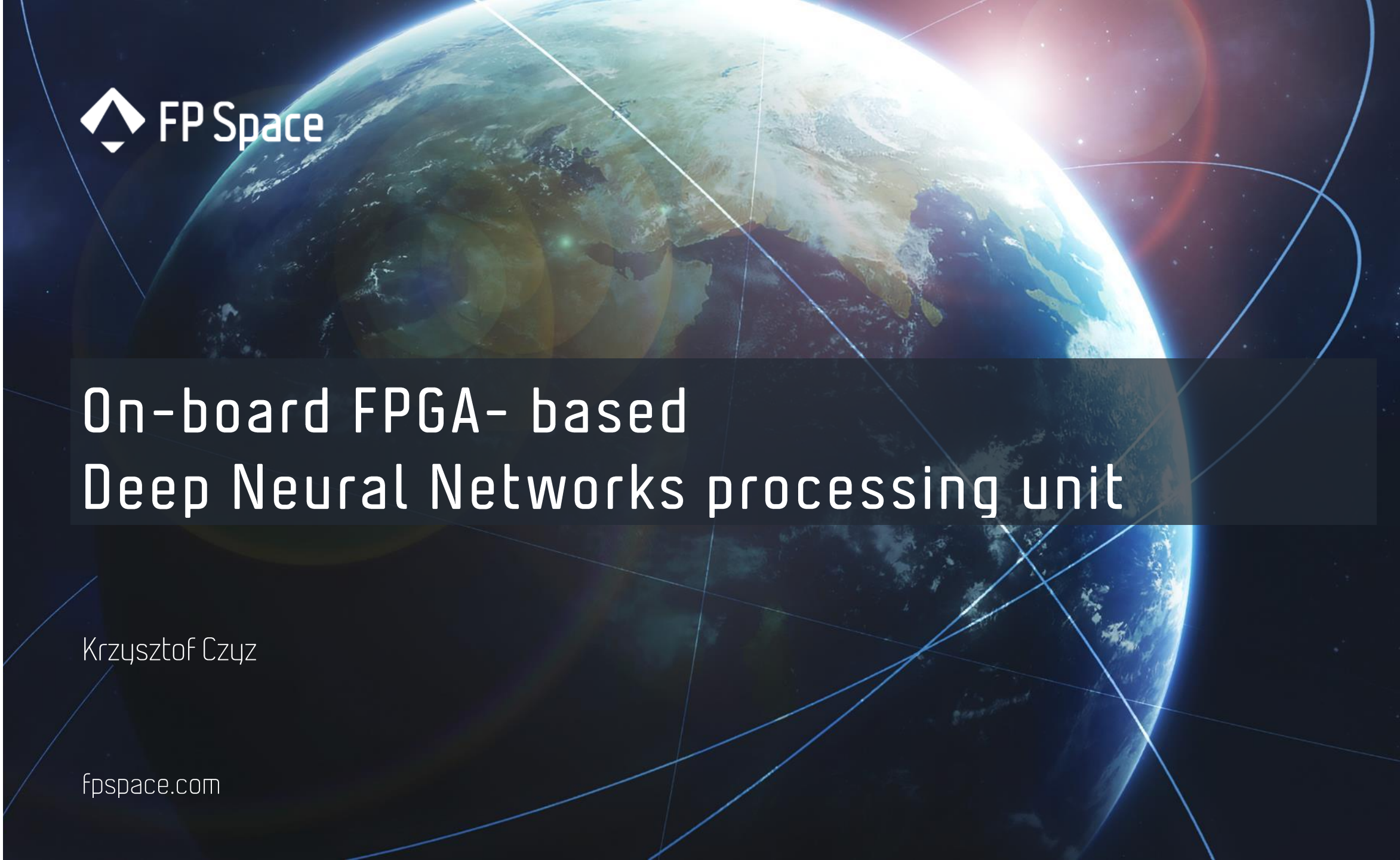


On-board FPGA- based Deep Neural Networks processing unit

Krzysztof Czyz

fpspace.com

2019

FP SPACE – WHO WE ARE AND WHAT WE DO…

Future Processing

Earth Observation
Application Development

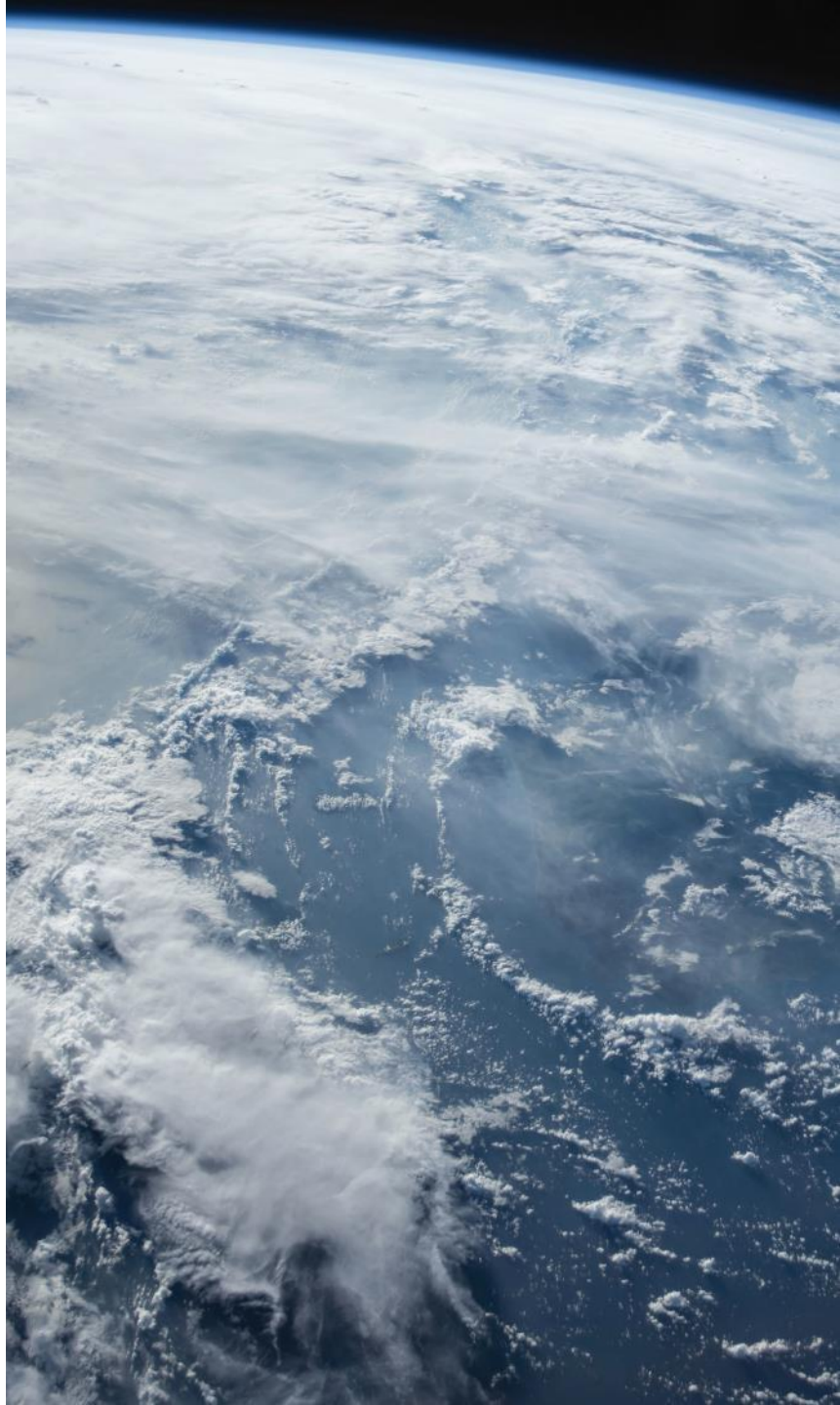
FP Space

KP LABS

Flight Software, Big Data
and Machine Learning

FPINSTRUMENTS

Electronics Design
and Production



INTUITION 1

HYPERCAM INSTRUMENT DEMOSTRATION MISSION

Hyperspectral instrument for Earth observation with increased spectral resolution enabling in orbit autonomous operation and automatic processing and classification of satellite data based on new algorithms for segmentation and classification of satellite images using deep convolutional networks.

Intuition 1

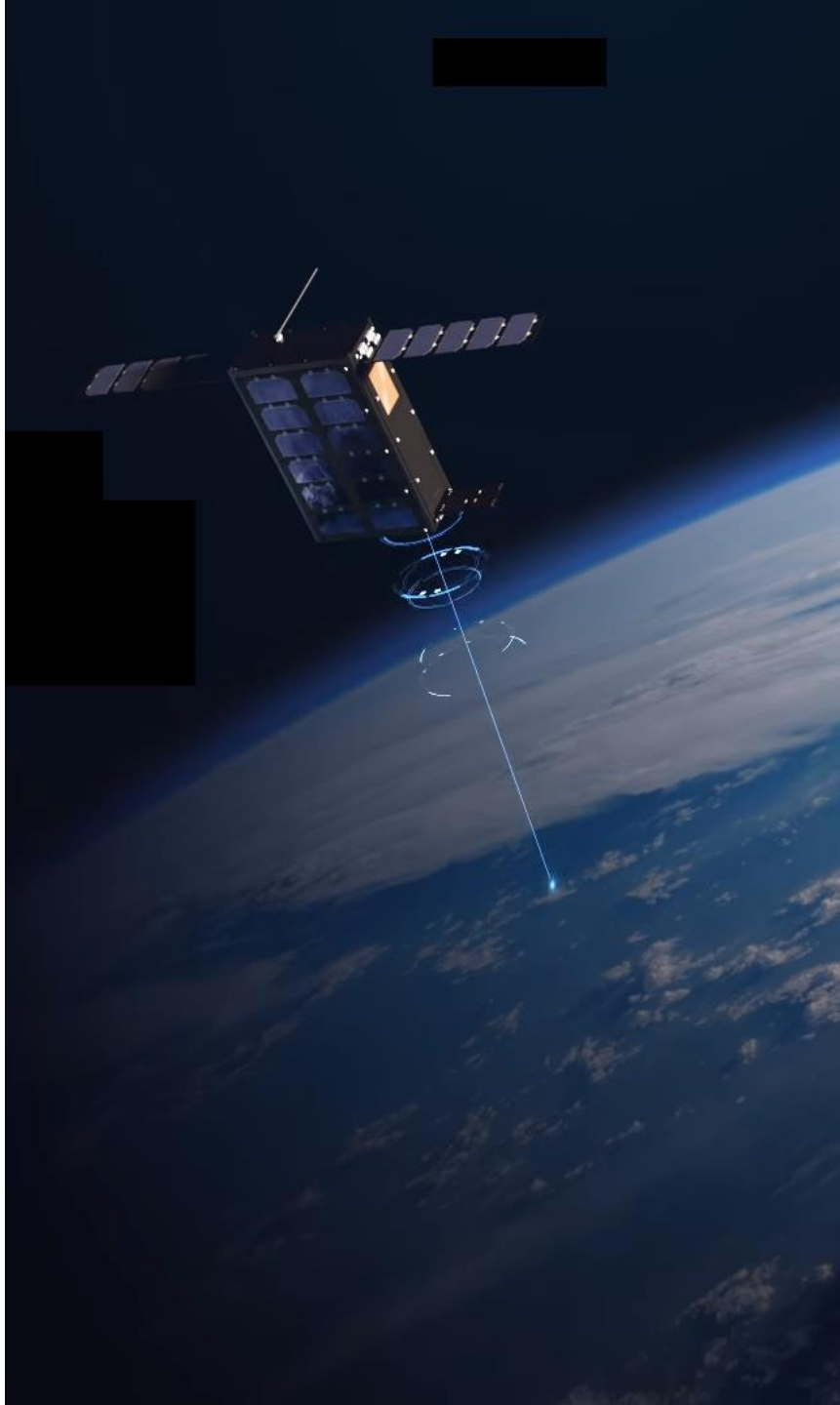
Hypercam instrument demostration mission

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

- CMOS sensor with embedded Fabry-Perot spectral filters
- ground sampling distance: 25m
- spectral range: 450 nm to 900 nm
- no. of spectral bands: >75
- swath: 15 km
- RAW data per second: 0.7 GiB

A high performance processing unit for Deep Learning Acceleration:

- Scalable architecture providing overall performance >1 TOPS
- On-board data segmentation and classification
- HSI data compression
- Dynamic schedule optimization



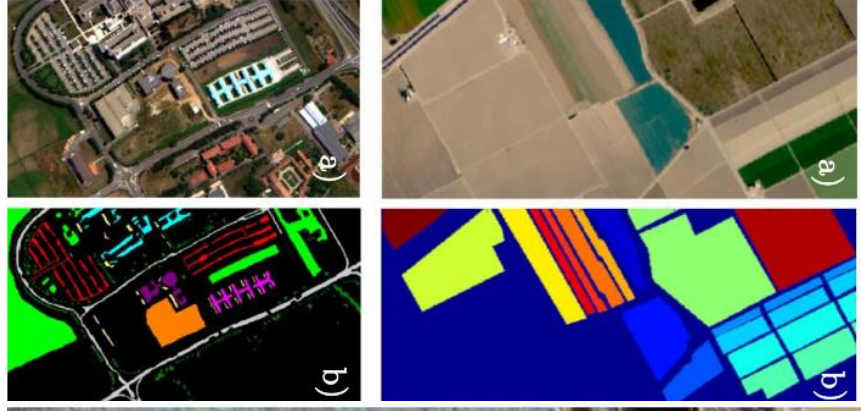
Intuition 1

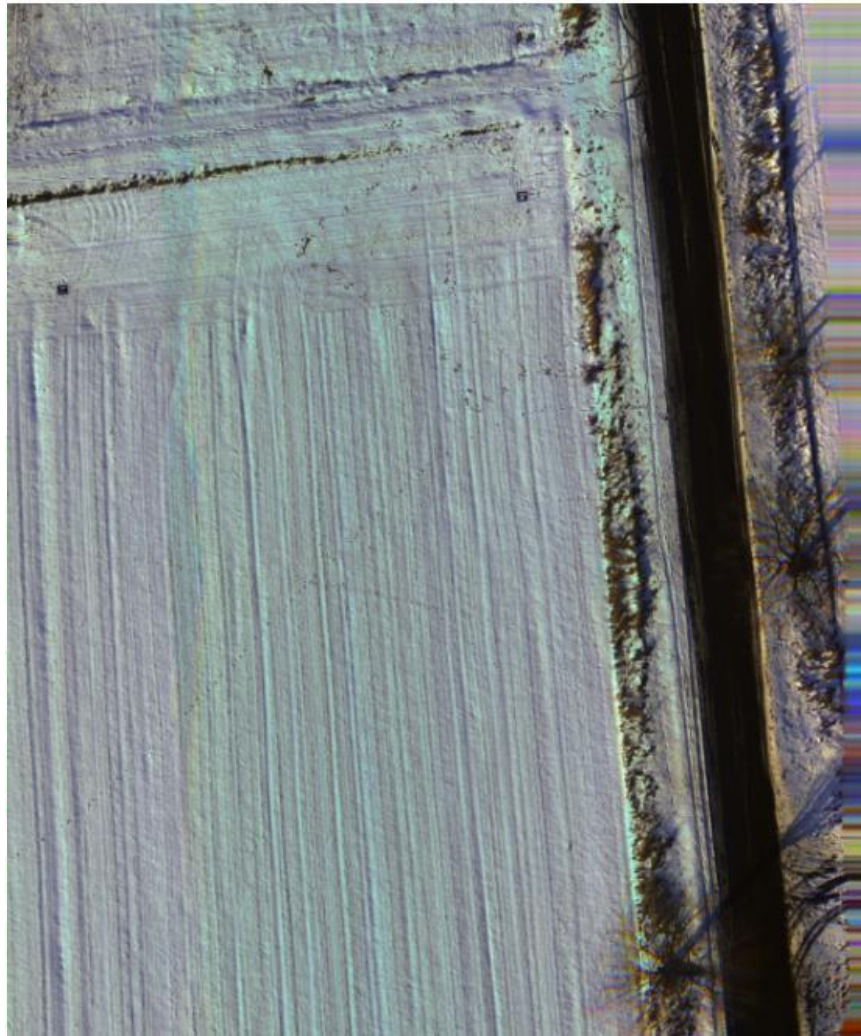
Hypercam instrument demostration mission

A algorithms for on-board HSI segmentation and classification:

- Segmentation of multiband and HSI
- Reduction of HSI by effectively select most-important bands
- Handling difficult HSI (data imbalance, representativeness)
- Automated design of deep neural nets for HSI segmentation
- HSI ground-truth data using UAV

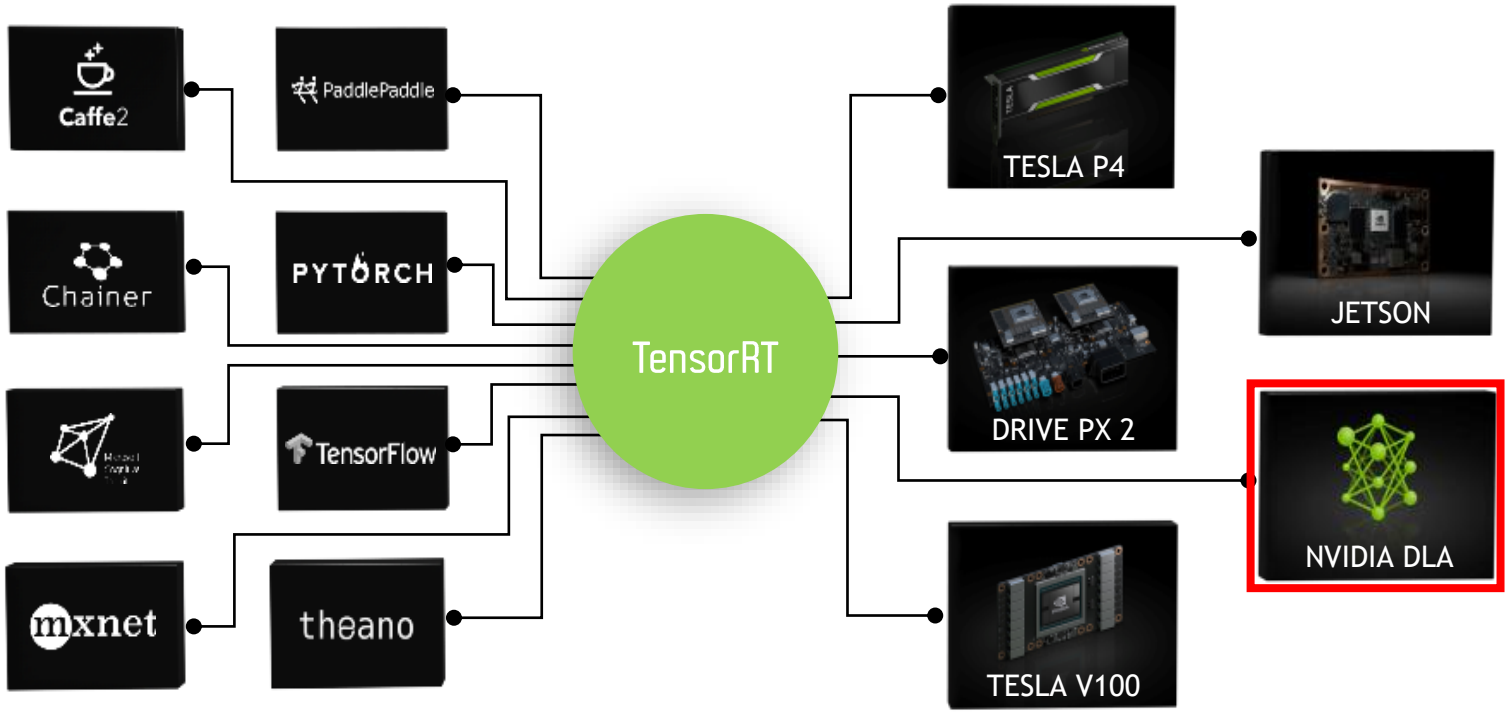
But, there is a price to be paid, these methods are computationally intensive and require supercomputing resources - that can be challenging, especially on-board of a space craft.





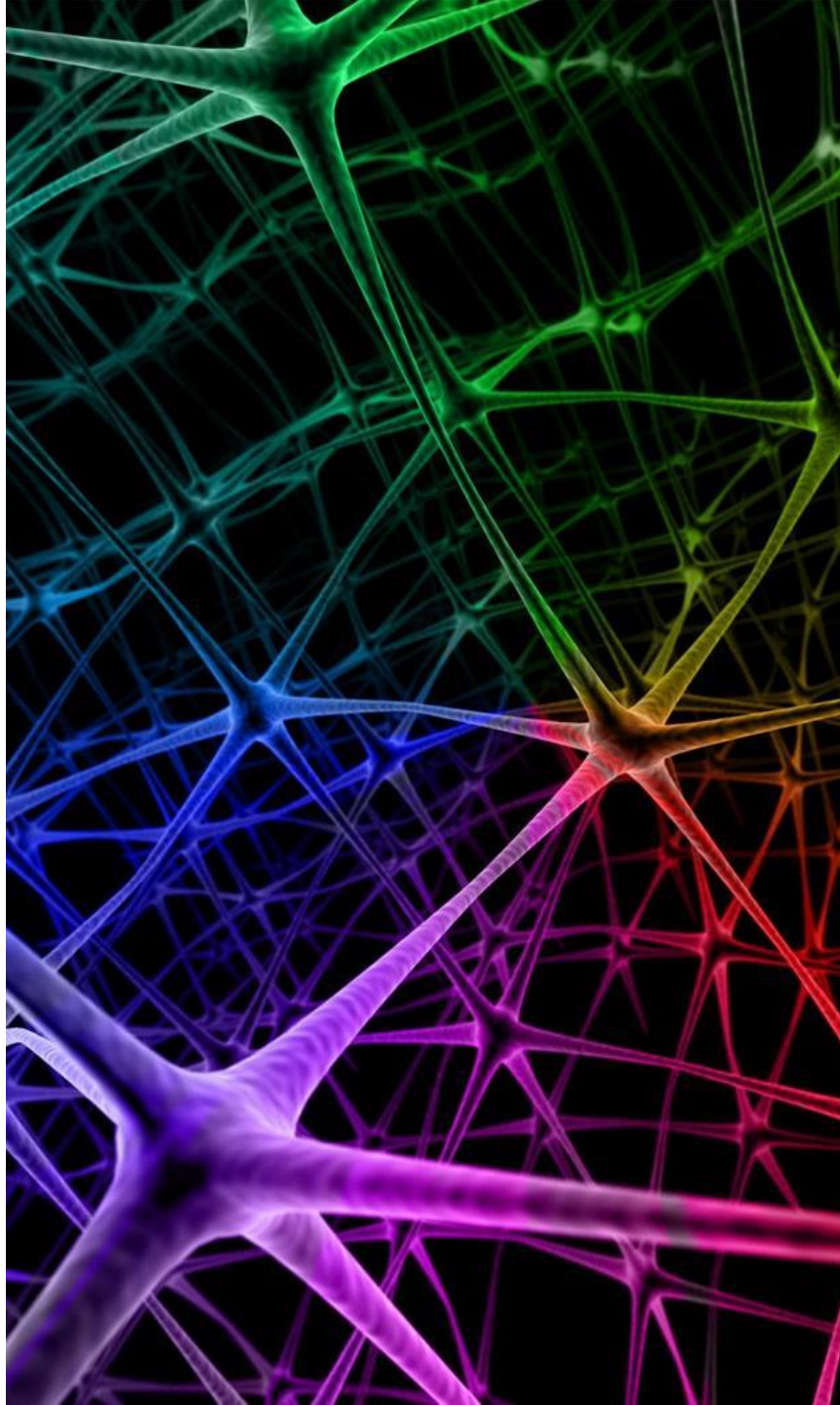
Frameworks and Hardware

ANYONE SEEN A SUPERCOMPUTER?



NVIDIA

source:AI At the Edge , Rammy Bahalul, nVidia



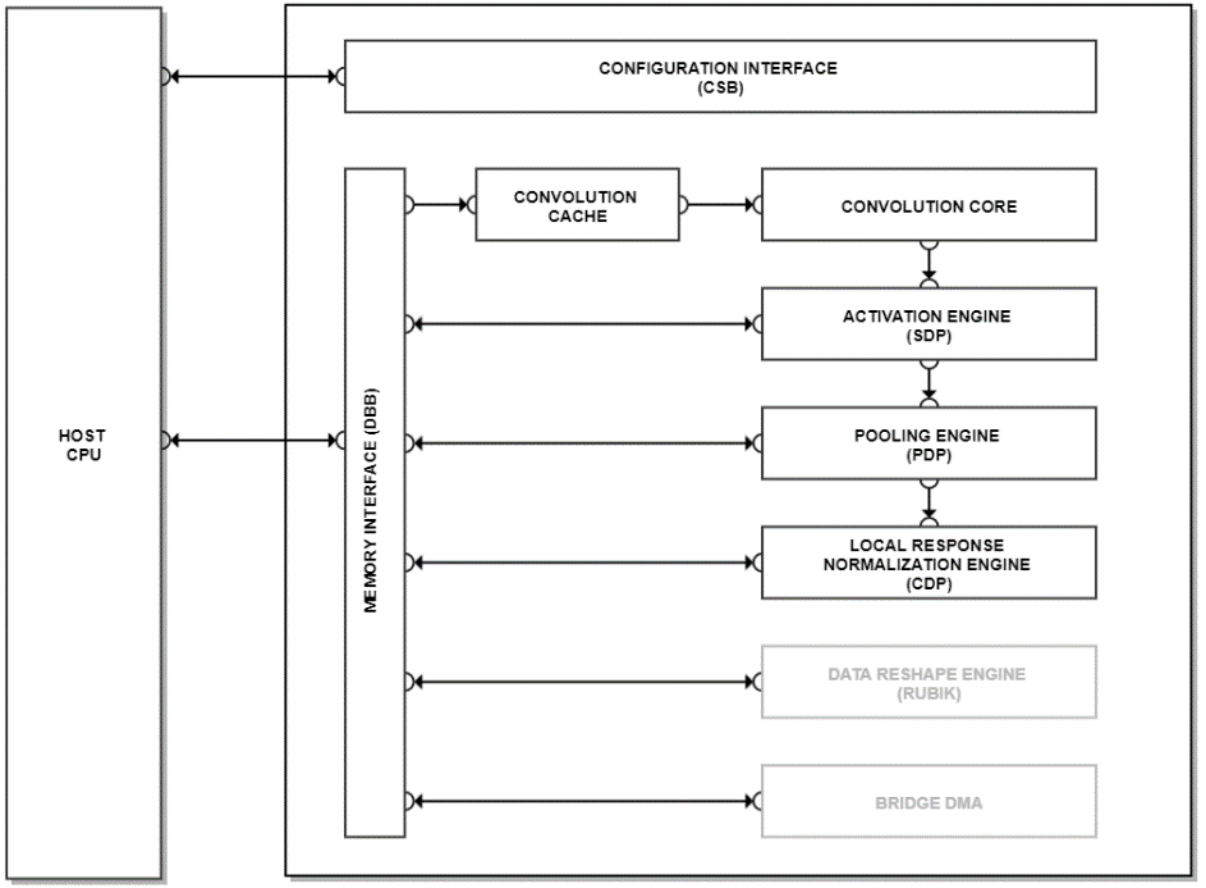
FP Space

NVIDIA Deep Learning Accelerator

WHAT IS IT? A HARDWARE PART

An open source neural network highly parallel deep learning accelerator created by NVIDIA providing:

- **Matrix Convolution Cores** (64-2048 cores)
  Direct/Winograd/Multi-Batch Convolution FC layers
- **Single Data Processors** (1-16 cores)
  Activation functions (from ReLU to non-linear)
  Bias addition
  Batch normalization
  Element-wise operations
- **Planar Data Processors** (1-16 cores)
  Pooling (min, max, average)
- **Channel Data Processors** (1-16 cores)
  Local normalization functions

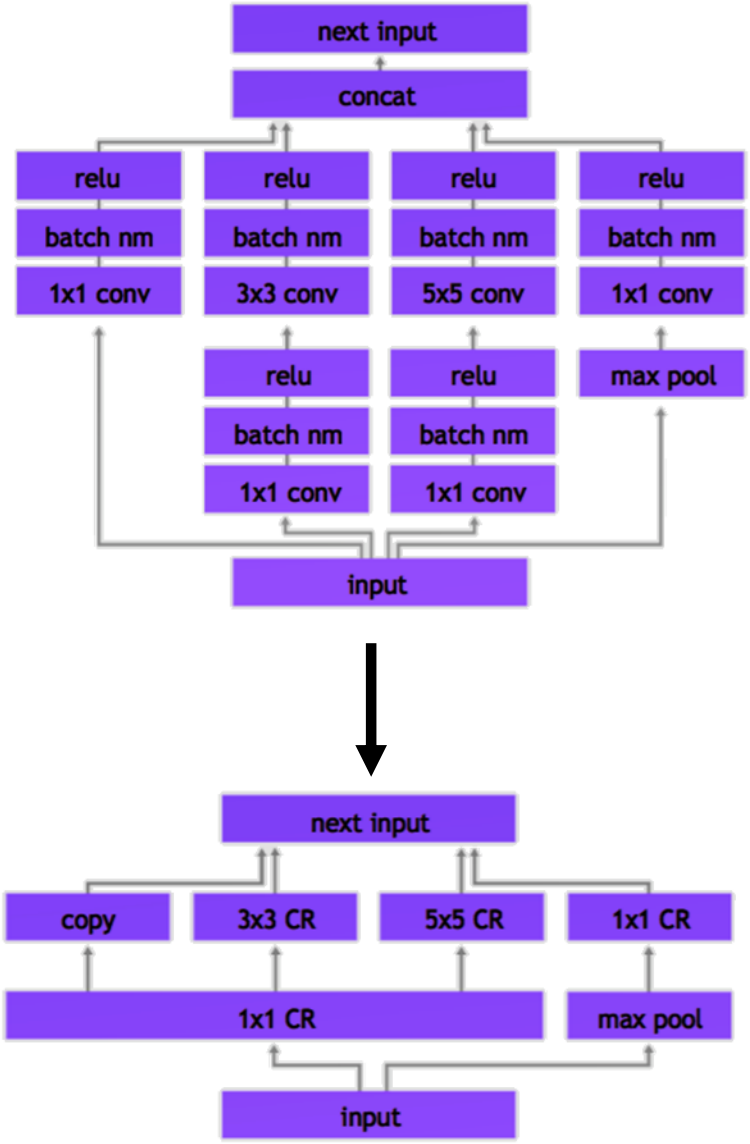


NVIDIA Deep Learning Accelerator

WHAT IS IT? A HARDWARE PART

An open source neural network highly parallel deep learning accelerator created by NVIDIA providing:

- **Matrix Convolution Cores** (64-2048 cores)
  Direct/Winograd/Multi-Batch Convolution FC layers
- **Single Data Processors** (1-16 cores)
  Activation functions (from ReLU to non-linear)
  Bias addition
  Batch normalization
  Element-wise operations
- **Planar Data Processors** (1-16 cores)
  Pooling (min, max, average)
- **Channel Data Processors** (1-16 cores)
  Local normalization functions



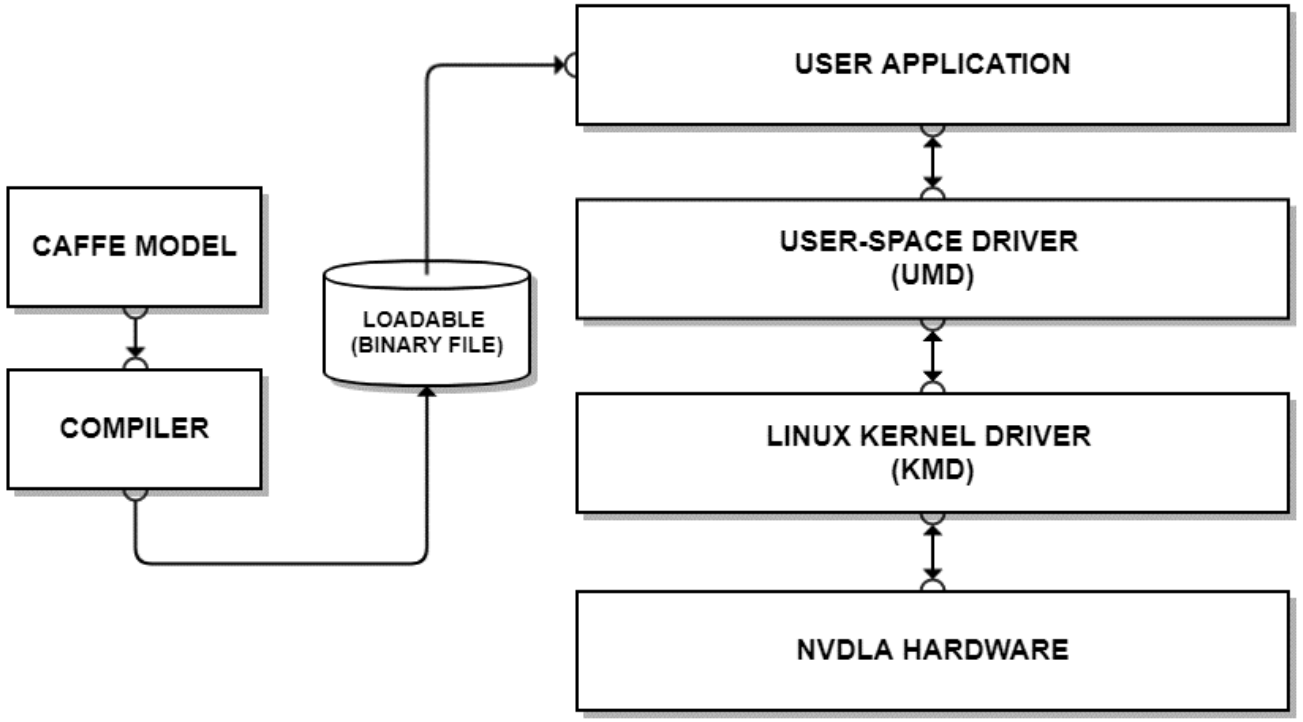
FP Space

NVIDIA DEEP LEARNING ACCELERATOR

WHAT IS IT? A SOFTWARE PART

nvDLA software ecosystem:

- **Caffe model**
  Use well established framework and deep learning data representations
- **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



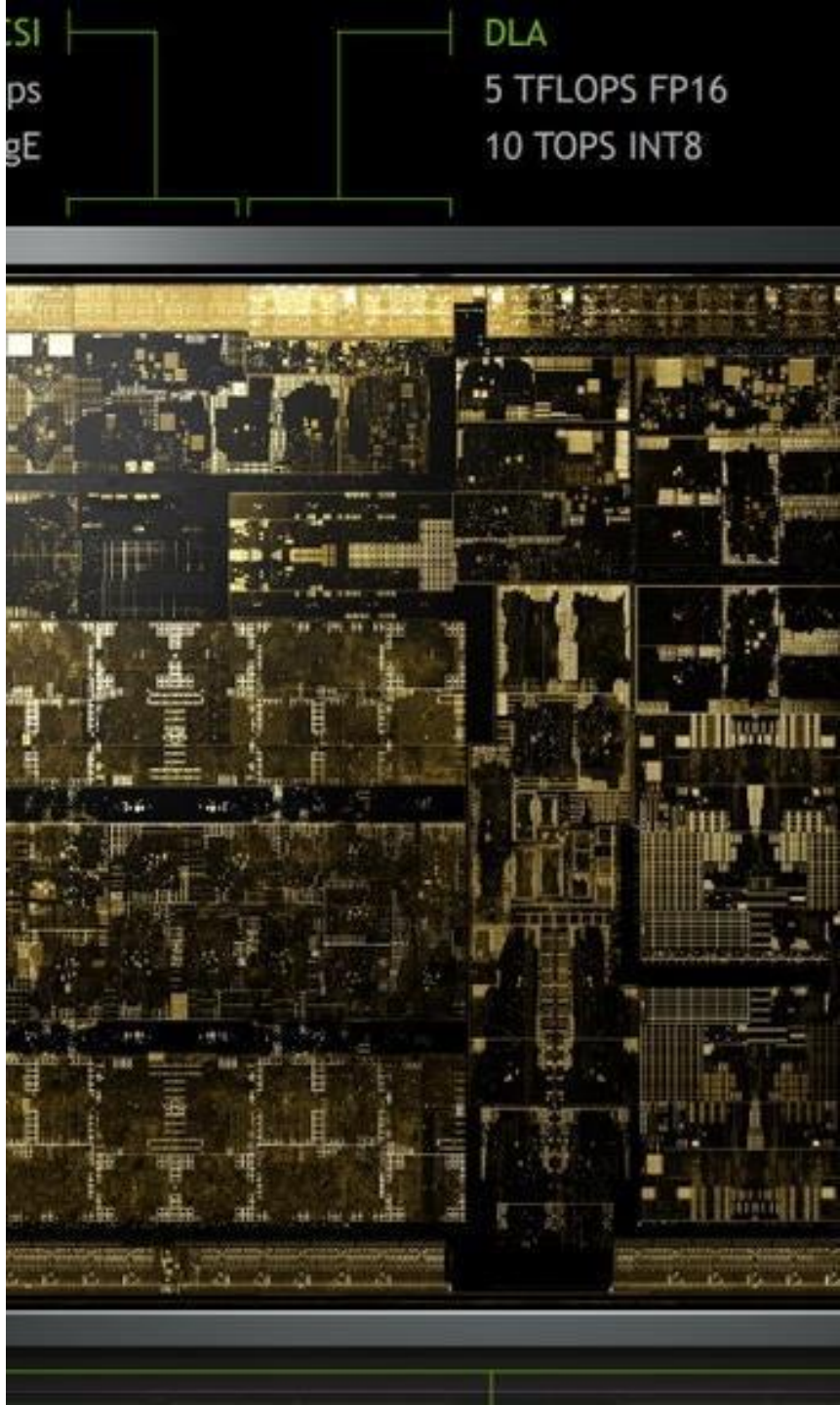
NVIDIA Deep Learning Accelerator

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 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)
- 2019 Q1
  release of DNN compiler for fixed point version of nvDLA



NVIDIA DEEP LEARNING ACCELERATOR

FPGA IMPLEMENTATION OF NVDLA

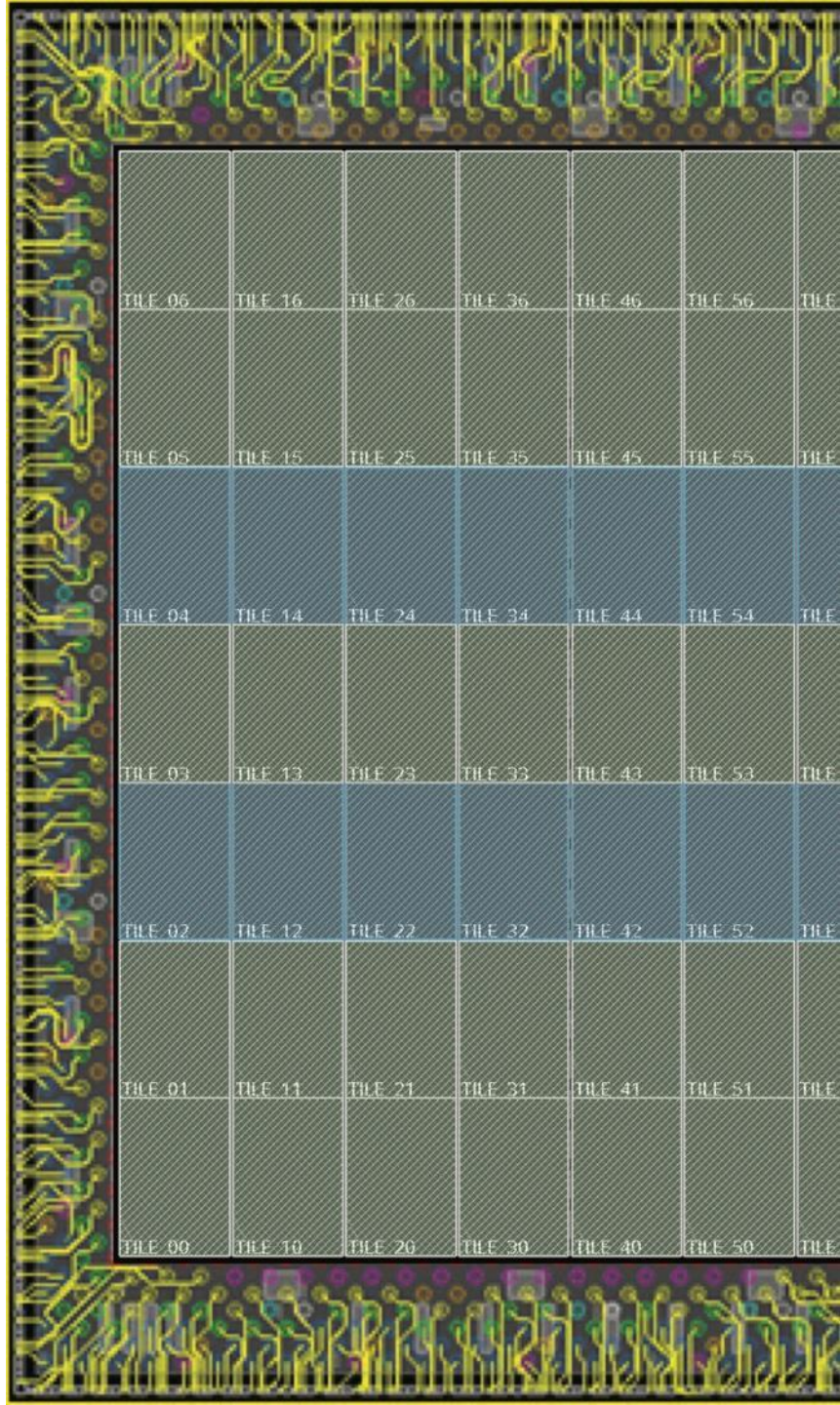
Goal: fit triple redundant nvDLA engine into FPGA

PROS:

- very good customization
- reasonable computing power of single nvDLA instance
  AlexNet 150 FPS (~33% of TX2i), GoogLeNet 75 FPS (38% of TX2i)
  for 512 MAC cores running @ 150 MHz
- FPGA mature technology in space

CONS:

- DNN compiler for fixed point nvDLA to be published in Q1 2019,
- the only reasonable data representation is fixed-point INT8
- higher power consumption than Jetson TX2i
- nvDLA is a complex IP, consuming a lot of FPGA area



NVIDIA DEEP LEARNING ACCELERATOR

FPGA IMPLEMENTATION OF NVDLA

Xilinx ZU15EG FPGA vs KU040

ZU15EG

- PMU with TMR, 2xR5 Cores with lock-step, 4xA53 Cores
- Good SEU performance, but exposed to SEL
- Good power efficiency

KU040

- No hardcoded cores (LEON seems to be a good choice)
- Higher sensitivity to SEU, but no SEL issues

NVIDIA DEEP LEARNING ACCELERATOR

FPGA OPTIMISATION ISSUES

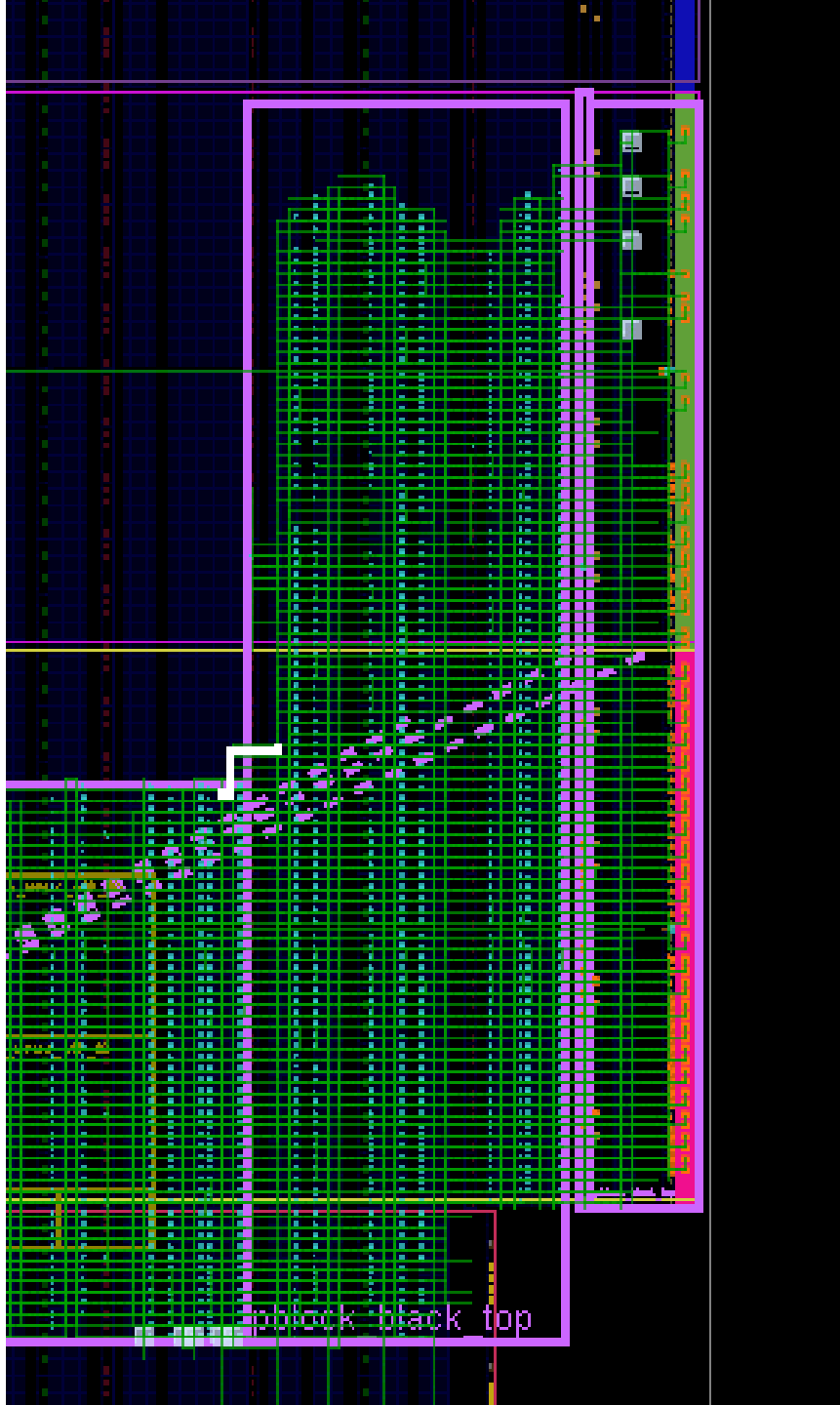
Goal:

fit 3x nv_11 in TMR mode on ZU15EG device

OR
fit 2x nv_06 in DMR mode on ZU15EG device
clock frequency at least 150 MHz

Possible solution:

- very high utilization of generic LUT is observed in the design, but almost no utilization of special function LUT like Distributed RAM (few hundreds) or Shift Register (literally three).
- a lot of FIFO queues are implemented as register chain – the reimplementation to use more efficient FPGA structures (BRAM, DRAM or SRL) can gain 3% up to 7% (depending on architecture model) extra space!



FP Space

NVIDIA DEEP LEARNING ACCELERATOR

FPGA RESOURCE UTILIZATION

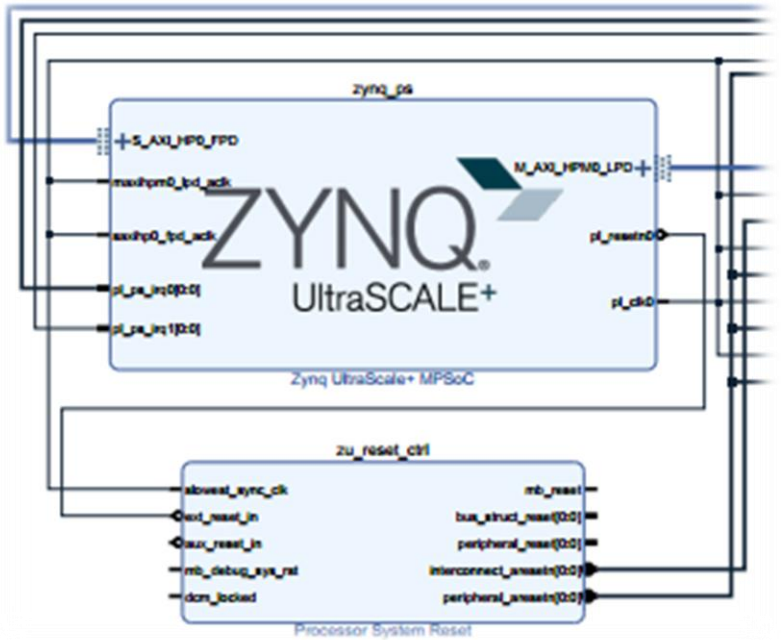
MODEL	Weight Compression	Winograd Convolution	Batch processing	SDP Non-linear	SDP Bias addition	SDP Batch norm.	SDP Element-wise	PDP Engine	CDP Engine	# of MAC	Conv. cache (kB)	SDP BS Throughput	SDP BN Throughput	SDP EW Throughput	PDP Throughput	CDP Throughput	Memory latency	Memory max. burst	Total LUTs	BRAM	DSP48	LUT usage (%, ZU15EG)	FF usage (%, ZU15EG)	LOCKSTEP possible	TMR possible	LUT usage (%, ZU17EG)	FF usage (%, ZU17EG)	LOCKSTEP possible	TMR possible
nv_01	■	■	■	■	■	■	■	■	■	512	512	16	16	4	8	8	1024	4	283495	209	239	**83%**	**32%**			**67%**	**26%**		
nv_02	■	■	■	■	■	■	■	■	■	512	512	4	4	1	2	2	1024	4	179960	197	83	**53%**	**25%**			**43%**	**20%**	✓	
nv_03			■	■	■	■	■	■	■	512	512	4	4	1	2	2	1024	4	171645	197	79	**50%**	**25%**			**41%**	**20%**	✓	
nv_04			■		■	■		■	■	512	512	4	4	-	2	2	1024	4	156351	194	65	**46%**	**22%**			**37%**	**18%**	✓	
nv_05			■		■	■		■	■	512	128	4	4	-	2	2	1024	4	155878	194	65	**46%**	**22%**			**37%**	**18%**	✓	
nv_06			■		■	■		■	■	512	128	1	1	-	1	1	1024	4	143205	194	40	**42%**	**21%**	✓		**34%**	**17%**	✓	
nv_07			■		■	■		■		512	128	1	1	-	1	1	1024	4	124613	192	35	**37%**	**18%**	✓		**29%**	**14%**	✓	✓
nv_08			■		■	■		■		512	128	1	1	-	1	1	64	4	124620	187	35	**37%**	**18%**	✓		**29%**	**14%**	✓	✓
nv_09			■		■	■		■		512	128	1	1	-	1	1	64	1	124087	187	35	**36%**	**18%**	✓		**29%**	**14%**	✓	✓
nv_10			■		■	■		■		512	32	1	1	-	1	1	64	1	118654	123	35	**35%**	**16%**	✓		**28%**	**13%**	✓	✓
nv_11			■		■	■				512	32	1	1	-	1	1	64	1	101371	119	32	**30%**	**12%**	✓	✓	**24%**	**10%**	✓	✓
nv_12			■			■				512	32	1	1	-	1	1	64	1	95772	117	30	**28%**	**11%**	✓	✓	**23%**	**9%**	✓	✓
nv_13						■				512	32	1	1	-	1	1	64	1	95541	117	30	**28%**	**11%**	✓	✓	**23%**	**9%**	✓	✓

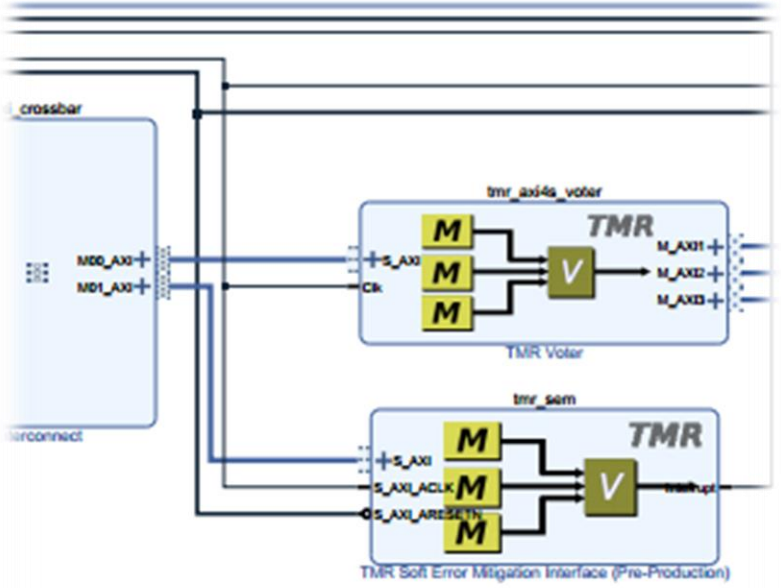
NVIDIA DEEP LEARNING ACCELERATOR

TRIPLE MODULAR REDUNDANCY

- NVDLA is a complex IP, consuming a lot of physical area and processing a lot of data for extended period.

- This increases the likelihood of Single Event Upset (SEU) which can lead to corrupted output data or functional failure.

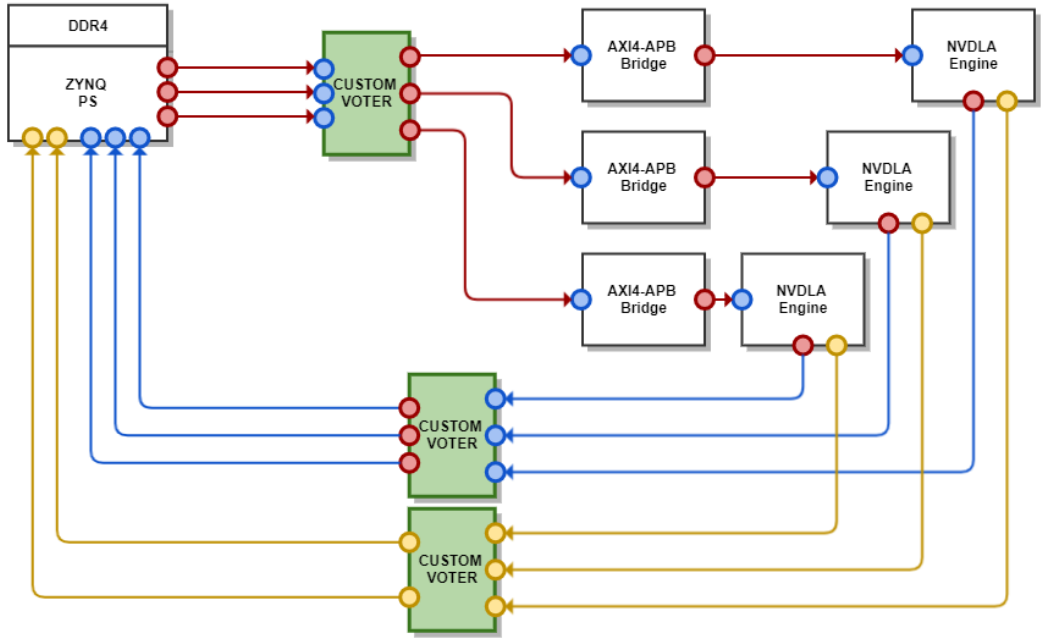




NVIDIA DEEP LEARNING ACCELERATOR

TRIPLE MODULAR REDUNDANCY
MULTIPLE MEMORY INTERFACES AND ADVANCED TMR VOTER

Block TMR

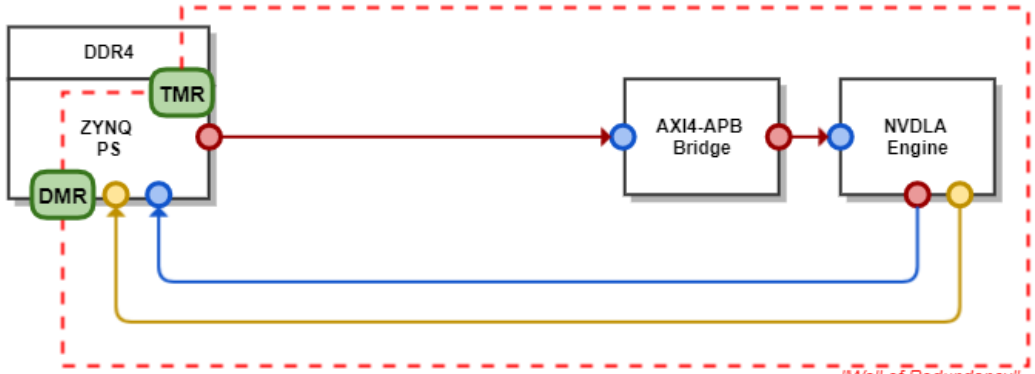


Pros:

- Error detection mechanism (well, FPGA part at least)
- Input/output data copys reside in DDR4

Cons:

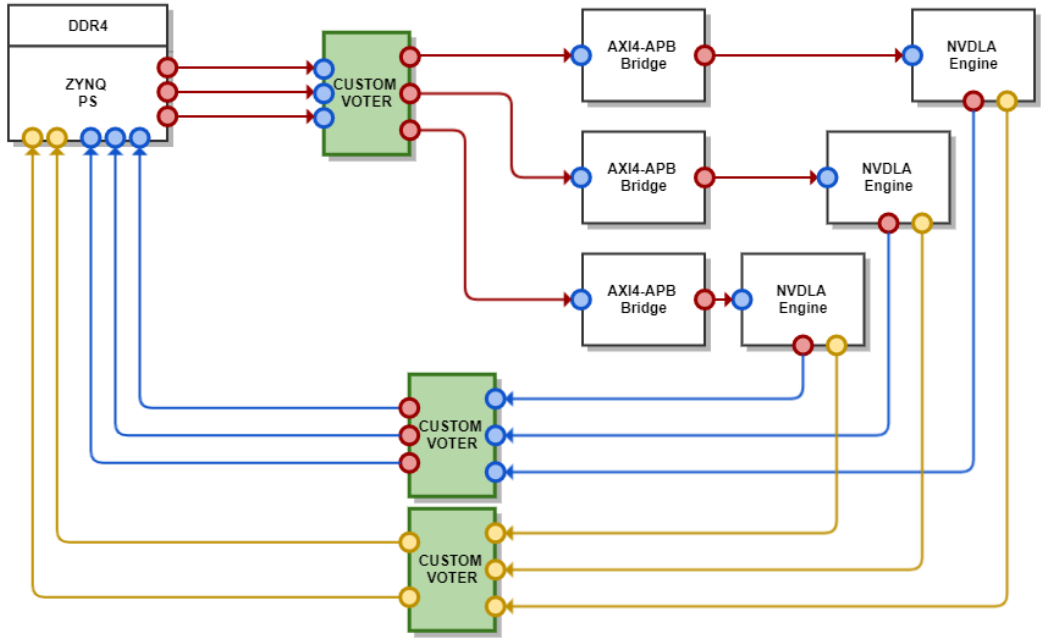
- Does not correct nvDLA operation (only partial results)
- Complicated TMR design – AXI4 may have distinct behaviour on each channel due to accessing the same DDR4 controller
- DDR4 bandwidth and latencies might be an issue (concurrent)
- DDR4 increased usage (3x)



NVIDIA DEEP LEARNING ACCELERATOR

TRIPLE MODULAR REDUNDANCY
MULTIPLE MEMORY INTERFACES AND ADVANCED TMR VOTER

Distributed TMR (to be implemented?)

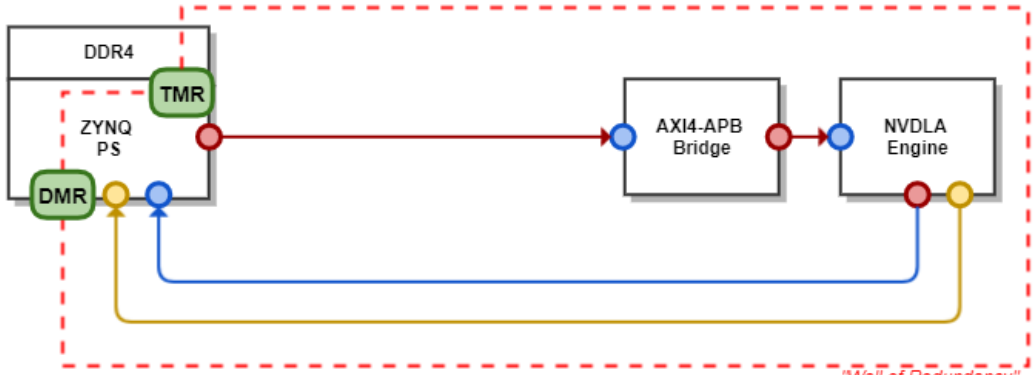


Pros:

- Implemented for individual Processors/Cores of nvDLA
- Improved resistance to SEE
- Masks configuration bit errors

Cons:

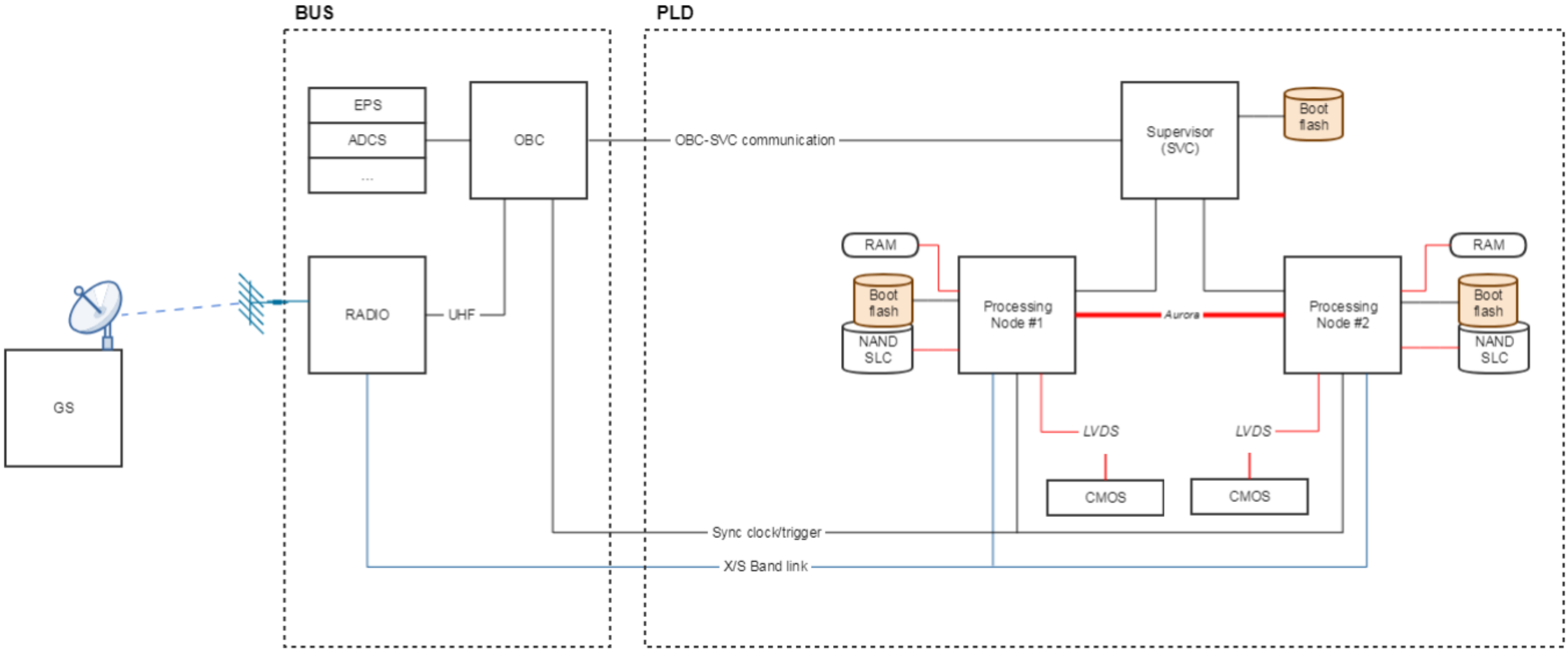
- Complicated design
- High resource utilisation



FP Space

HyperCAM payload

BLOCK DIAGRAM

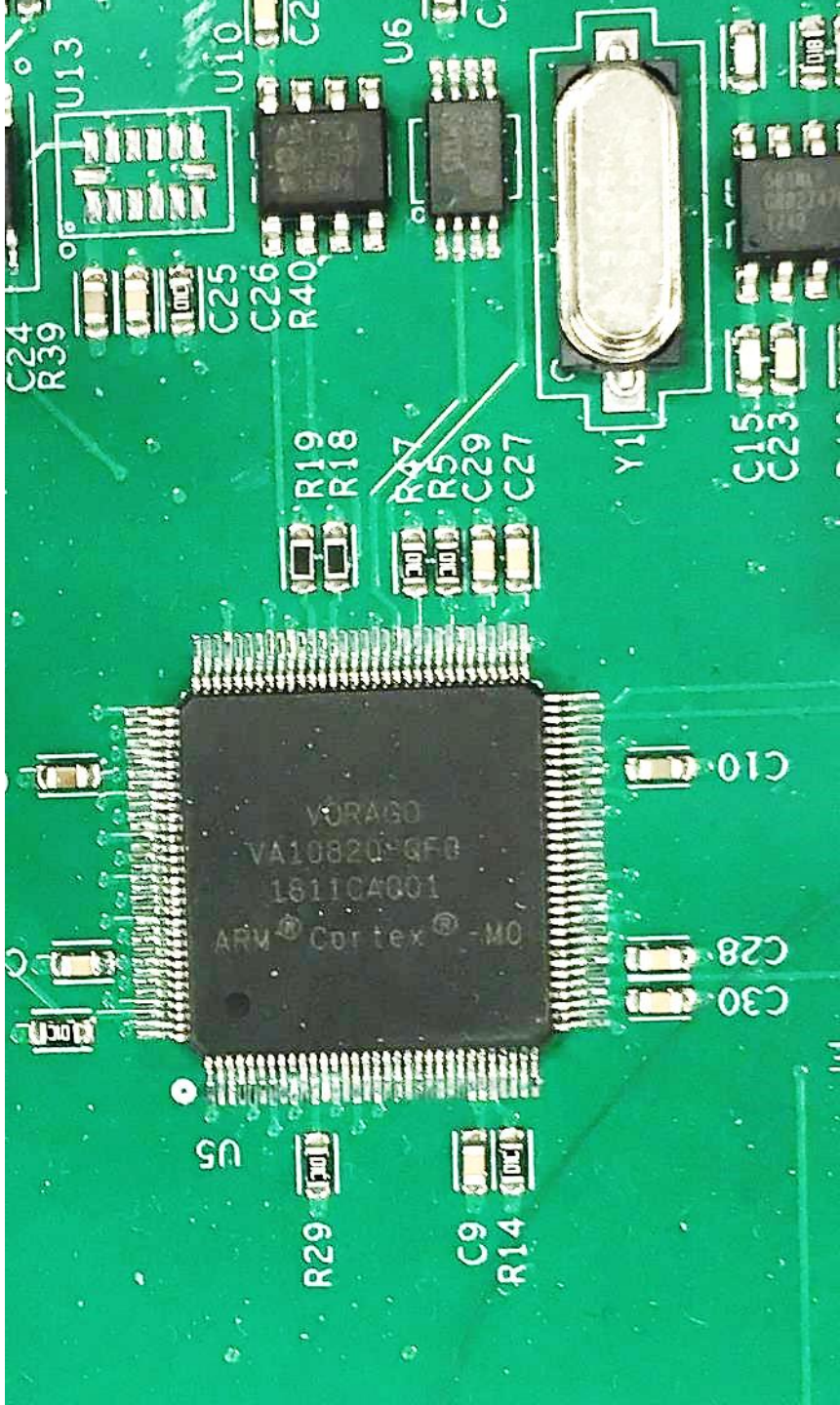


Supervisory Circuit

VORAGO RADHARD MCU

Supervisory circuit / payload controller functions:

- Communication bridge between PLD and OBC
- Task scheduling for computing unit
- Power, thermal and fault management
- Error logging
- Firmware management
- Safe mode



COGNITION

ON-BOARD PROCESSING DEMONSTRATION…

A rideshare mission, initial on orbit tests of DNN processing unit.

- Architecture providing performance >200 GOPS
- 8 GiB of RAM memory
- Panchromatic camera
- Power consumption < 20 W
- Size U1

… to be launched in early 2020

Intuition 1

Hypercam instrument demostration mission

A high performance processing unit dedicated for segmentation of the hyperspectral images in orbit:

- Scalable architecture providing overall performance >1 TOPS
- 16 GiB of RAM memory
- Power consumption < 30 W
- Flash based data storage 0.5 TiB
- Triple Modular Redundancy

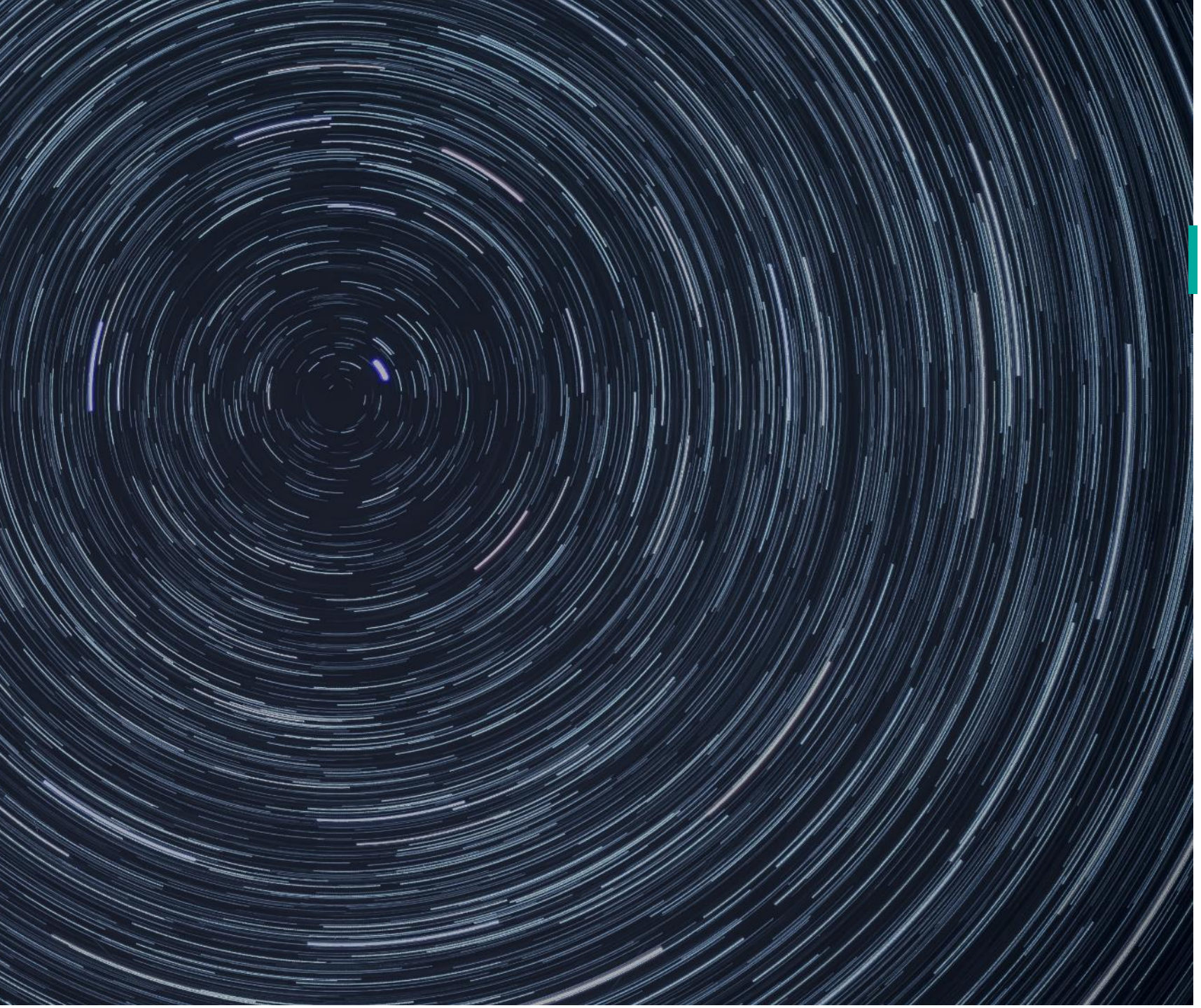
...supercomputing powers for a small sat

INTUITION 1

HYPERCAM INSTRUMENT DEMOSTRATION MISSION

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.

... to be launched in 2023



FP Space

THANK YOU

Krzysztof Czyż,
FP Instruments sp. z o.o.
44-100 Gliwice,
kczyz@fp-instruments.com

fpspace.com