FPGA- based approach to efficient on-board data processing using deep neural networks.

Krzysztof Czyz*, Mateusz Maciag*, Jacek Lach*, Marcin Kurczalski*,
Jakub Nalepa**, Pablo Lorenzo Ribalta**

13/11/2018

EARTH OBSERVATION CHALLENGES

- large number of data generated by instruments
- data updated at a steadily higher frequency
- variety of data being generated (eg. hyperspectral, SAR)
- traditional EO enterprises provide raw images or do some basic levels of postprocessing and leave the end users to build derivatives.

source: MEETING TOMORROW 'S EARTH OBSERVATION CHALLENGES, www.luciad.com

Example: Hyperspectral EO Instrument

A high resolution hyperspectral instrument for EO operating at 600 km SSO:

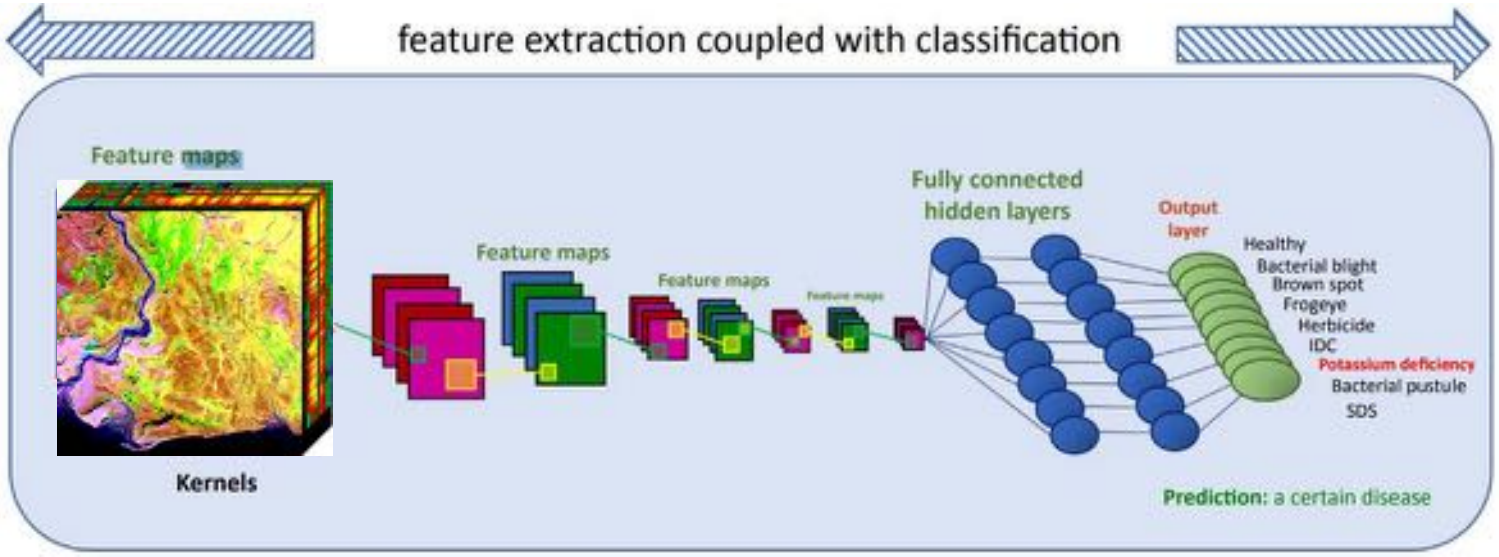
- swath: 32 km
- ground sampling distance: 12.5m
- spectral range: 450 nm to 900 nm
- no. of spectral bands: 75
- RAW data per second: 2.8GiB
  746 GiB per pass over the Europe
- pre-processed data per second: 0.22GiB
  34 GiB per pass over the Europe

Handling, storing and transferring of such amount of data to the ground is a challenge.

What if on-board processing is possible?

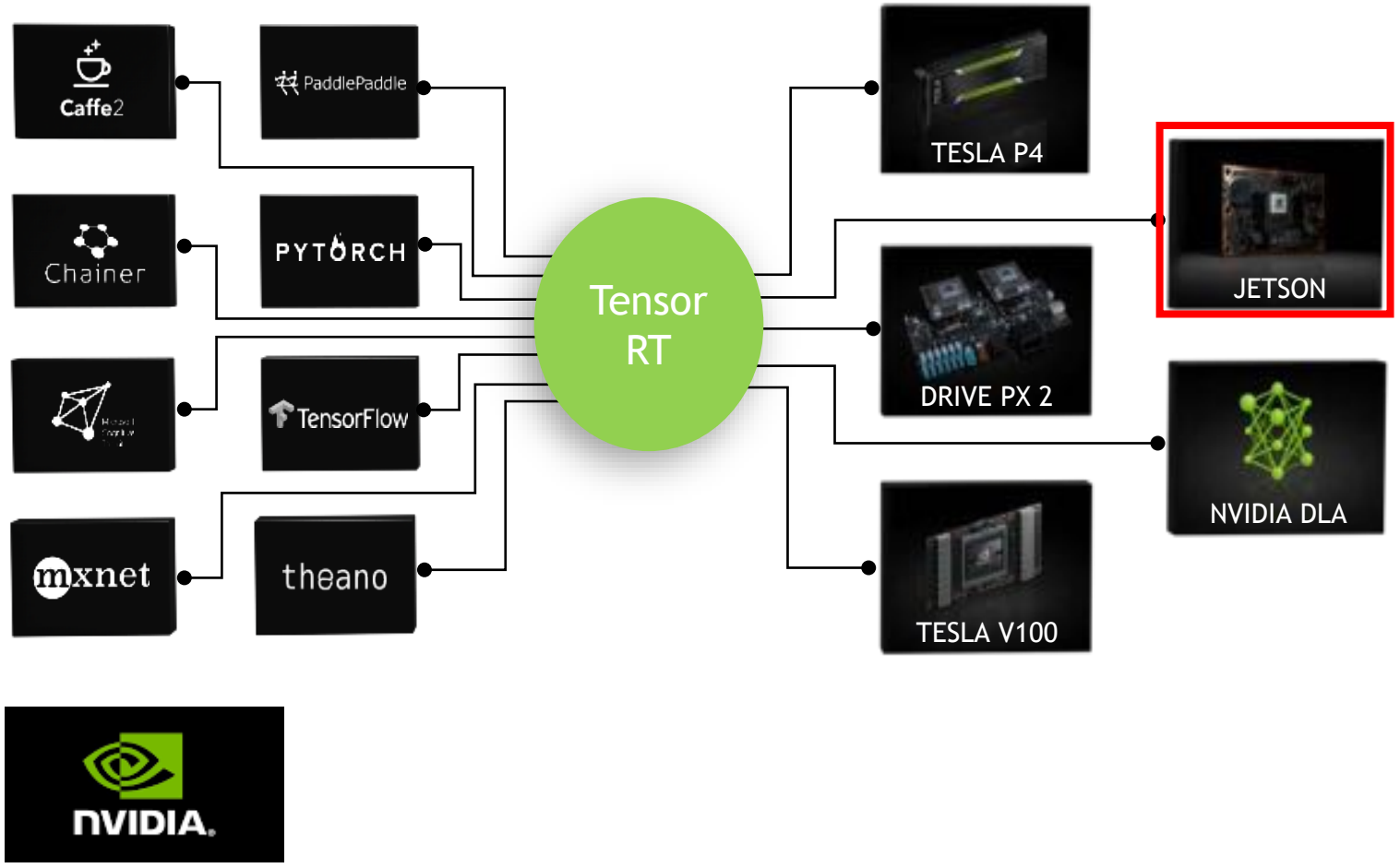
Deep Neural Network Classifier

On-board data classification can be a tool to reduce data volume by orders of magnitude simultaneously providing a rapid reaction to detected patterns.

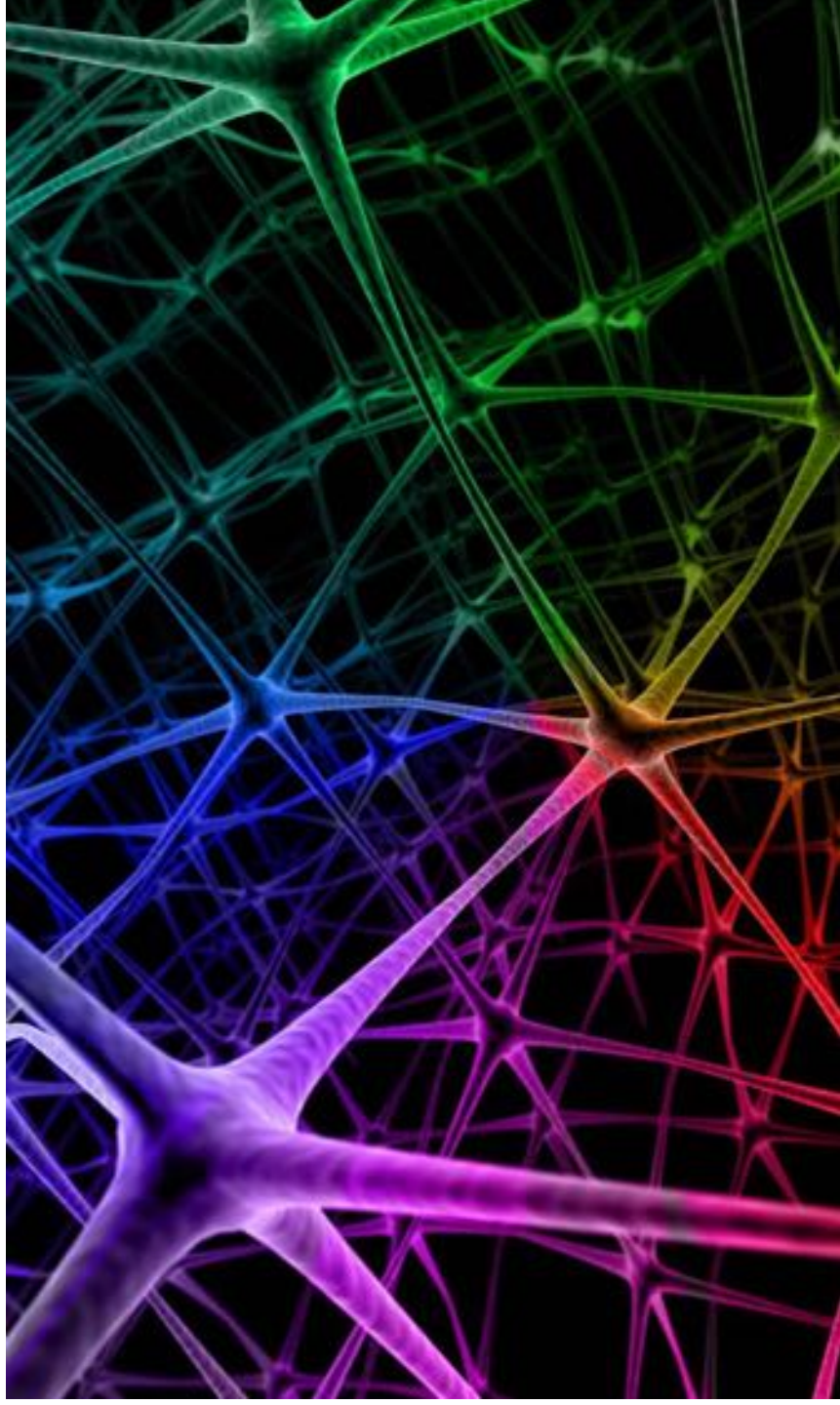


How to do it? Is there any tool enabling such approach on orbit?

DNN: Frameworks and Hardware



source:AI At the Edge , Rammy Bahalul, nVidia



JETSON TX2i

SUPERCOMPUTER FOR AI AT THE EDGE

PROS:

- computing performance >1 TFLOPS
  AlexNet 463 FPS, GoogLeNet 196 FPS
- power efficiency <10W
- good tools

CONS:

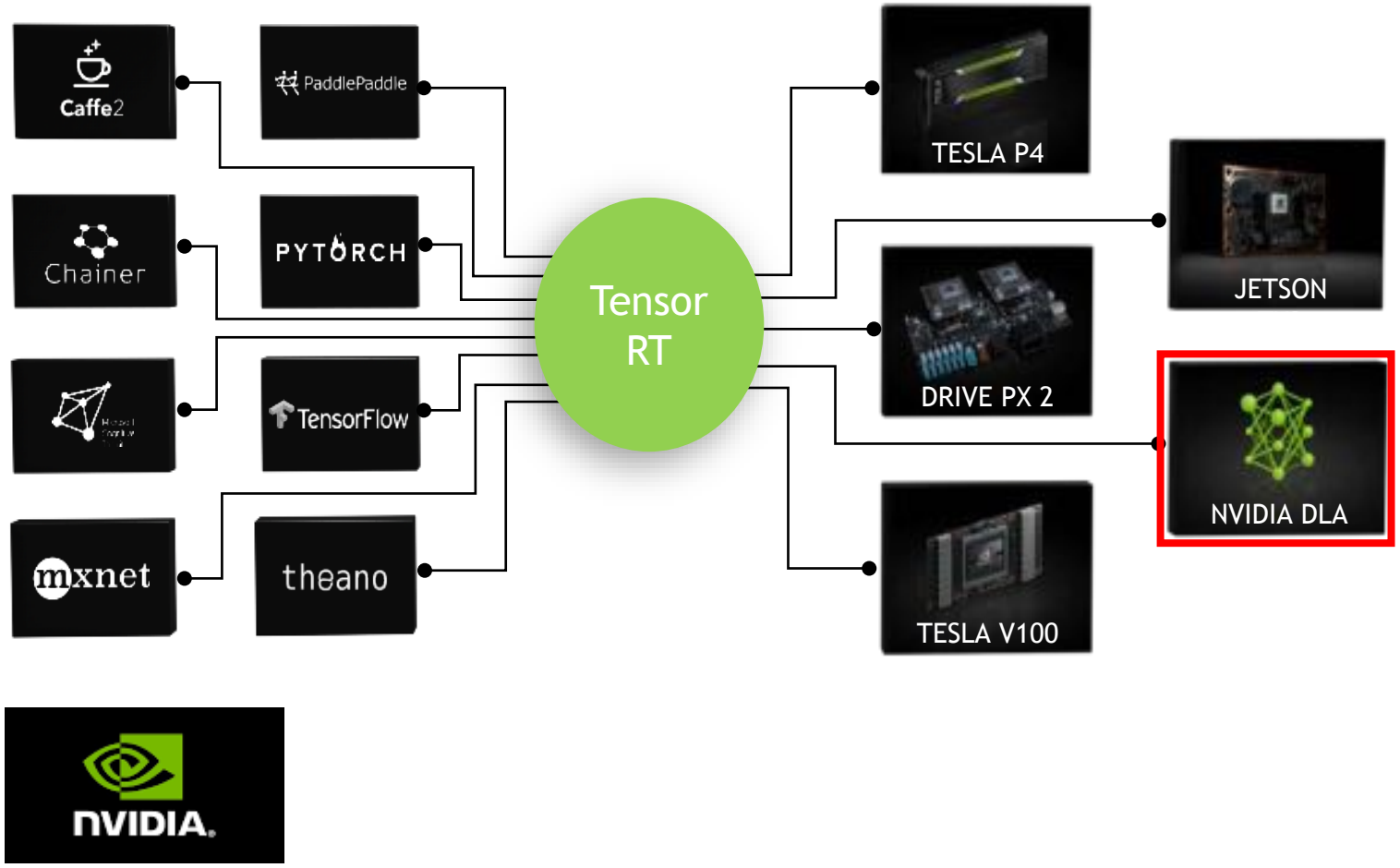
- Tegra X2 available as Jetson module
  limited customisation
- fragile flash memory on board (eMMC)
- no SEU resistance (including power rails)

cuDNN, TensorRT

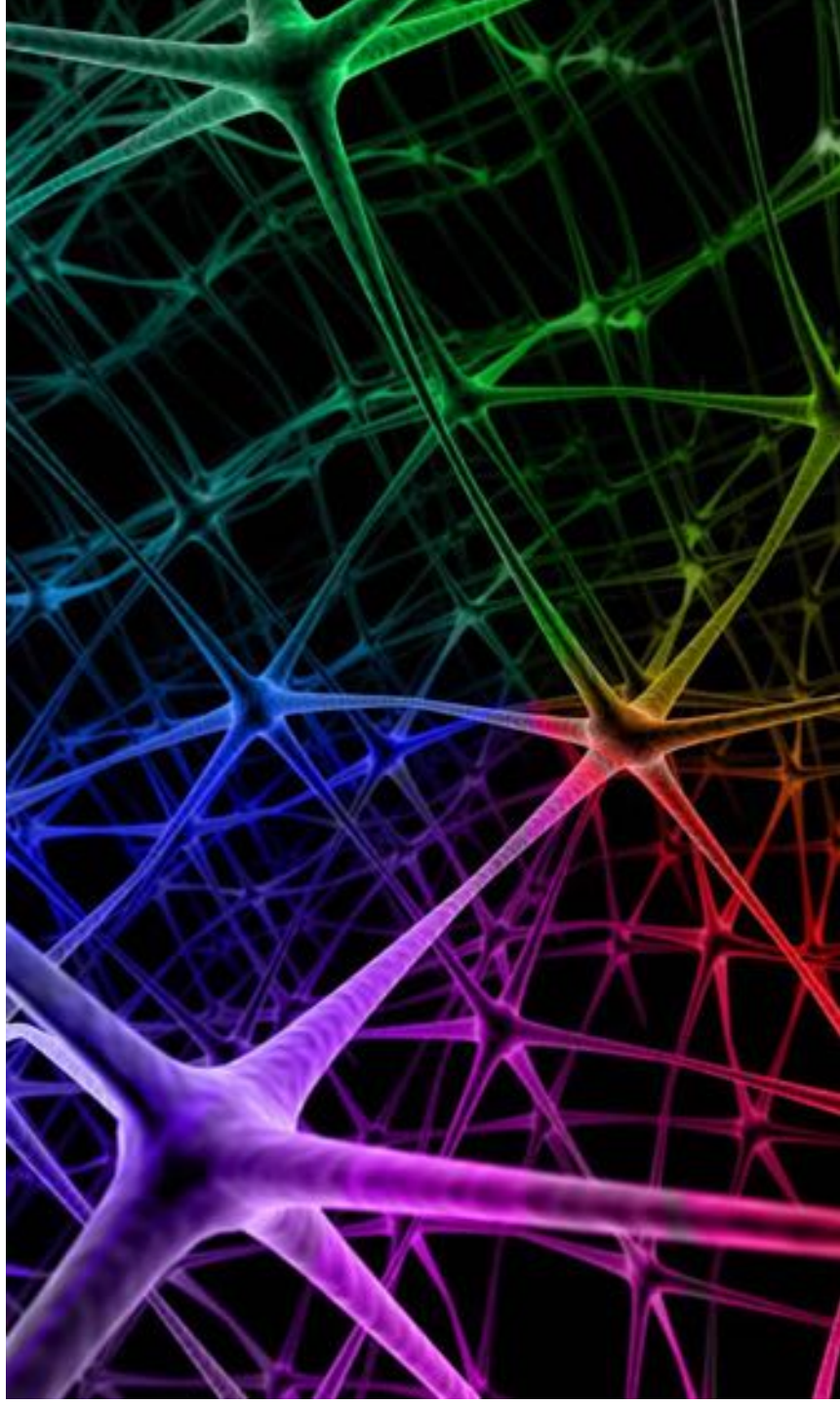
CUDA

Linux, ROS

DNN: Frameworks and Hardware



source:AI At the Edge , Rammy Bahalul, nVidia



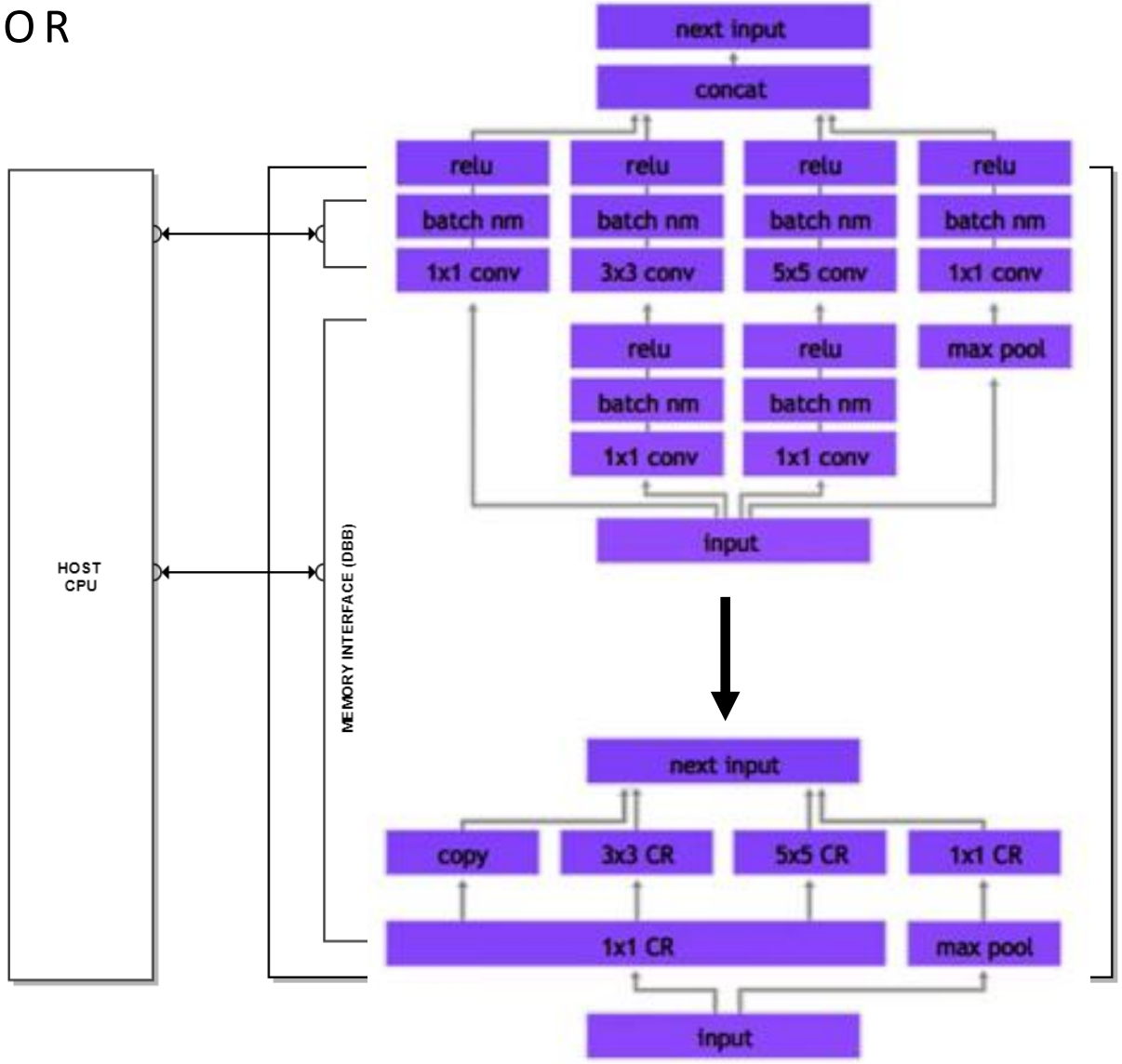
FP Space

NVIDIA Deep Learning Accelerator

WHAT IS IT? HARDWARE

A CPU independent neural network accelerator created by NVIDIA providing:

- Convolution Core
  Direct/Winograd/Multi-Batch Convolution
  FC layers
- Single Data Processor
  Activation functions (from ReLU to non-linear)
  Bias addition
  Batch normalization
  Element-wise operations
- Planar Data Processor
  Pooling (min, max, average)
- Channel Data Processor
  Local normalization functions



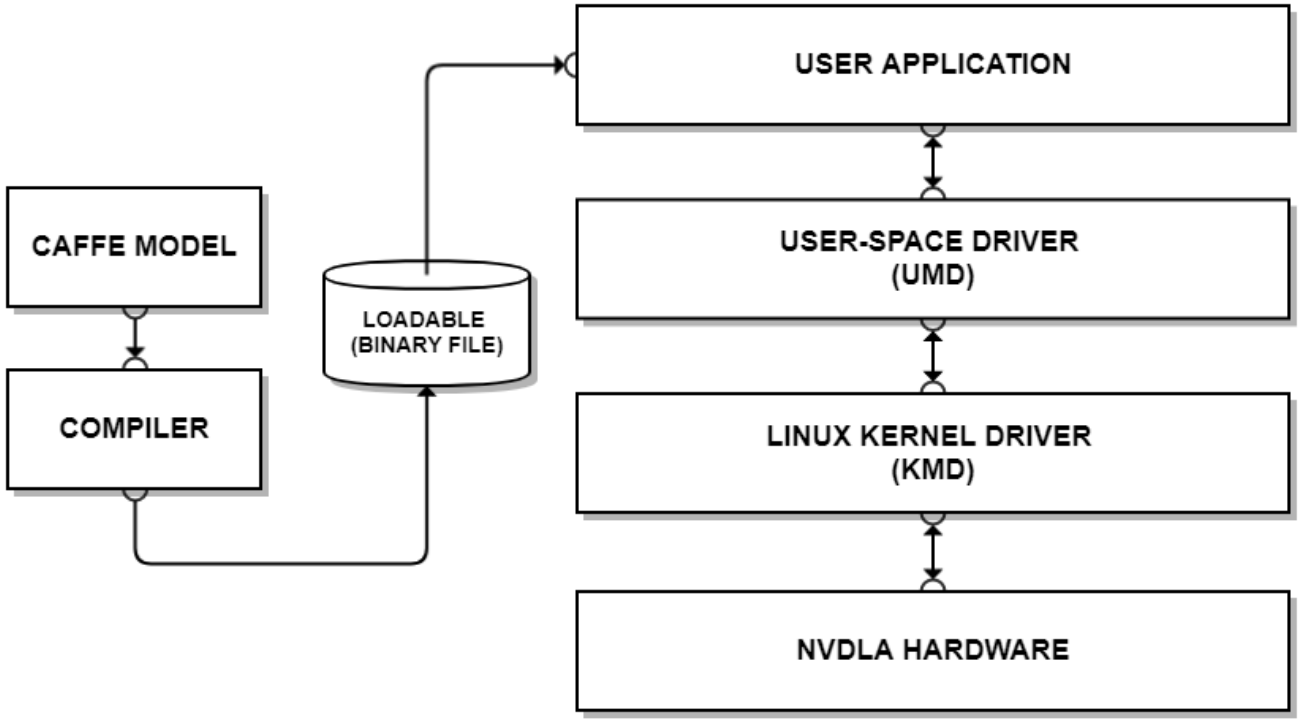
FP Space

NVIDIA DEEP LEARNING ACCELERATOR

WHAT IS IT? SOFTWARE

nvDLA software ecosystem:

- Caffe model
  Use well established framework and deep learning data represenations
- Compiler
  Translates DNN layers to list of low-level operations.
- User Application
  Run inference
- UMD/KMD drivers
  Hardware Abstraction Layer to the deep learning accelerator



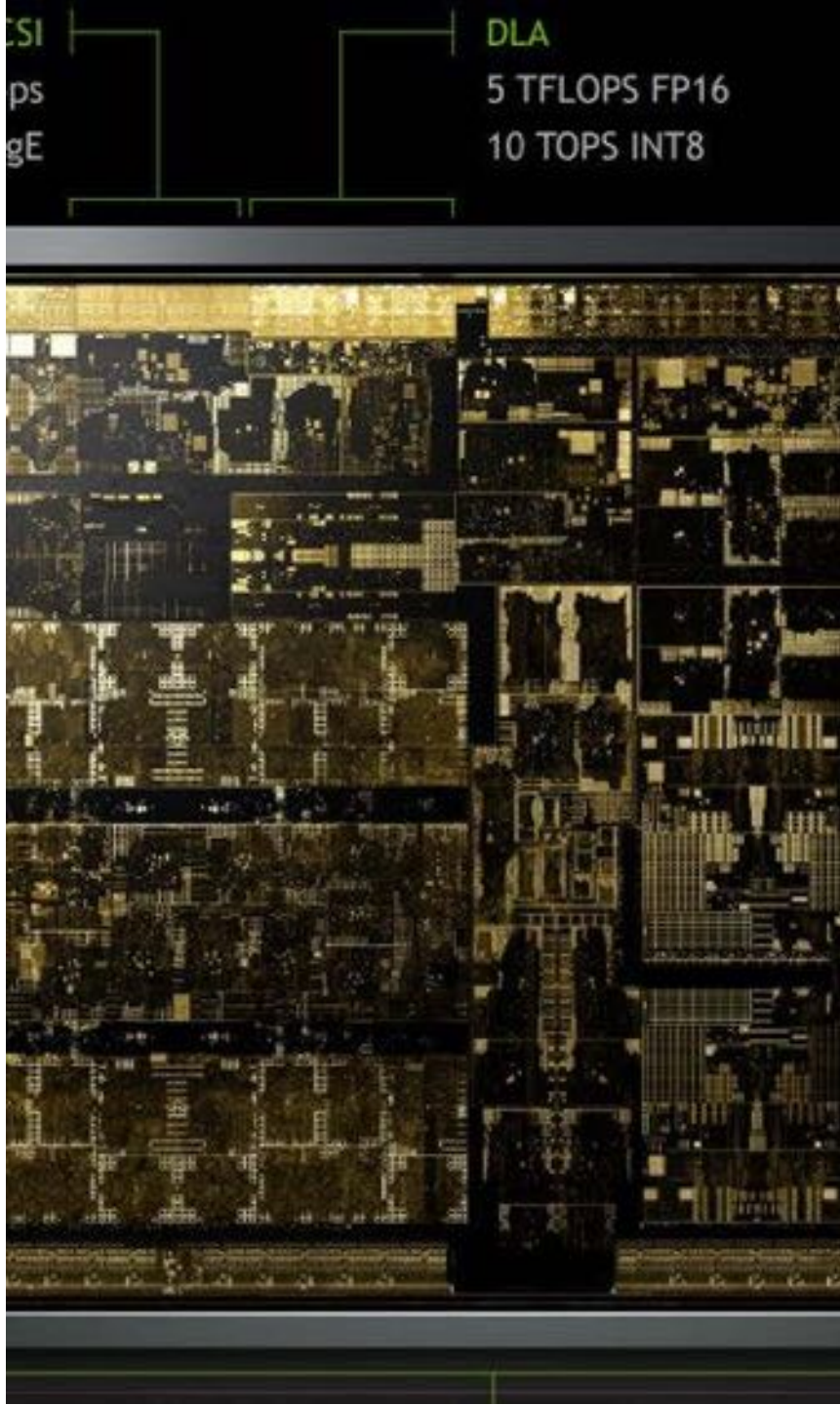
NVIDIA Deep Learning Accelerator

WHAT IS IT? HARDWARE IMPLEMENTATION

nvDLA is meant to be implemented in hardware i.e. Xavier SOC.

NVIDIA has decided to go to open source with nvDLA in 2017 and has released Verilog RTL:

- End of 2017
  fully featured, floating point, non-configurable accelerator (NV_FULL)
- 2018 Q1/Q2
  small, fixed-point (INT8), non-configurable accelerator (NV_SMALL)
- 2018 Q3
  added degrees of freedom in architecture configuration (NV_CUSTOM)



NVIDIA DEEP LEARNING ACCELERATOR

FPGA IMPLEMENTATION OF NVDLA

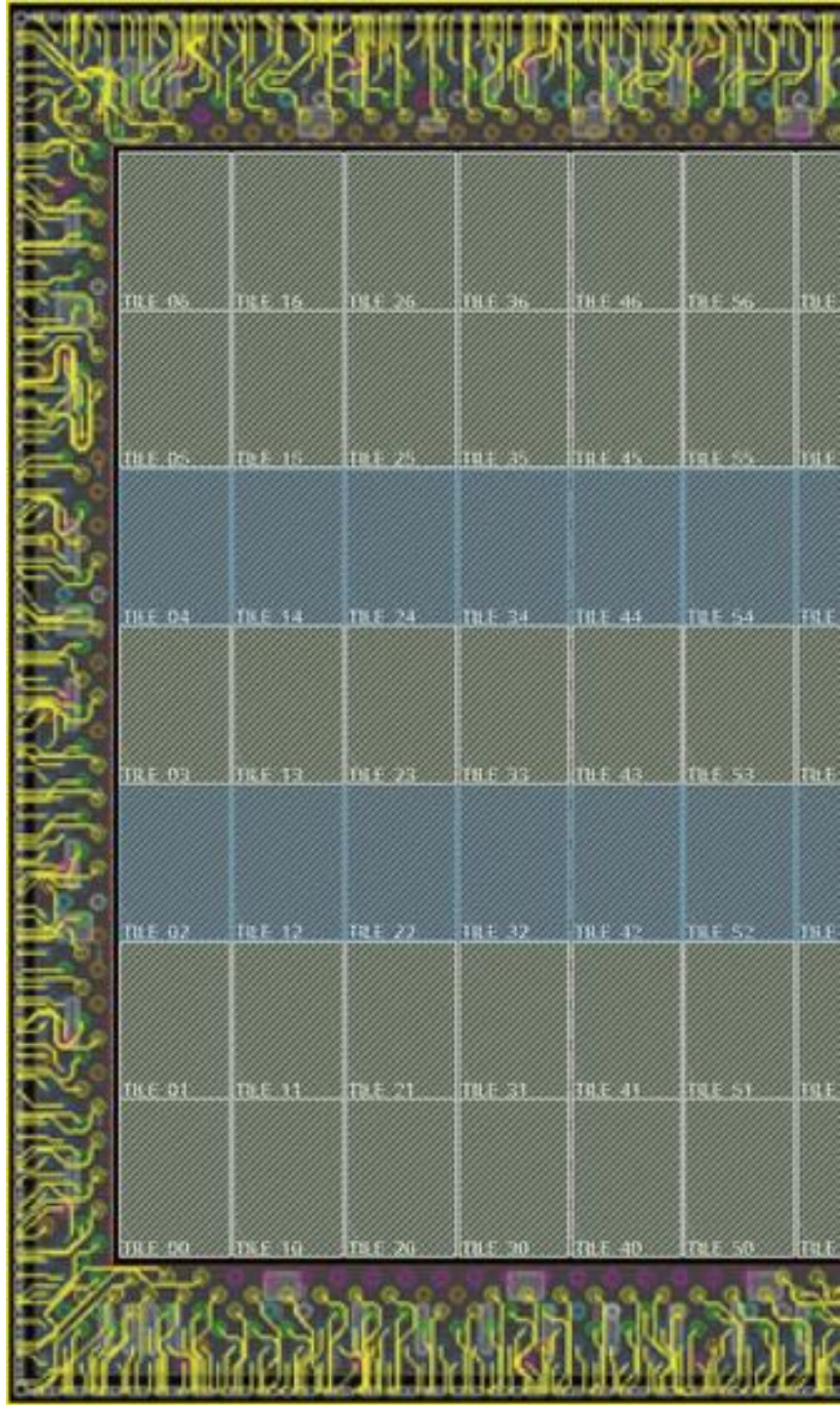
Goal: fit triple redundant nvDLA engine into Xilinx ZU15EG FPGA

PROS:

- customization
- resonable computing power of single nvDLA instance
  AlexNet 149 FPS (~33% of TX2), GoogLeNet 74 FPS (38% of TX2)
  for 512 MAC cores running @ 150 MHz
- FPGA mature technology in space

CONS:

- lack of DNN compiler for fixed point nvDLA,
- the only reasonable data representation is fixed-point INT8
- higher power consumption than Jetson TX2i
- nvDLA is a complex IP, consuming a lot of physical area



NVIDIA DEEP LEARNING ACCELERATOR

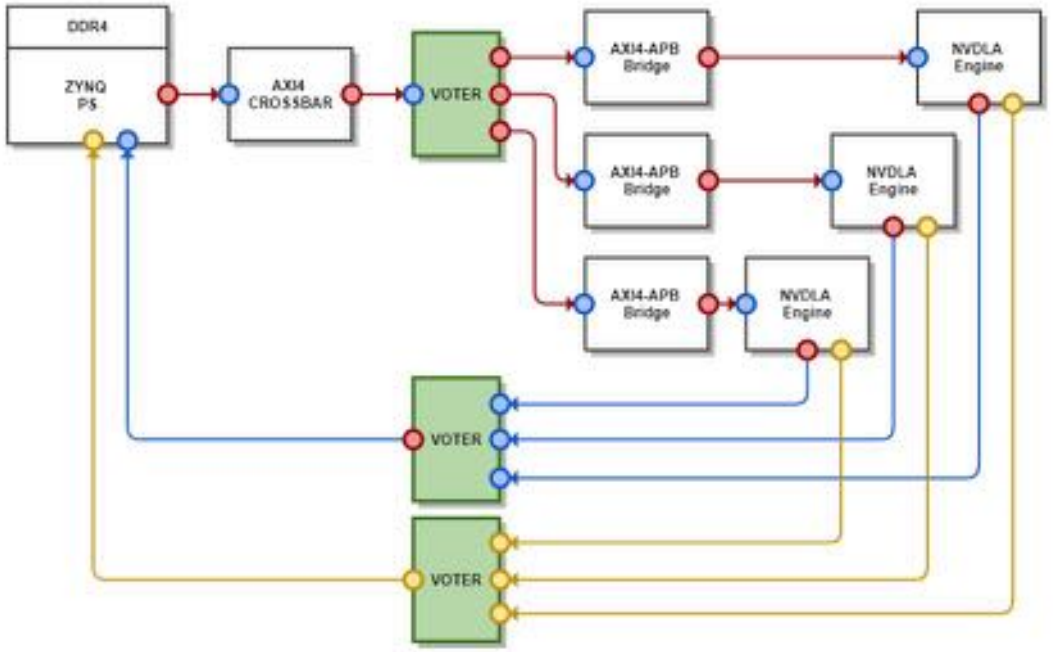
TRIPLE MODULAR REDUNDANCY – SIMPLE TMR VOTER

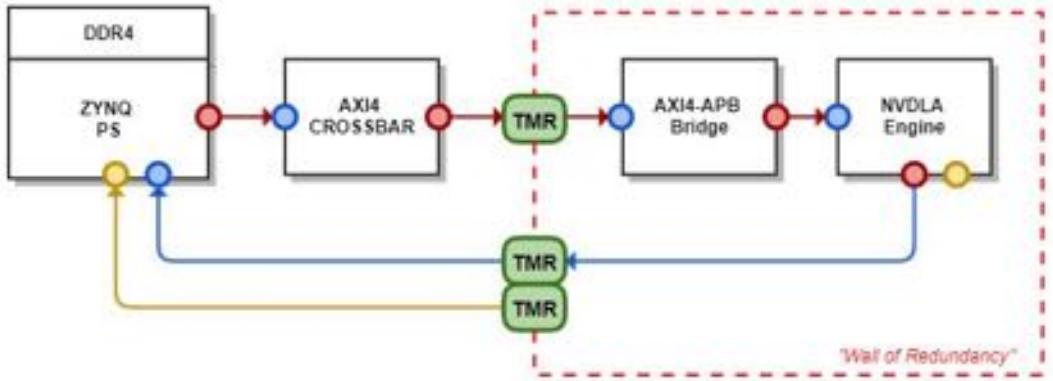
Pros:

- Uses off-the-shelf TMR solution (IP core from Xilinx library).

Cons:

- Single point of failure!





NVIDIA DEEP LEARNING ACCELERATOR

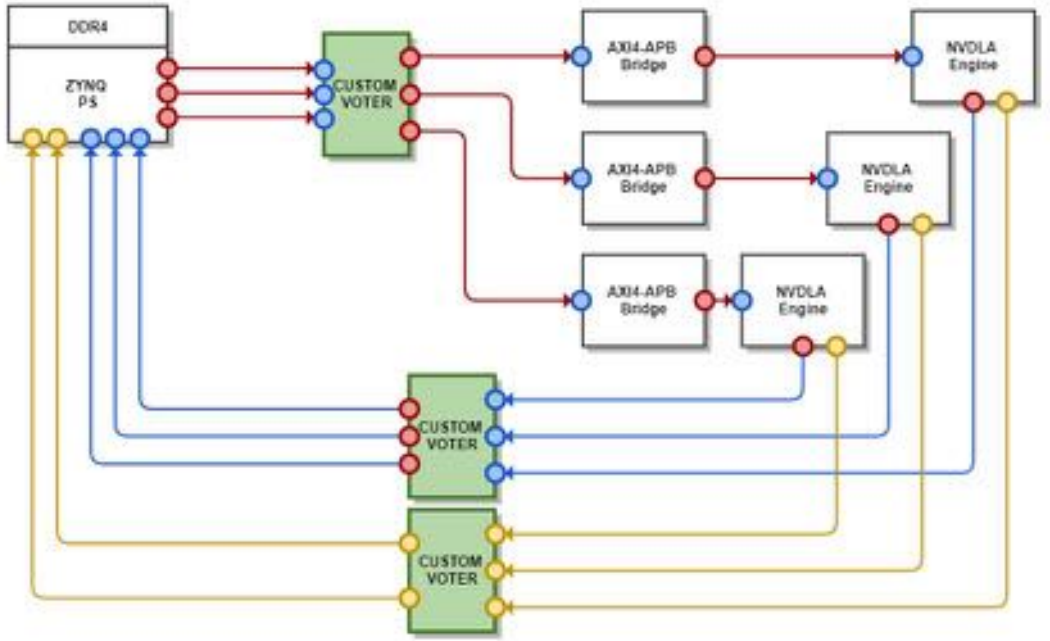
TRIPLE MODULAR REDUNDANCY – MULTIPLE MEMORY INTERFACES AND ADVANCED TMR VOTER

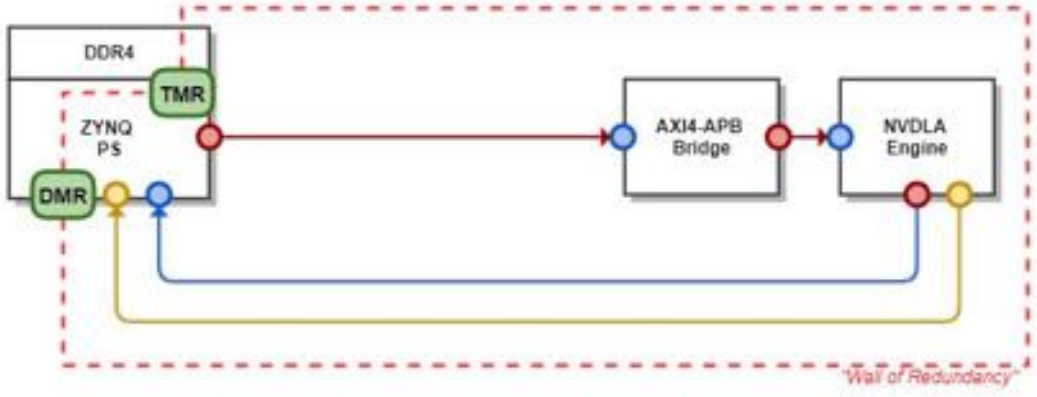
Pros:

- Fully redundant system (well, FPGA part at least)
- Input/output data copy reside in DDR4

Cons:

- Complicated TMR design – AXI4 may have distinct behaviour on each channel due to accessing the same DDR4 controller.
- DDR4 bandwidth and latencies might be an issue (concurrent)
- DDR4 increased usage (3x)





HyperPU processing unit

Development of processing unit providing:

- 2 x Xilinx MPSOC ZU15EG FPGA
- 6 x nvDLA cores (2 redundant units each running in TMR)
- >200 GOPS per core

  (AlexNET 149 FPS, GoogLeNet 74 FPS)
- >1200 GOPS per whole unit
- 2 x 16 GiB of RAM memory
- array of SLC flash memory for data storage

HyperCAM hyperspectral instrument

Innovative hyperspectral system for observing the Earth with increased spectral resolution enabling automatic processing and selection of satellite data in orbit based on new algorithms for segmentation and classification of satellite images using deep convolutional networks.

FP Space

THANK YOU

Krzysztof Czyż
kczyz@fpspace.com

Gliwice, Poland

sales@fpspace.com

www.fpspace.com