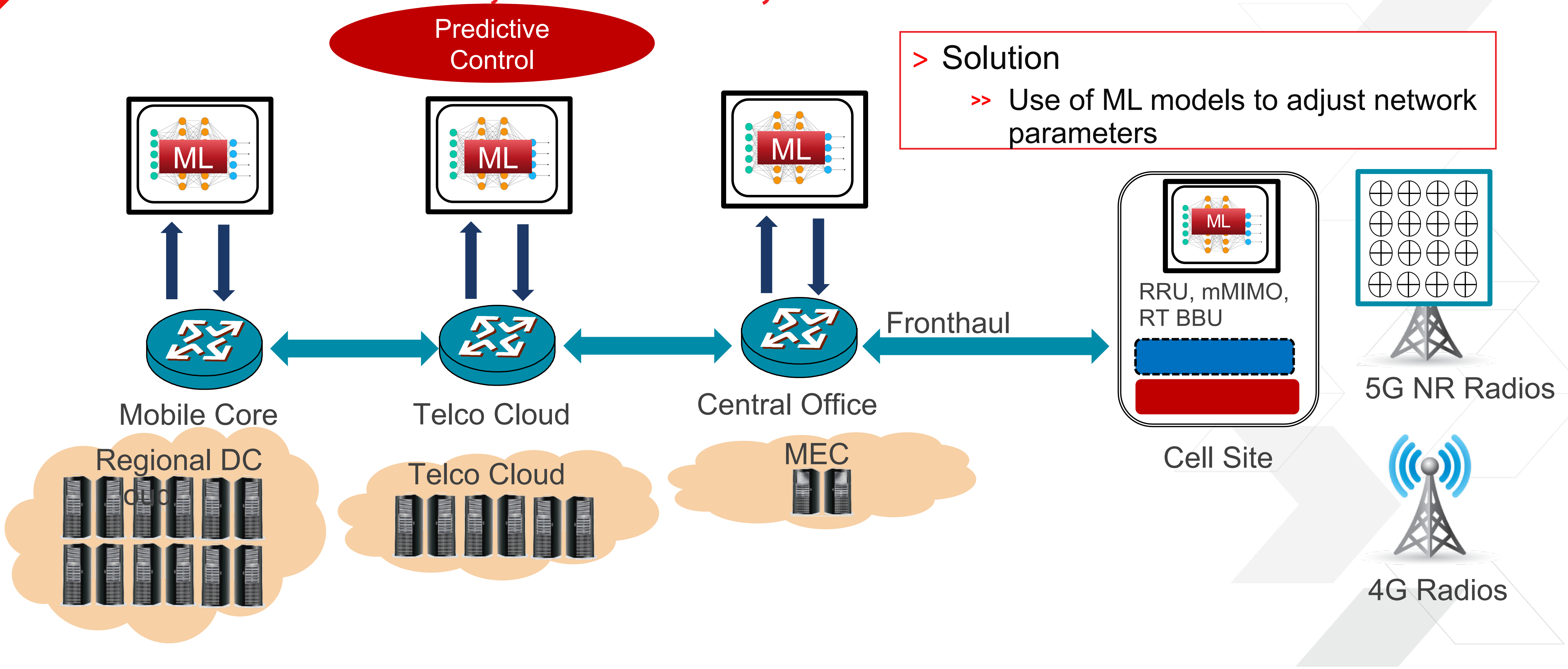
> Problem
  >> Manual and reactive adjustment of network parameters is slow



XILINX.

AI/ ML in 5G NR, Fronthaul, Access

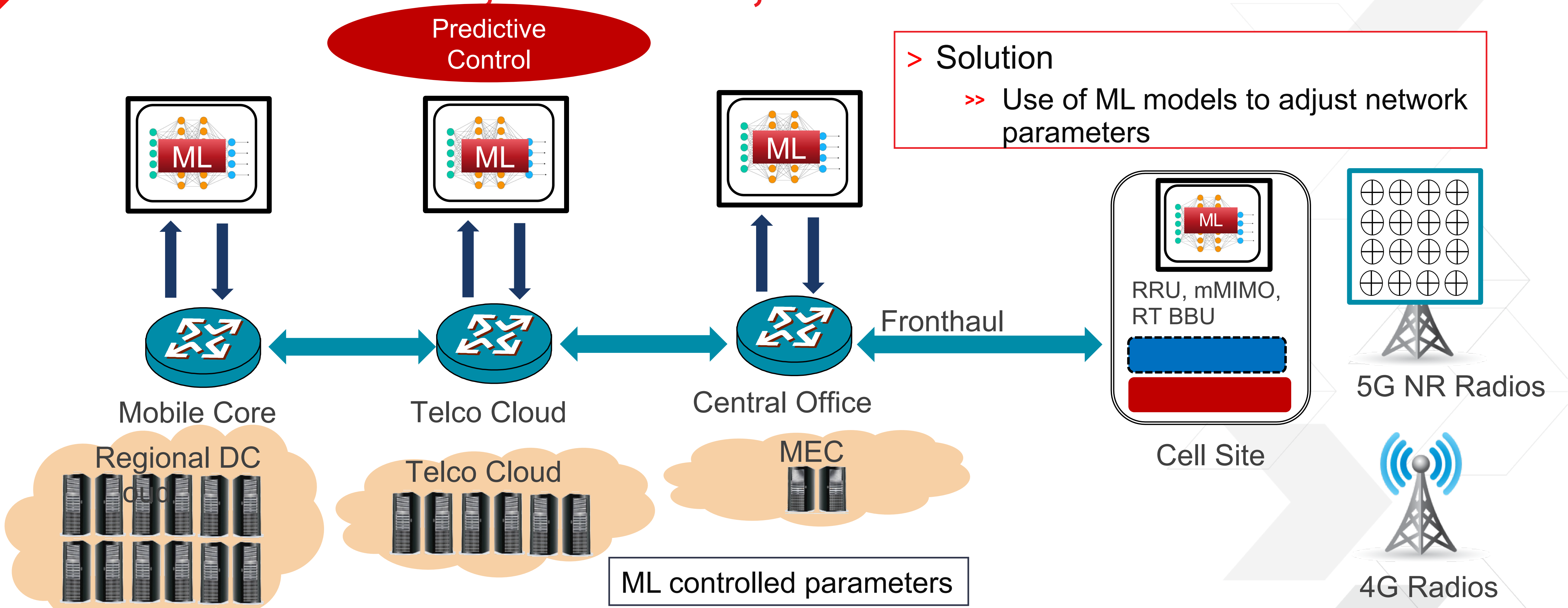
> Solution
  - Use of ML models to adjust network parameters



XILINX

AI/ ML in 5G NR, Fronthaul, Access

> Solution
>> Use of ML models to adjust network parameters



- Network Telemetry
- Network Management and optimization
- Network Security

- QoS
- Content delivery
- Network Flow analytics

- Power, Resource
- Route selection
- Traffic steering

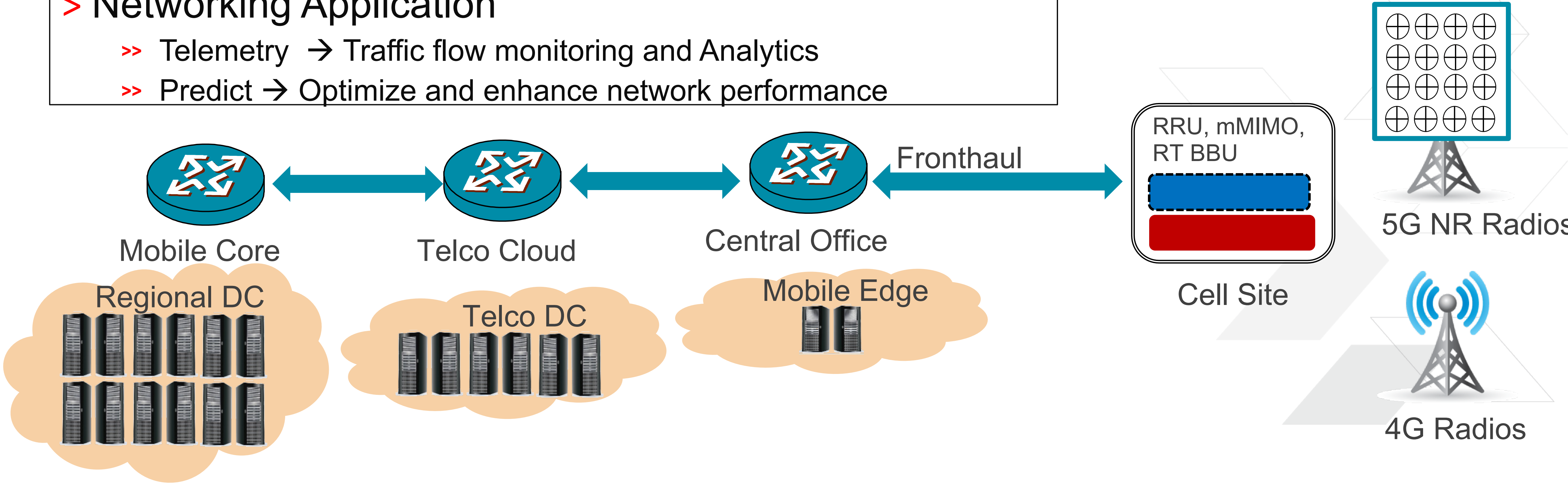
- M-MIMO
- Beam Steering

- DPD
- Radio Energy efficiency

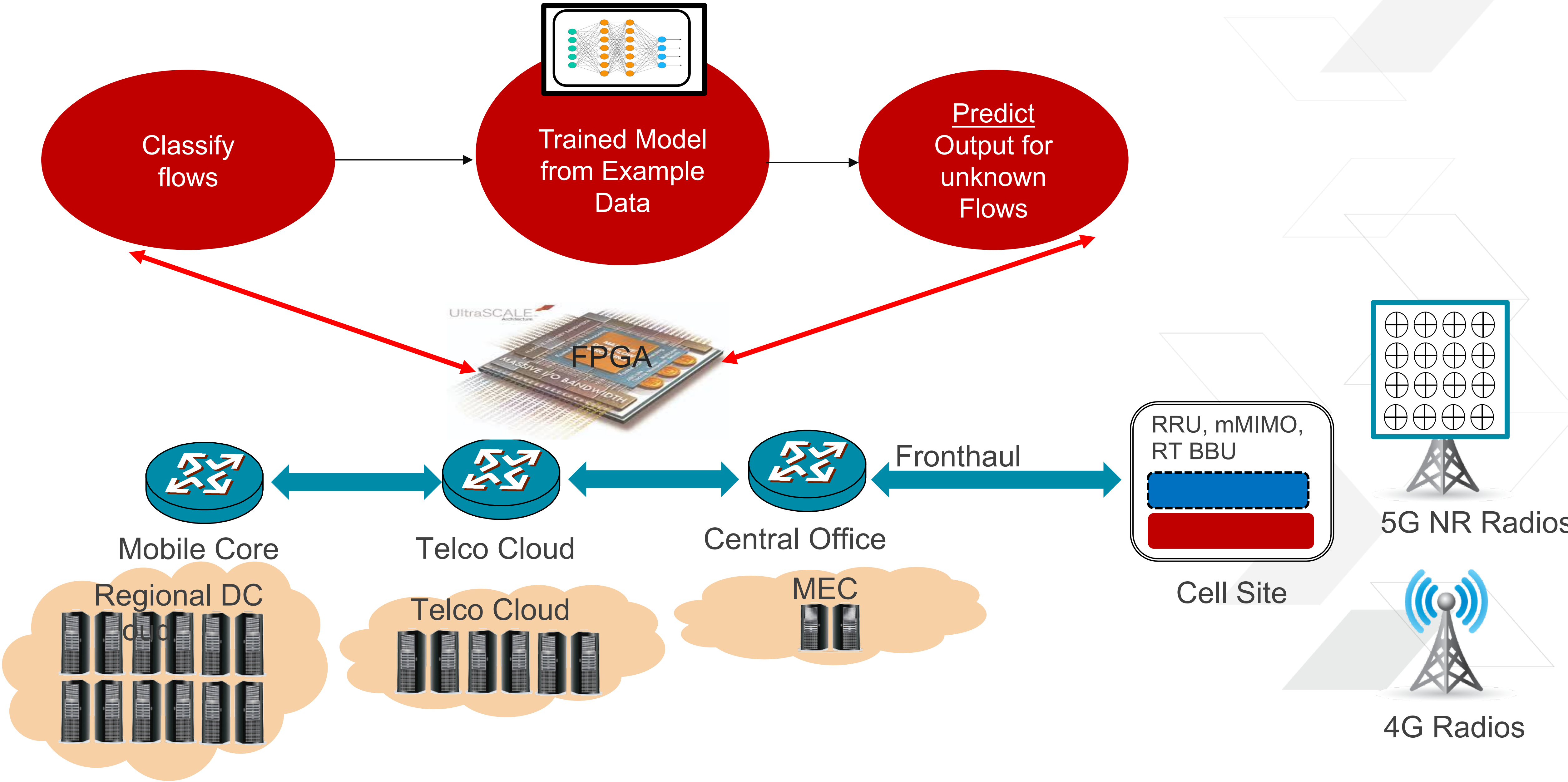
XILINX.

Why ML in Networks

- > Example of networking data (Why ML?)
  - Millions of different flows (Video, Social, Security)
  - Too Complex for manual optimization
  - Traditional analytics (Netflow, sFlow, CLIs) can not handle complexity
- > Networking Application
  - Telemetry → Traffic flow monitoring and Analytics
  - Predict → Optimize and enhance network performance

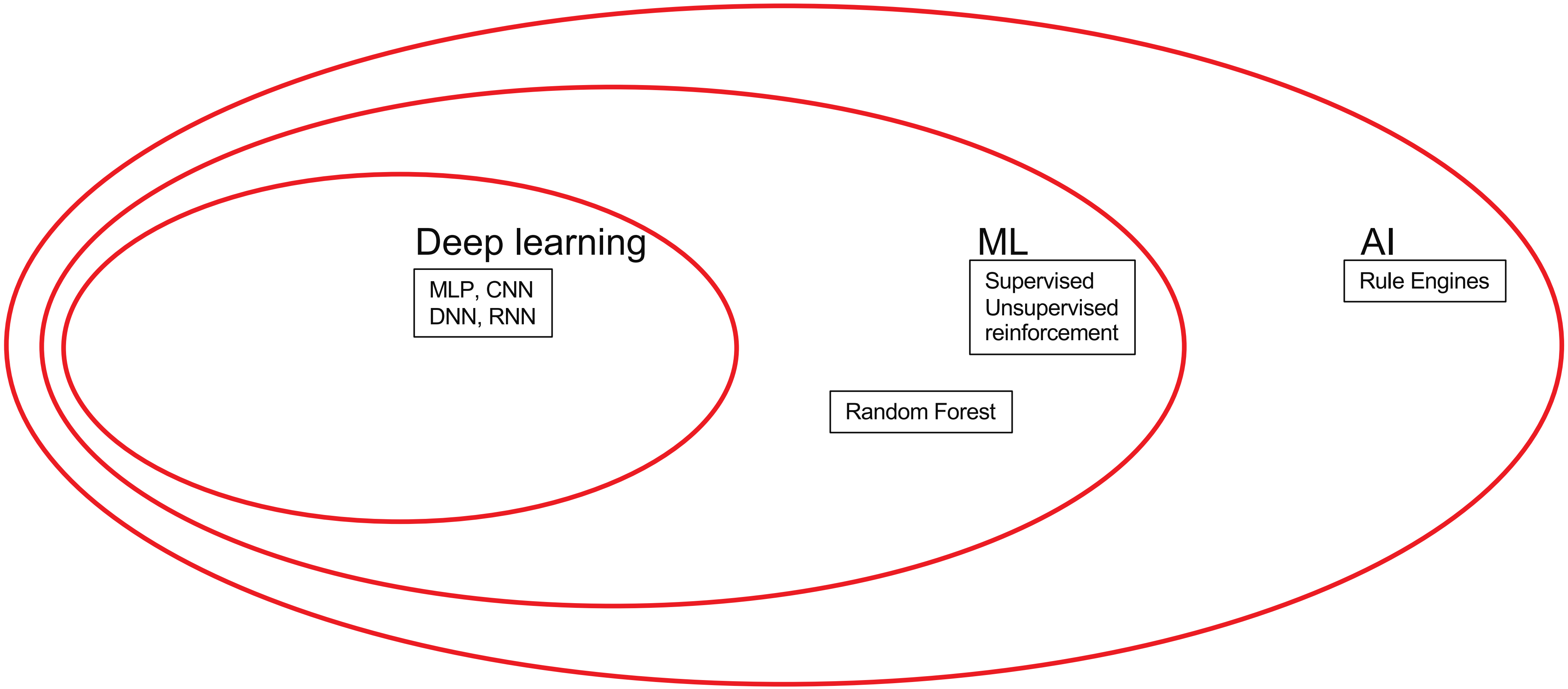


Deploying ML in FPGAs for Telcos

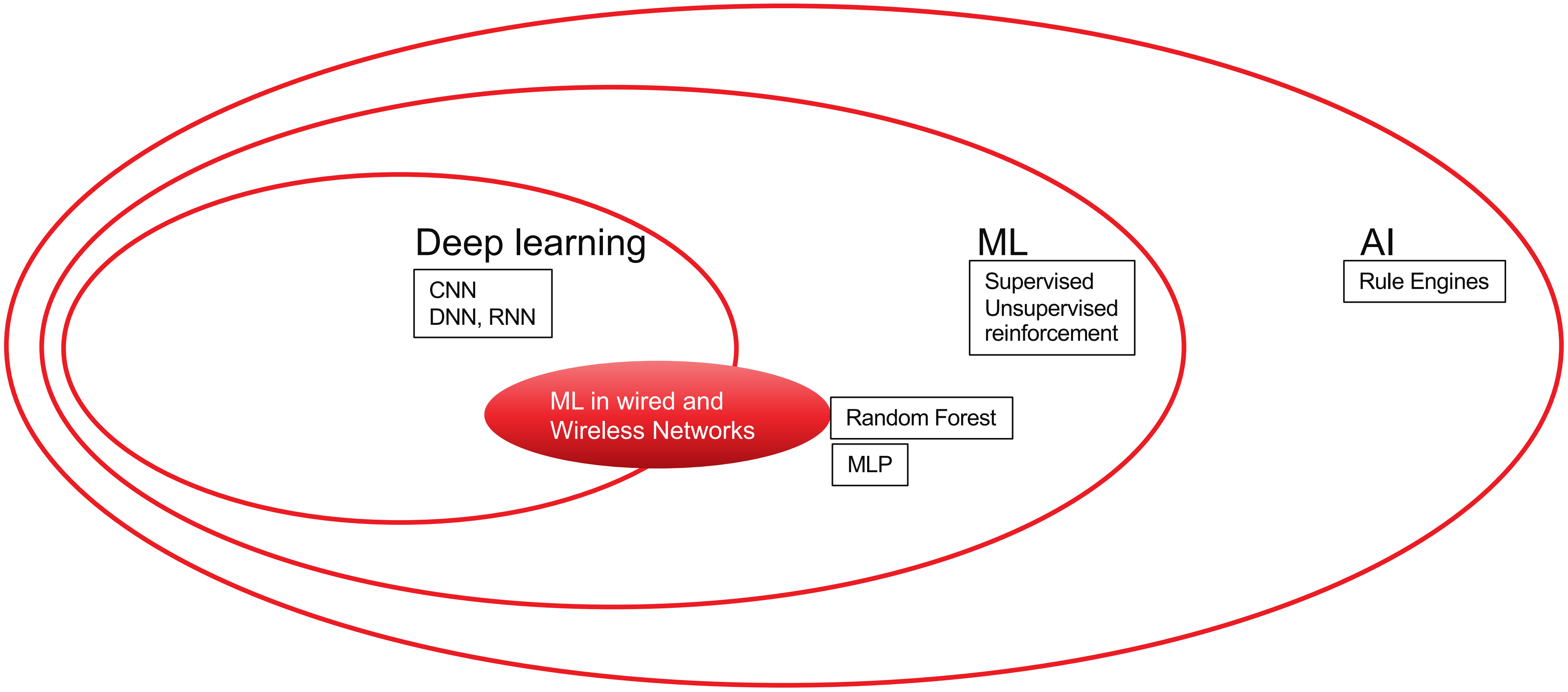


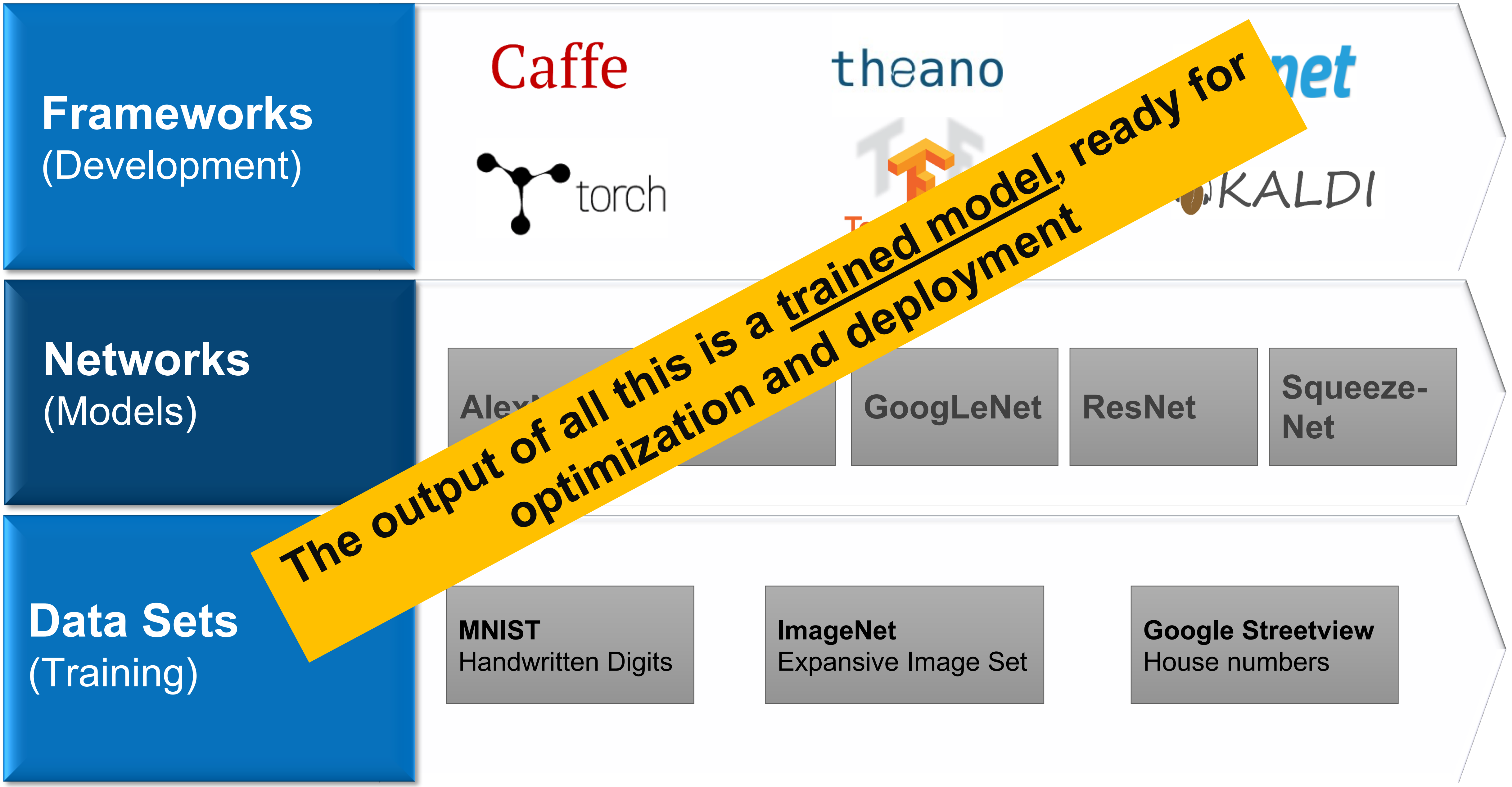
XILINX.

Scope area for ML in networks



Scope area for ML in networks

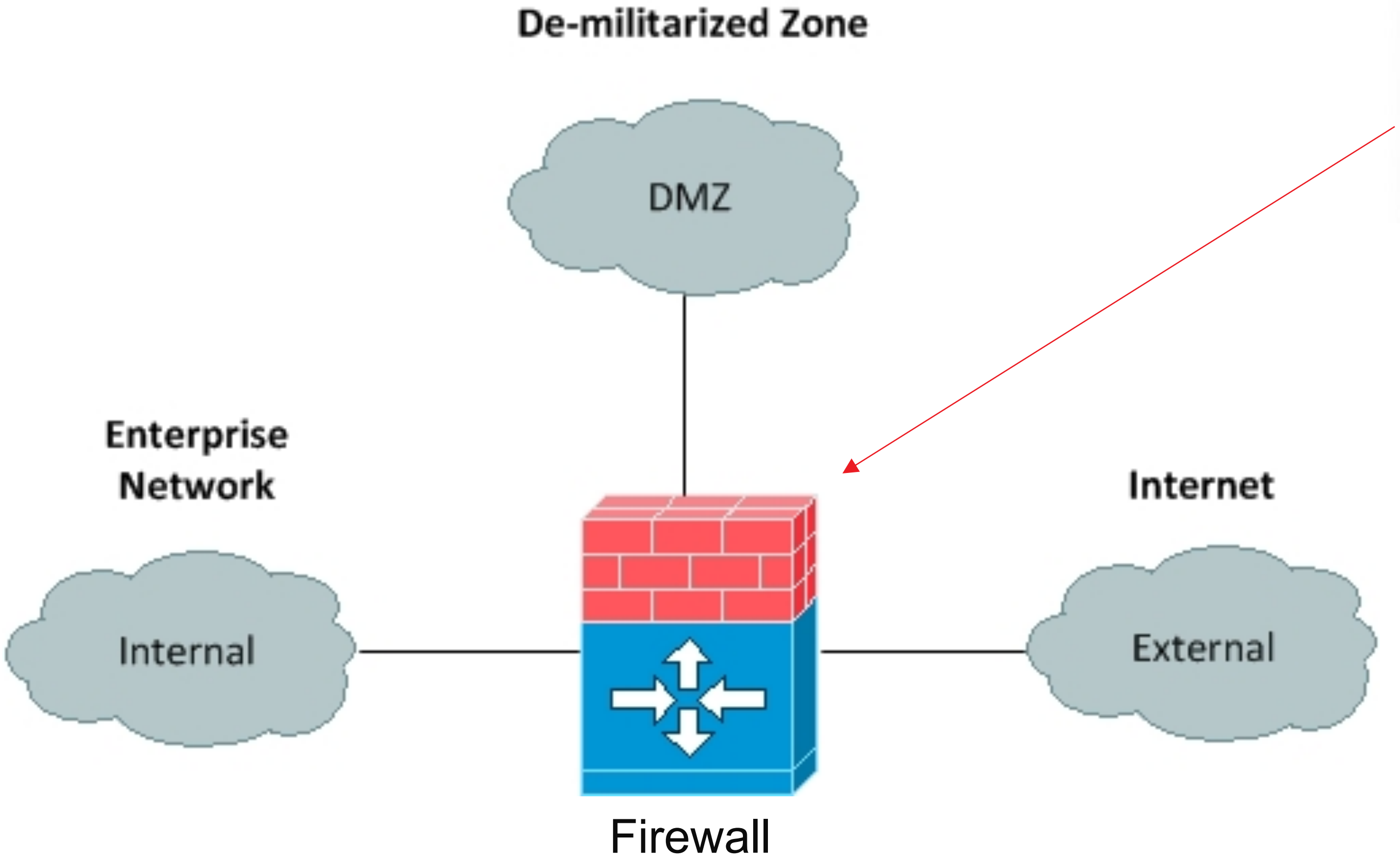




Frameworks
(Development)
Caffe
theano
net
torch
KALDI
Networks
(Models)
GoogLeNet
ResNet
Squeeze-
Net
Data Sets
(Training)
MNIST
Handwritten Digits
ImageNet
Expansive Image Set
Google Streetview
House numbers
The output of all this is a trained model, ready for optimization and deployment

XILINX

ML for Network Security



Firewall Appliance

> Inspect and take action on traffic from/to Enterprise network.

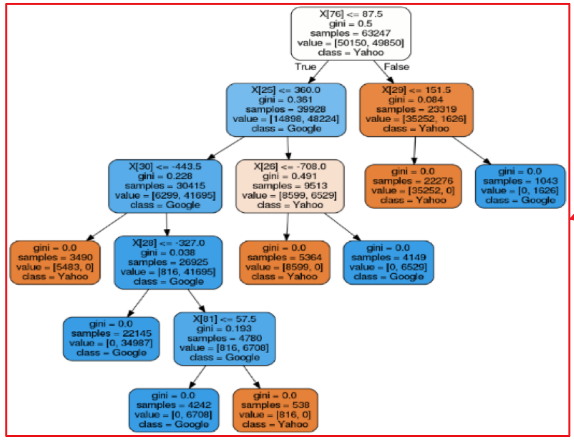
> Built with
  - NPUs, FPGAs, CPU, Custom ASIC, TCAMs

> Includes
  - MACSec, IPSec, SSL, RegEx, Application level security

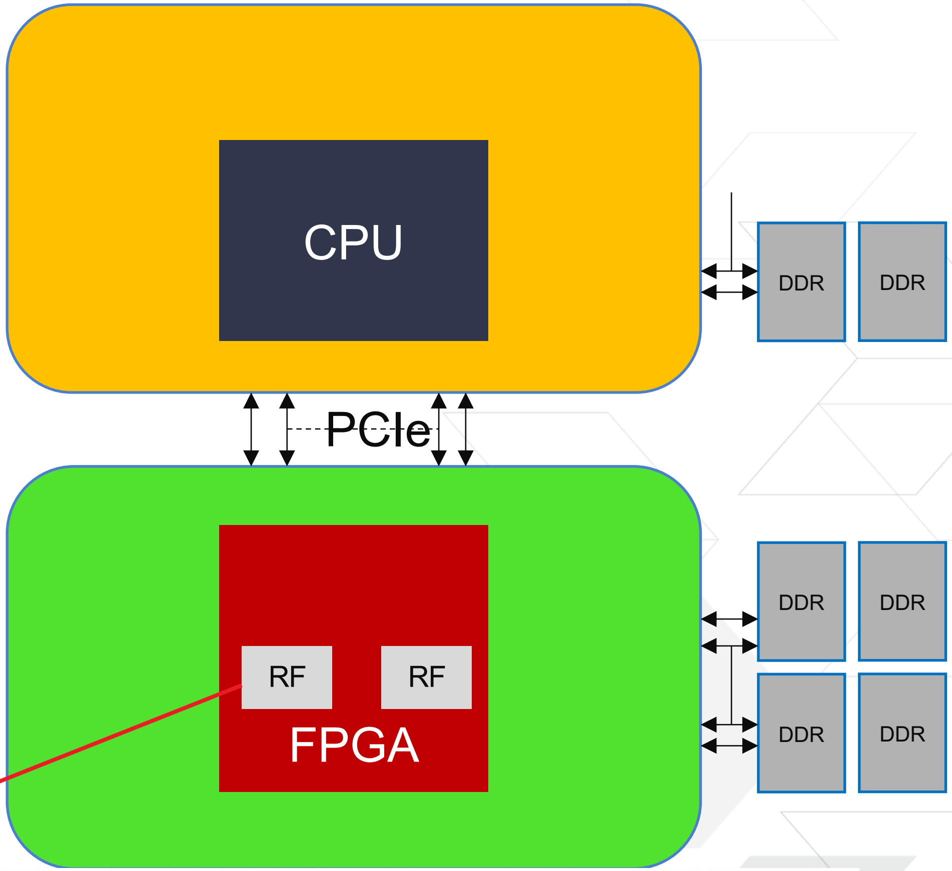
> Why ML
  - Policies and threats always changing

Random Forest for Malware detection

> Why Random Forest for Network security
  - Low resource utilization
  - Easier to train
  - Multiple kernels can be cascade to achieve higher performance
  - Low latency
  - Can utilize SRAM/DRAM



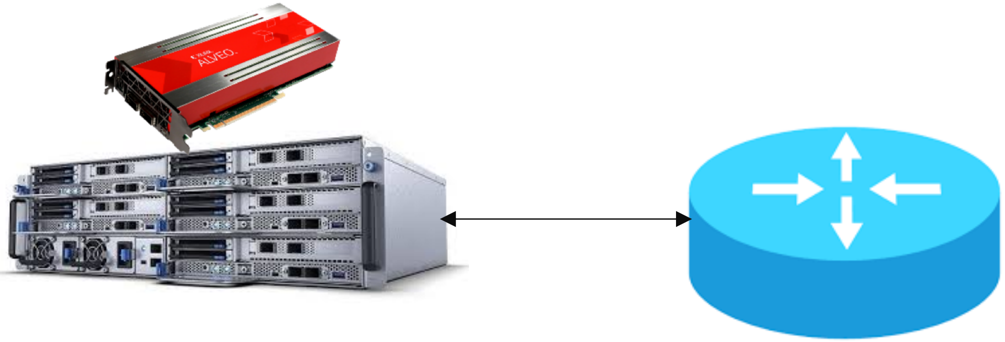
Random Forest ML Model

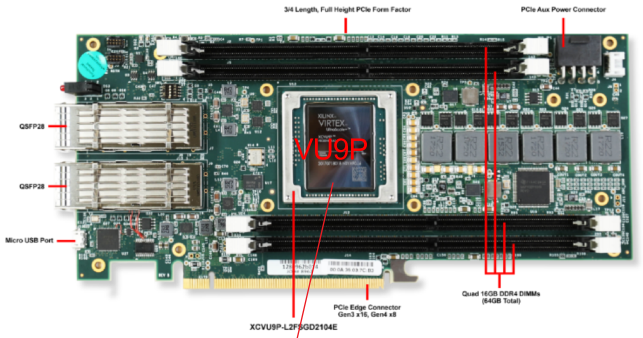


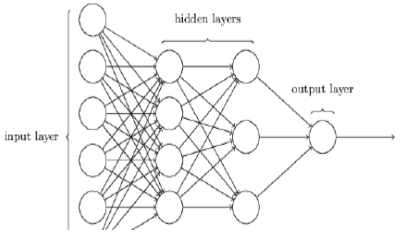
XILINX.

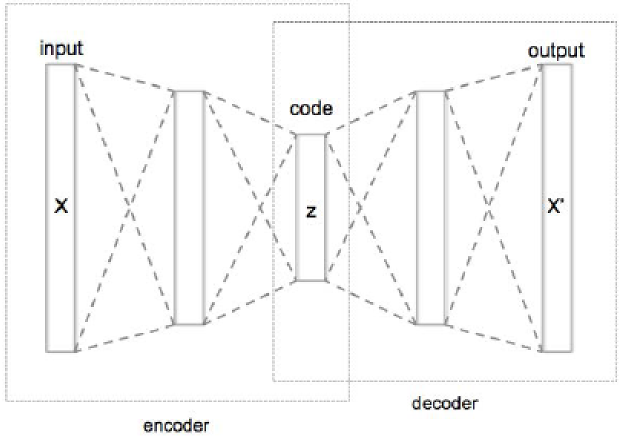
ML for QoE in Access Networks

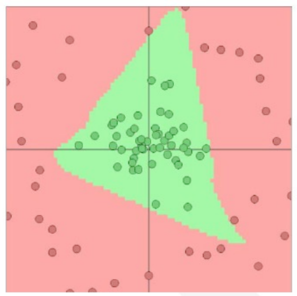
- Telco edge cloud application.
- Detect the traffic flows causing congestion
- ML Algorithm
  - Auto Encoder (Similar to MLP)
- Written using C and running on programmable Accelerator



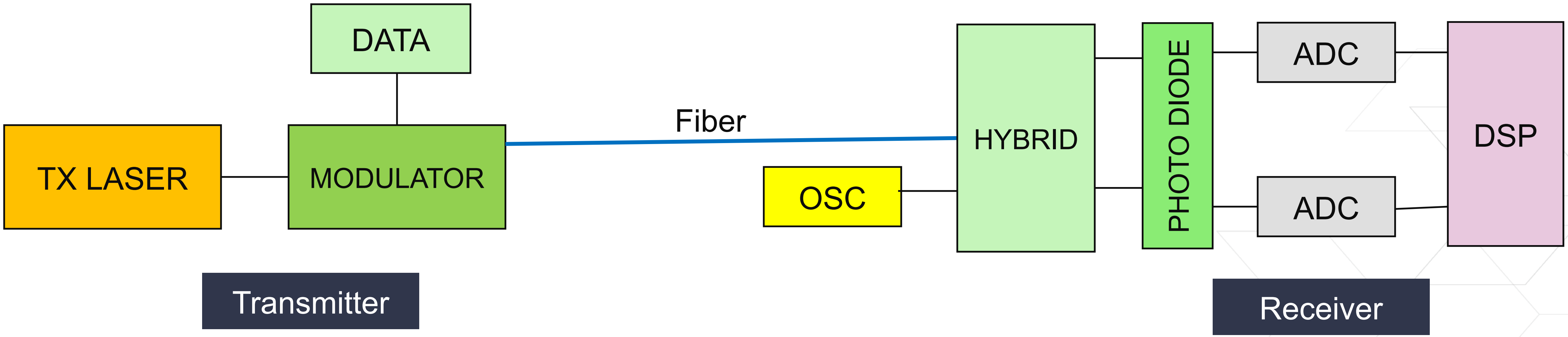






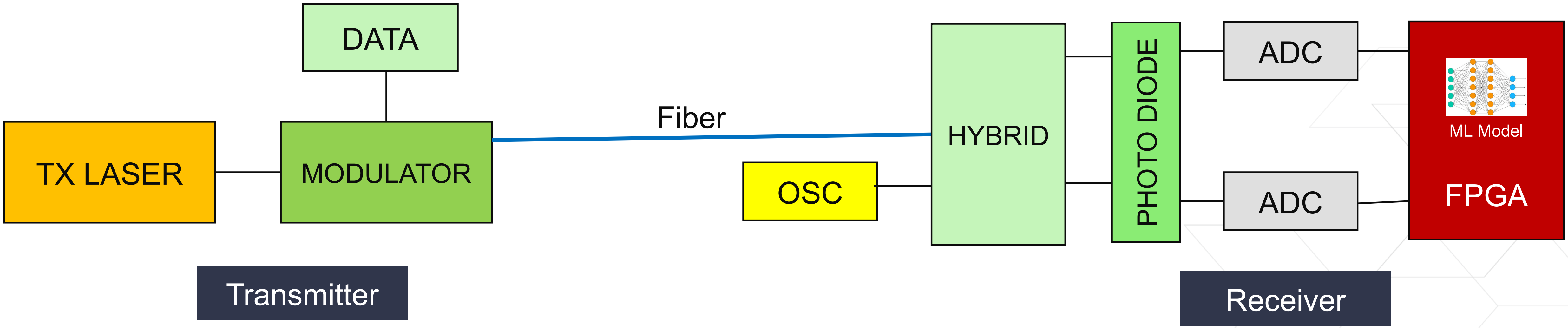


ML for Optical Transmission



> **Problems in Optical Transmission**
>> Chromatic dispersion
>> Polarization mode dispersion
>> Laser phase noise
>> Fiber non linearity

ML for Optical Transmission



> **Problems in Optical Transmission**
- Chromatic dispersion
- Polarization mode dispersion
- Laser phase noise
- Fiber non linearity

> **Solution using ML**
- Direct detection using ML Algorithms

> **Advantage**
- Use of inexpensive fiber and transceiver for optical transmission

Types of ML for OTN

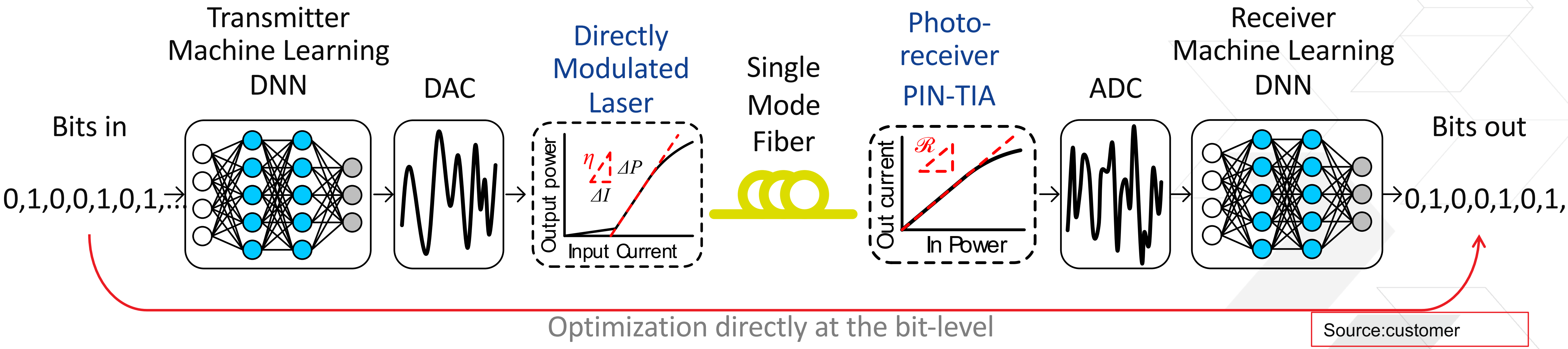
> Supervised learning
  >> Model is trained and deployed based on path, wavelength, modulation and corresponding BER. Trained model is used to create new paths

> Unsupervised learning
  >> Model identifies anomalies in the data
    – Wavelength, path, BER and modulation

> Reinforcement learning
  >> Model learns through the feedback/effect of modifying the traffic parameters (power, modulation etc.)
  >> Delayed reward with trial-and-error

ML in Optical Transmitter/Receiver

- Goal → Best transmission quality from Tx to Rx
  - Adjust transmitter and receiver parameters to achieve lowest BER
  - Proactive adjustment of parameters instead of coherent optics

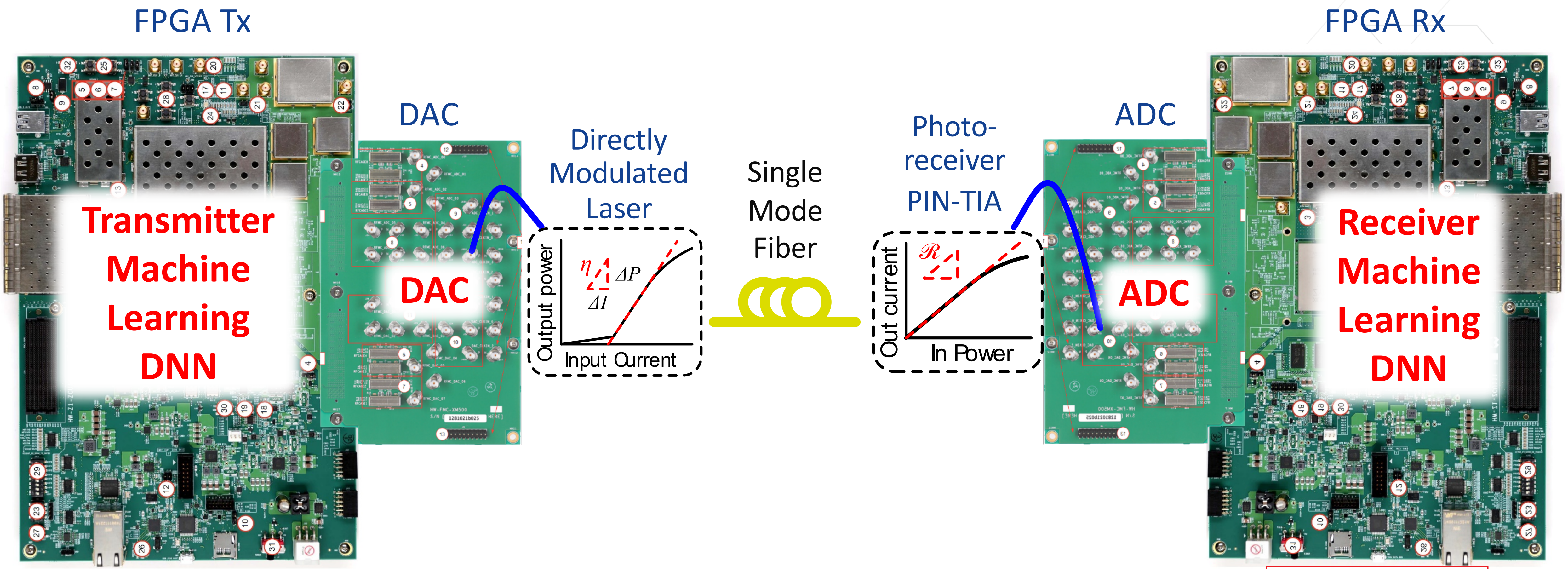


Source:customer

Advantage of Machine Learning → Optical data is available for iterative and quick learning

ML in OTN - Implementation

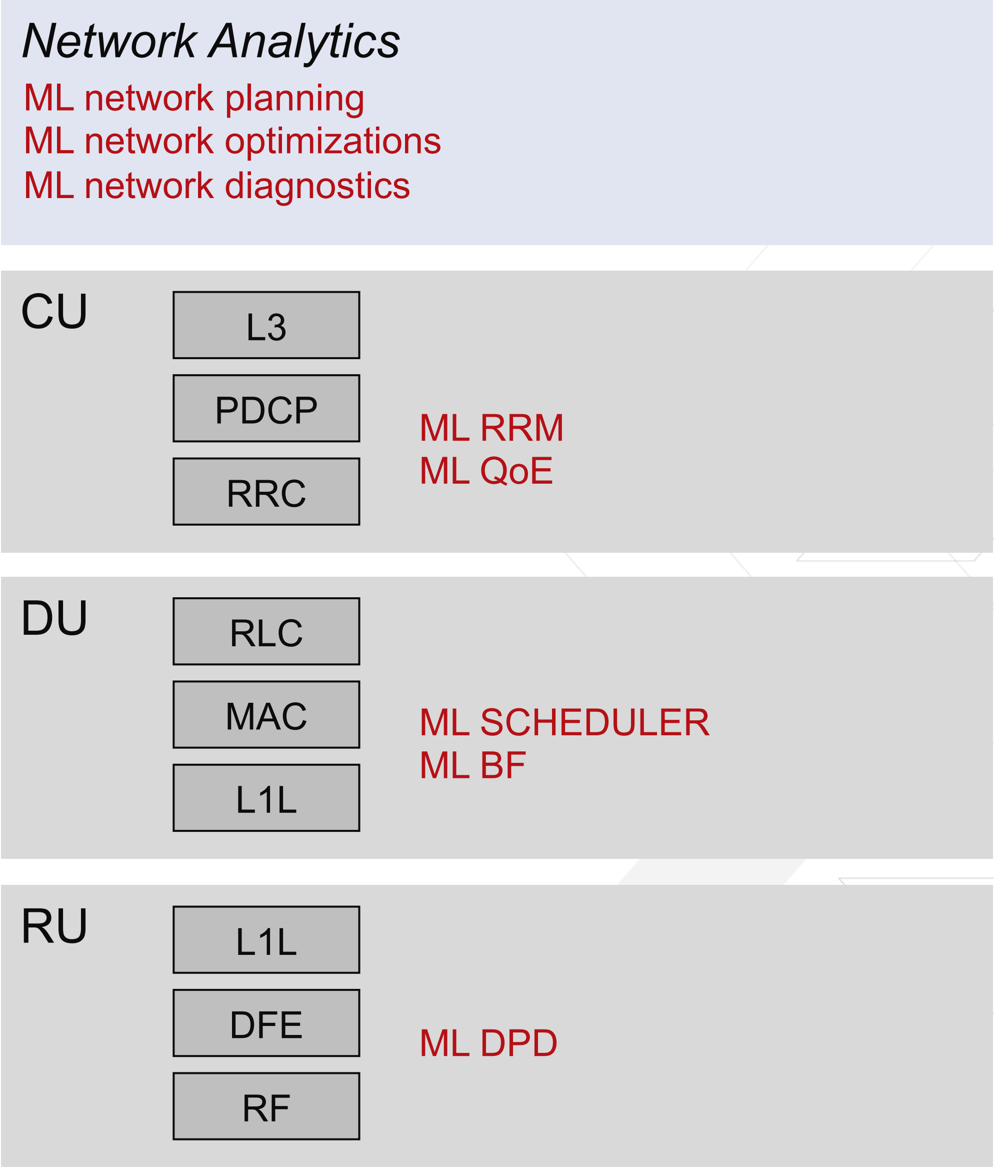
- Implemented on programmable device (using DACs and ADCs for Tx and Rx respectively)
- Low cost intensity modulation in transmitter and intensity detection in receiver
- BER of 10^{-6} at 12Gbps achievable with using ML in receiver.



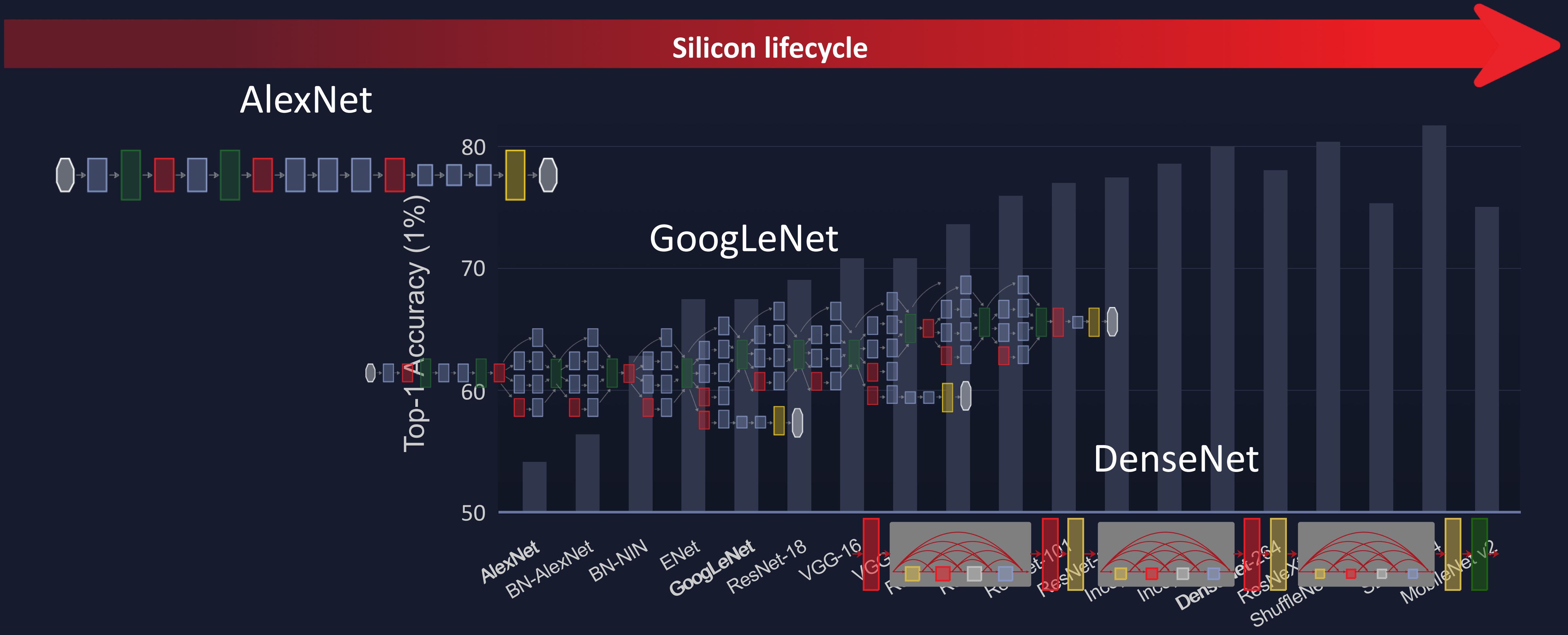
Source:customer

Examples of ML in wireless networks

- ML Network analytics to optimize OPEX/CAPEX and user experience
  - Automatically without user interaction
- Certain RRM behaviourals can be hard to model
- Massive MIMO scheduler complexity
- Beamforming optimization
- PA agnostic digital predistortion
- Improve Radio energy efficiency



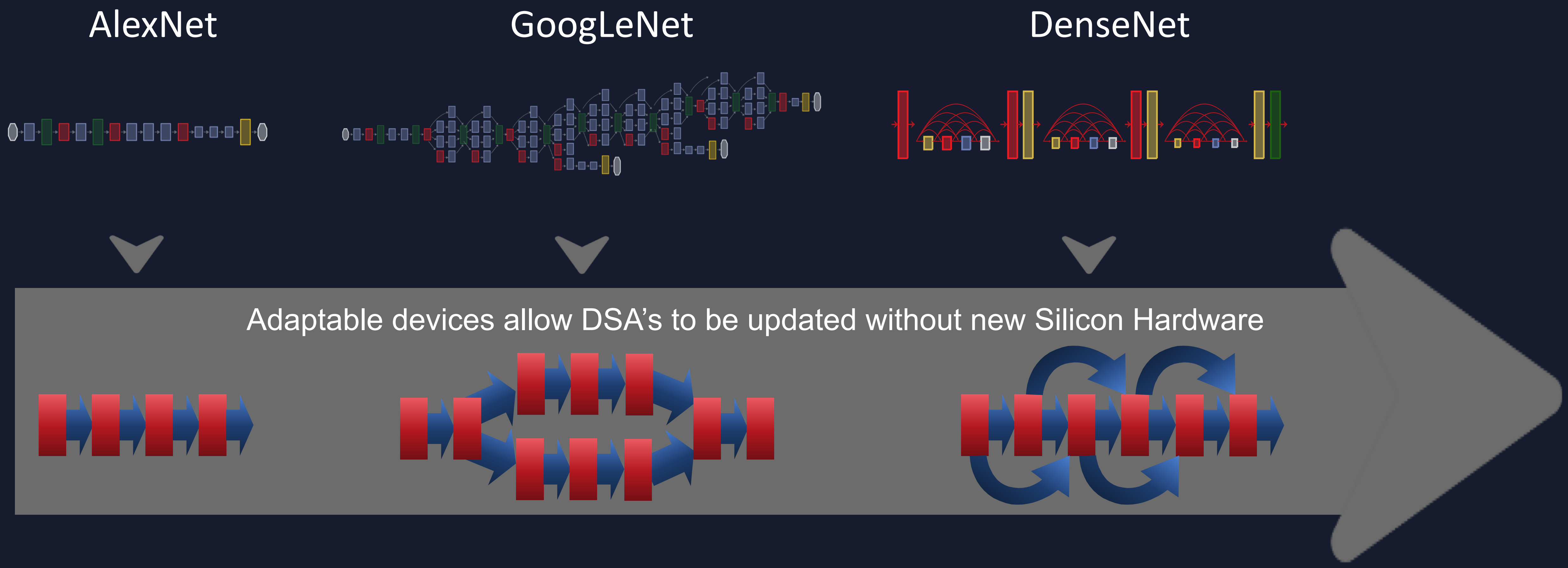
AI is Evolving Incredibly Fast



Silicon Hardware Design Cycle Can't Keep Up with the Rate of AI Innovation

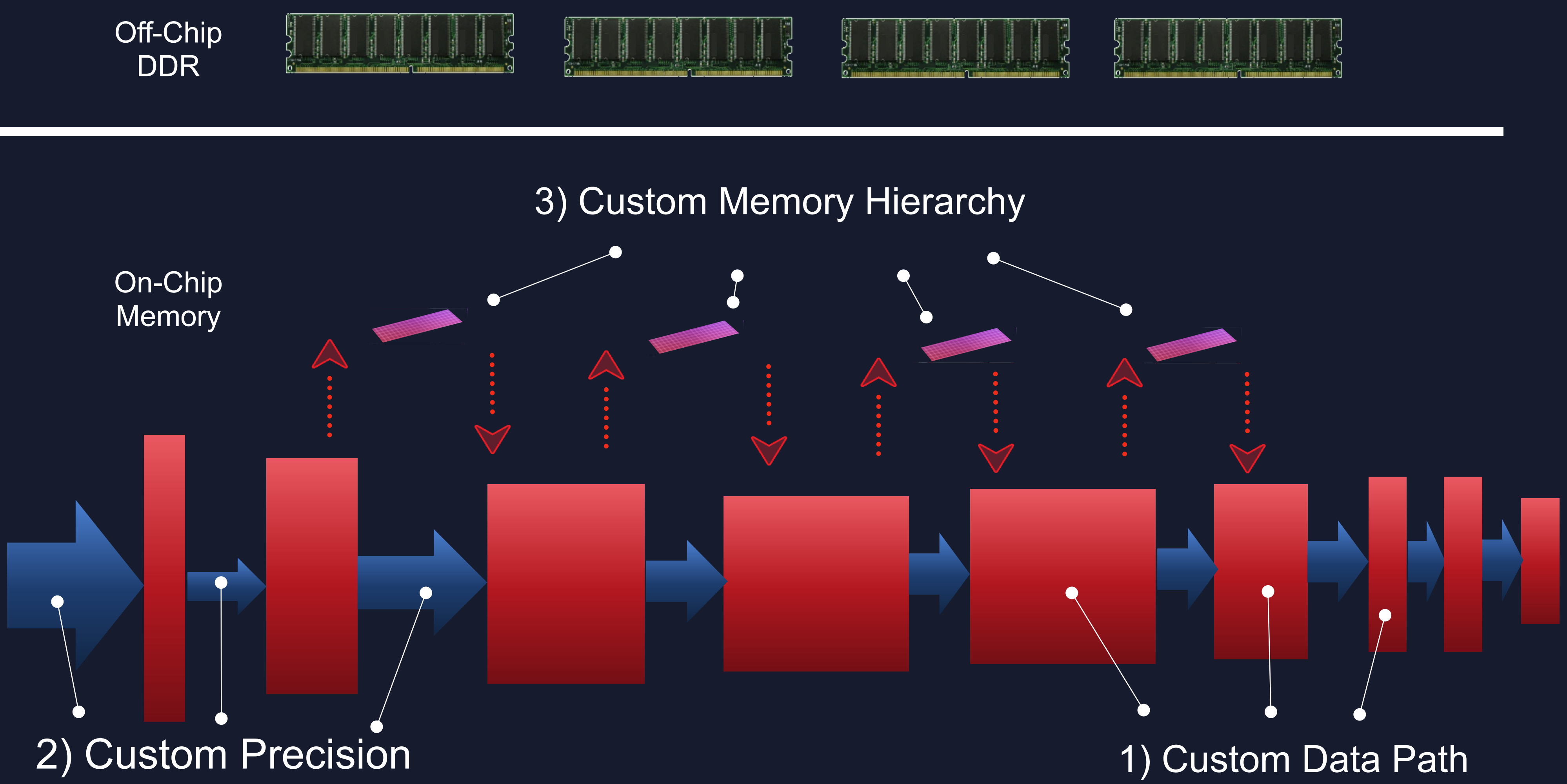
XILINX

Adaptive Hardware Enables flexible AI

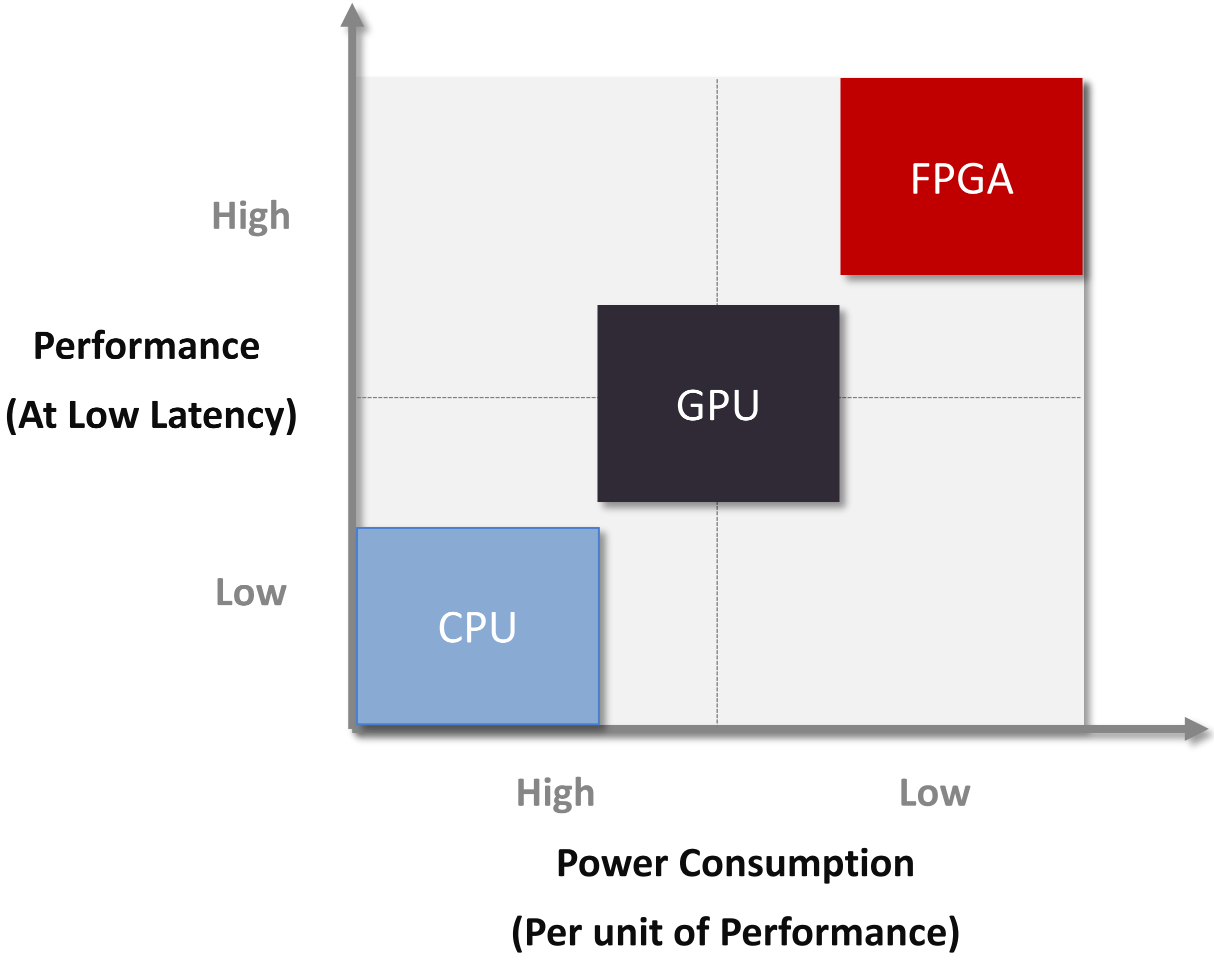


Only programmable hardware Can Keep Up with the Rate of AI Innovation

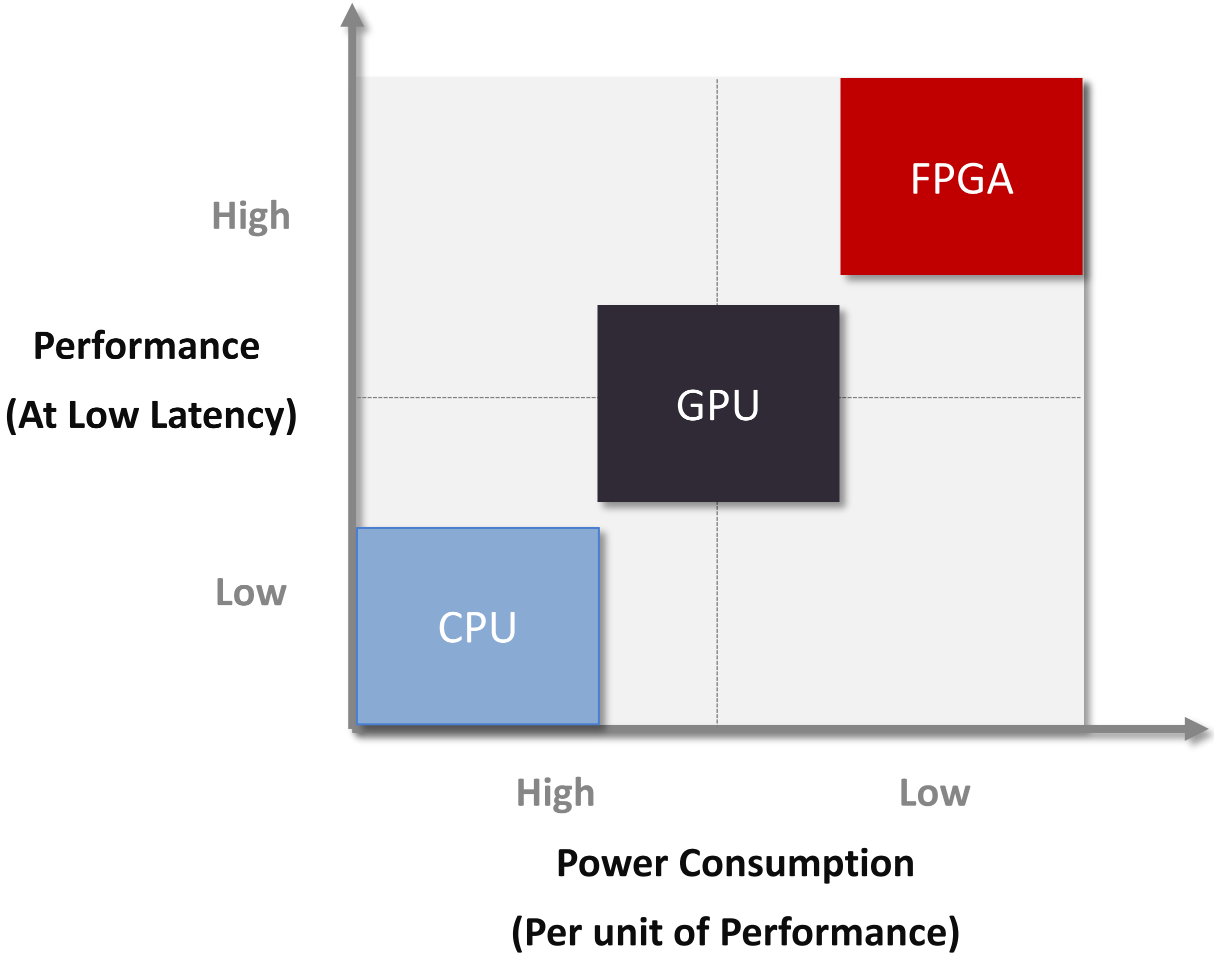
FPGA based architecture for ML



Machine learning platforms



Hardware based Machine learning



- ML based prediction in Networks requires
  - Low latency
  - High Performance

Hardware based ML give the Highest Performance at Low Latency, with optimal power

XILINX

Running DNN on Programmable hardware

Adaptable

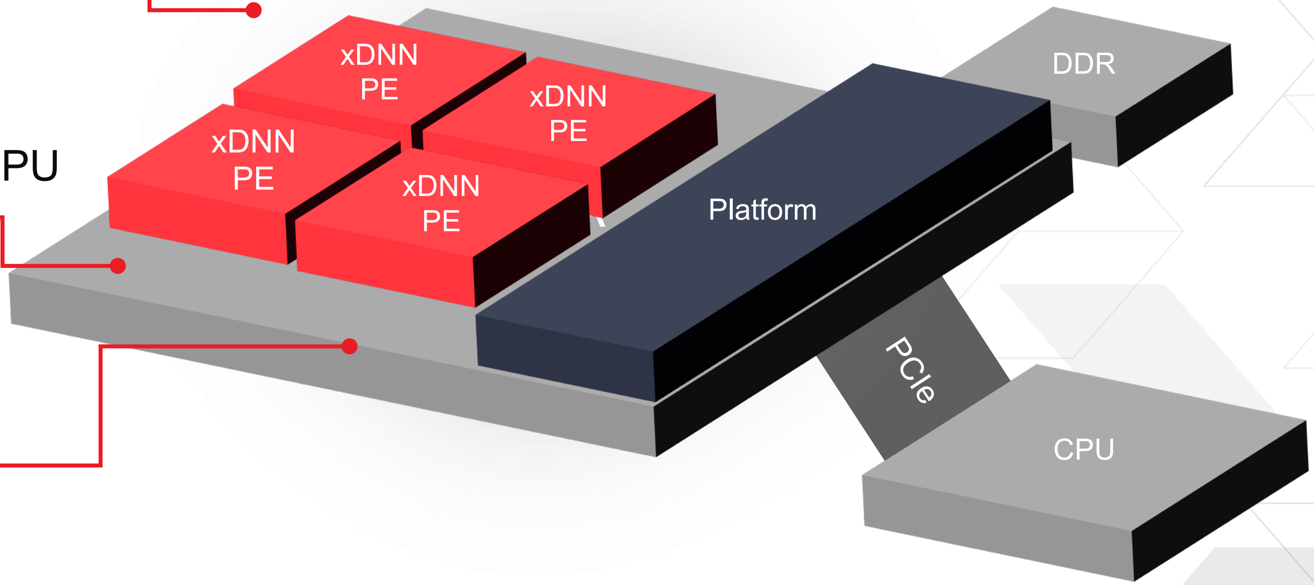
- AI algorithms are changing rapidly
- Adjacent acceleration opportunities

Realtime

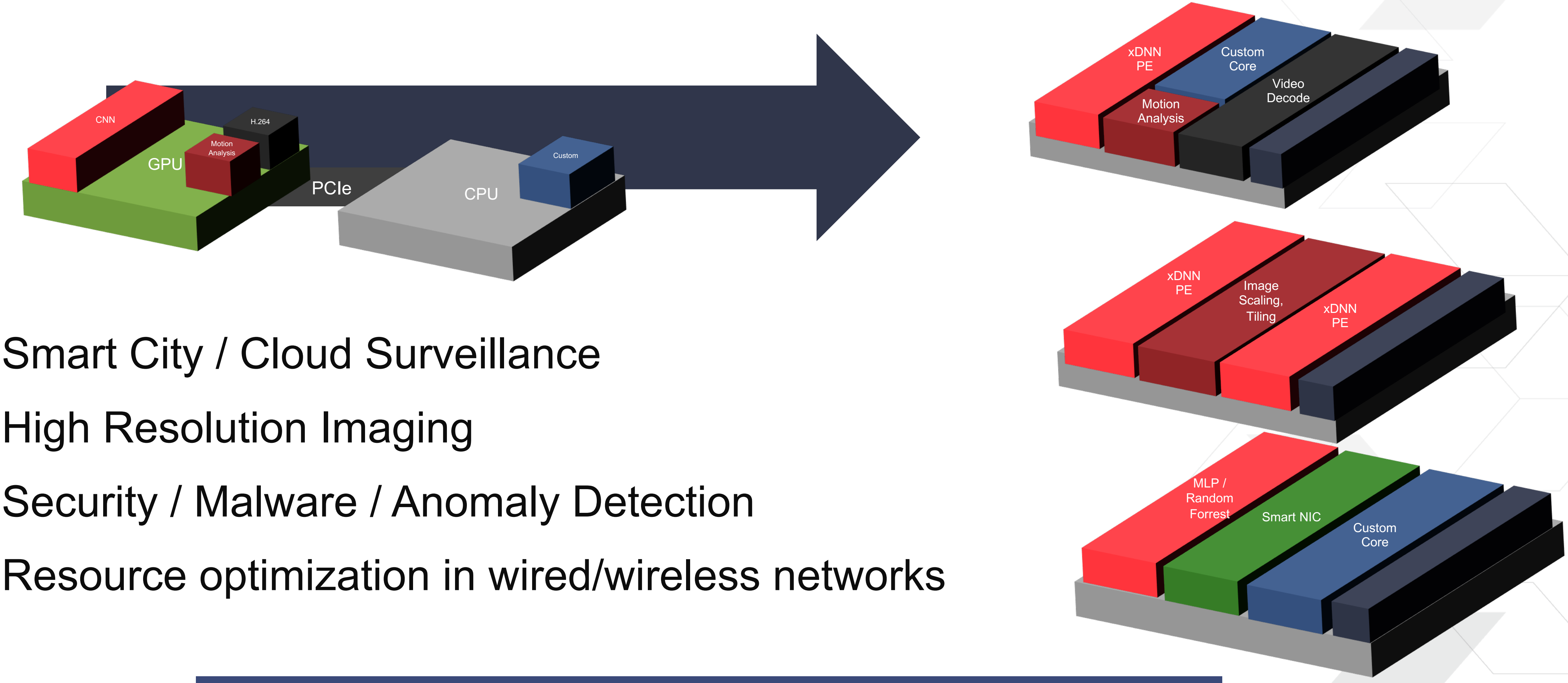
- 10x Low latency than CPU and GPU
- Data flow processing

Efficient

- Performance/watt
- Low Power



ML in Programmable hardware – Advantage

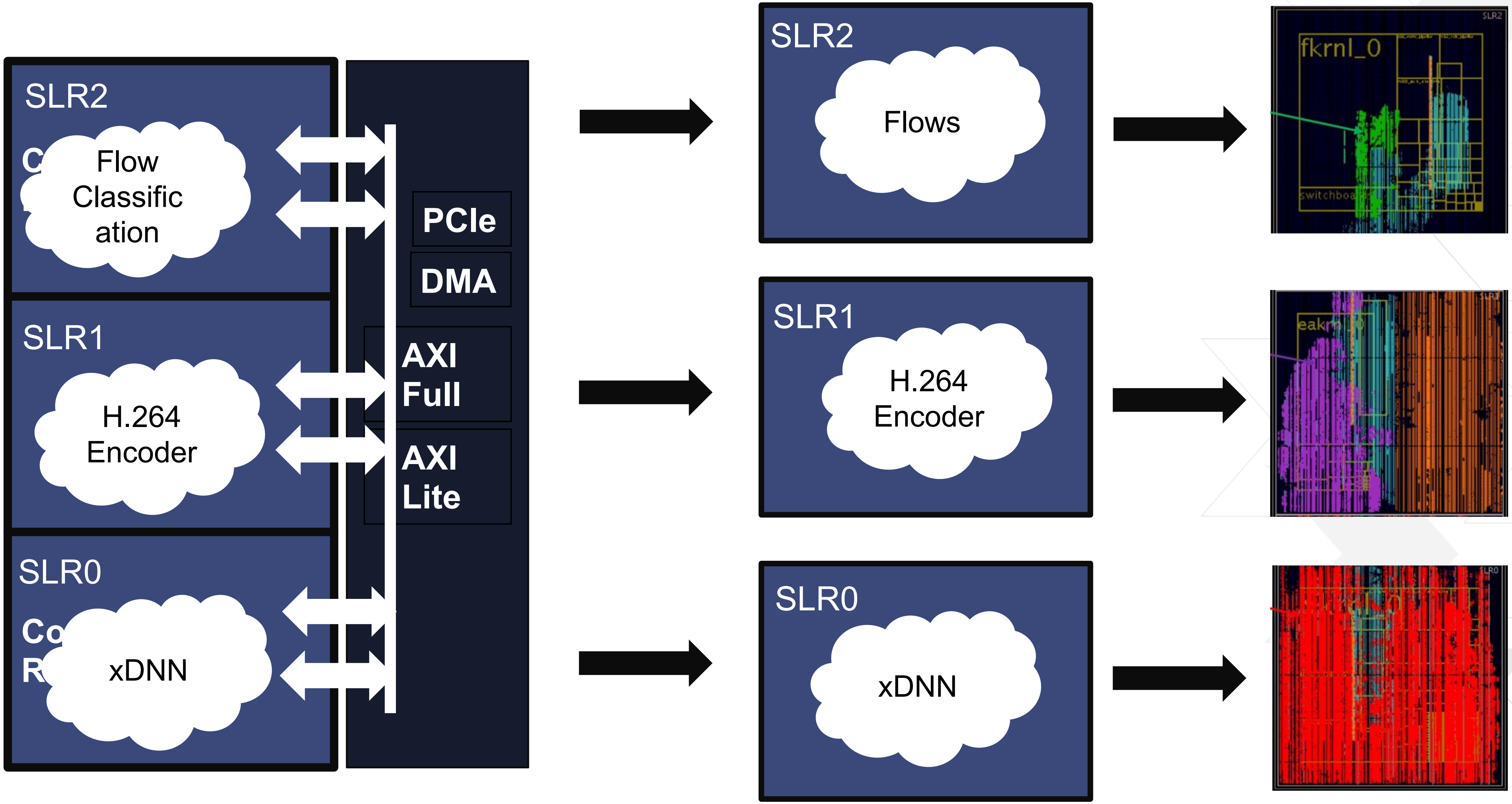


> Smart City / Cloud Surveillance
> High Resolution Imaging
> Security / Malware / Anomaly Detection
> Resource optimization in wired/wireless networks

FPGA Advantage: Latency, Cost, Area

XILINX

Multiple Kernels on programmable hardware



Paper | Release Notes (v0.6.2)

Cornell University

arXiv.org > cs > arXiv:1903.08066

Computer Science > Computer Vision and Pattern Recognition

Trained Uniform Quantization for Accurate and Efficient Neural Network Inference on Fixed-Point Hardware

Sambhav R. Jain, Albert Gural, Michael Wu, Chris Dick

(Submitted on 19 Mar 2019)

We propose a method of training quantization clipping thresholds for uniform symmetric quantizers using standard backpropagation and gradient descent. Our quantizers are constrained to use power-of-2 scale-factors and per-tensor scaling for weights and activations. These constraints make our methods better suited for hardware implementations. Training with these difficult constraints is enabled by a combination of three techniques: using accurate threshold gradients to achieve range-precision trade-off, training thresholds in log-domain, and training with an adaptive gradient optimizer. We refer to this collection of techniques as Adaptive-Gradient Log-domain Threshold Training (ALT). We present analytical support for the general robustness of our methods and empirically validate them on various CNNs for ImageNet classification. We are able to achieve floating-point or near-floating-point accuracy on traditionally difficult networks such as MobileNets in less than 5 epochs of quantized (8-bit) retraining. Finally, we present Graffitist, a framework that enables immediate quantization of TensorFlow graphs using our methods. Code available at this https URL .

Comments: 17 pages, 9 figures
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Machine Learning (cs.LG)
Cite as: arXiv:1903.08066 [cs.CV]
(or arXiv:1903.08066v1 [cs.CV] for this version)

Paper: https://arxiv.org/abs/1903.08066

Xilinx / aie-ml-stack Private

<> Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Releases | Tags

Pre-release

v0.6.2

17fa8cc

AIE ML Stack 0.6.2

sjain-stanford released this 5 days ago

Release 0.6.2

Updates

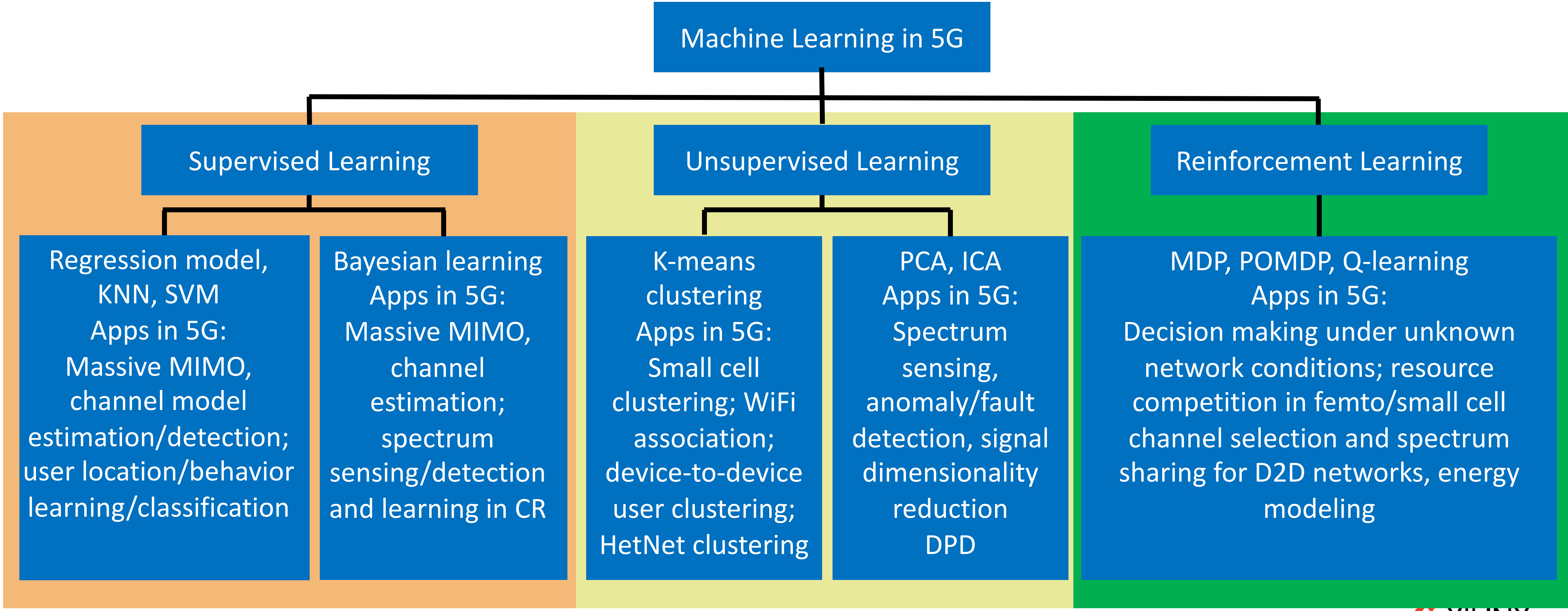
- Released training and validation scripts for Graffitist quantized networks
- Released calibration set generation script for provided networks

Complete Feature List

- Graffitist | Quantization (INT8)
- Graffitist | Retraining (INT8, INT4)
- Networks supported
  - Inception v{1, 2, 3, 4}
  - ResNet v1 {50, 101, 152}
  - VGG {16, 19}
  - MobileNet v{1, 2}
  - DarkNet 19
  - networks with topologies/compute layers similar to above
- Bit-approximate to Xilinx AI Engine (AIE)

Machine Learning in 5G

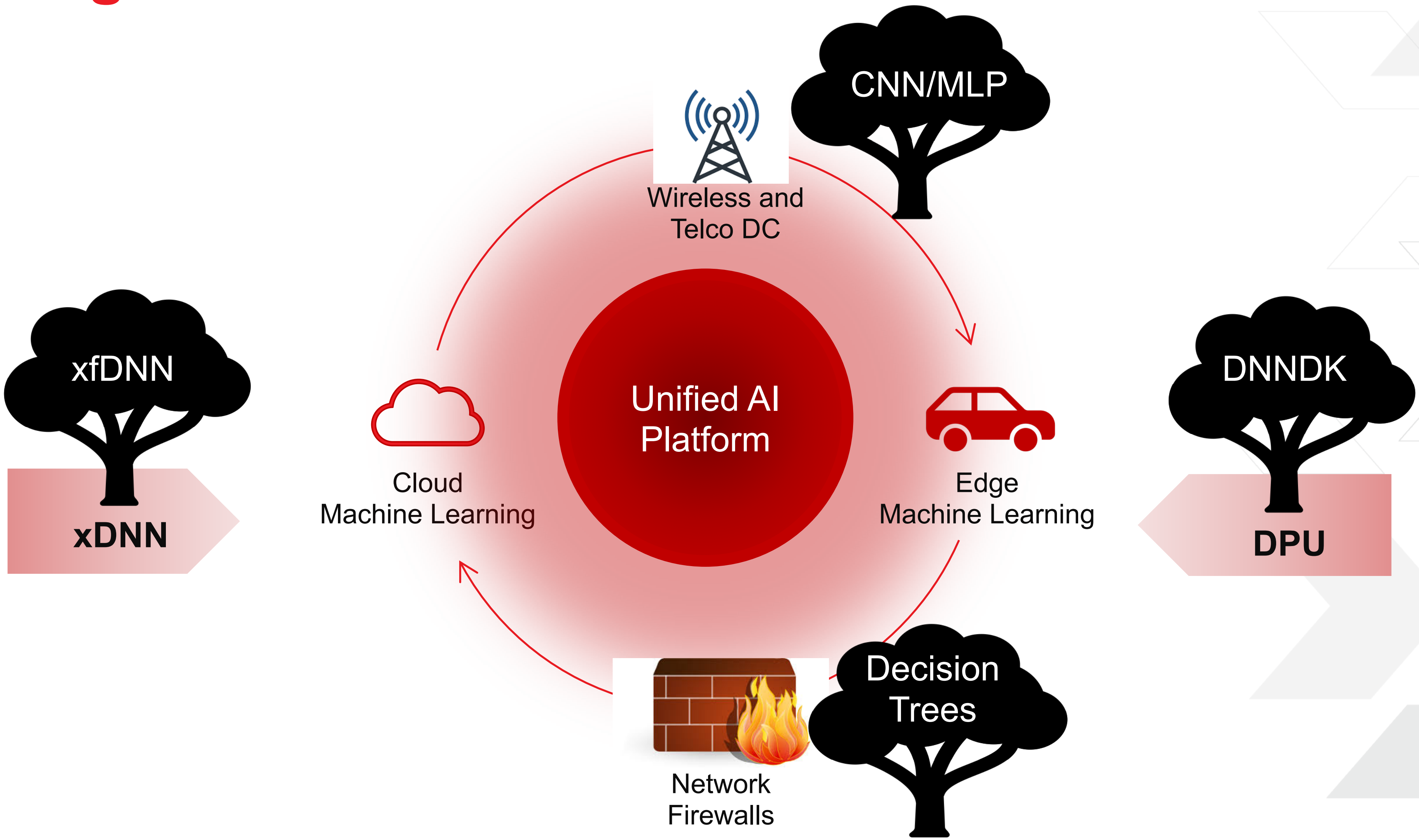
- ML tools enabling rapid deployment and integration of ML functions in radio and BB
- From ML network description in, for example, TensorFlow → Versal implementation



ML in Communications - Summary

- Next generation networks are complex, ML and AI will play a key role in
  - Management and control of network parameters (QoS, Filtering)
  - Security and threat prevention
  - Anomaly detection
  - Telemetry
- Supervised learning are most common use cases for applying ML in comms
  - Random forest
  - MLPs
- Telco DC in wired and wireless can easily deploy ML.
- In wireless networks – beamforming, Radio resource optimization can use ML.

Moving Towards Unified AI Platform



Adaptable solutions, compatible tools, uncompromised performance

XILINX.

Adaptable.
Intelligent.

XILINX