GSTP-MAKE

DeepCube

Final Report

30/05/2022

GSTP- Make: DEEPCUBE - Objectives

Segment 3: Implementation of National Priorities

- The goal of the project is to develop a service as a product to simplify fitting DNN (Deep Learning Neural Networks) in on board HW to make better use of AI on space missions.

- The objective of the activity is to acquire wider experience on DNN simplification strategies to provide services to the future customers.

- A DNN simplification software will be developed as a tool to support AGENIUM business strategy. AGENIUM Space' DeepCube software will simplify DNN models to reduce resources required for inference execution, considering the capabilities of existing HW (COTS and space-qualified HW).

GSTP- Make: DEEPCUBE - Objectives

- The role of the service is to support data processing engineers in reducing HW resources requirement of powerful DNN for image analysis, to be executed on-board (priority) but also on low performance on-ground HW:
  - 1)User trains his own neural network (ex. ensemble/very big network with millions of free parameters)
  - 2) User exports his model to a standard framework like TensorFlow,
  - 3) AGENIUM Space framework takes the model as input plus some labeled data to run inference on it and proposes several networks compatible with HW solutions.
  - 4) AGENIUM Space framework provide the best simplified model and uses third-party software for the specific code required to run inference on the selected HW.
  - 5) The produced model (if no code production is used or HW is not supported) can be used in third party SW for specific HW code production since we rely on known data structures embedded into compatible formats for most industrial formats
    - **ONNX, TensorFlow, Caffe…**

WBS

- DeepCube
  - **WP 1000 Management**
    - WP 1100 Project Management (including Quality)
    - WP 1200 Commercialization
  - **WP 2000 Framework for DNN Simplification V1**
    - WP 2100 Devices selection V1&V2
    - WP 2200 Use-cases selection V1&V2
    - WP 2 300 Implementation Use-Cases V1
  - **WP 3000 HW Testing**
    - WP 3100 SW4HW preparation V1
    - WP 3200 Test V1
    - WP 3300 Improving Testing Environment
    - WP 3400 Test V2
  - **WP 4000 Framework for DNN Simplification V2**
    - WP 4100 Simplification Rules updating
    - WP 4200 Implementation Use-Cases V2
  - **WP 5000 Tool Packaging**
    - WP 5100 Tool Packaging Environment V1
    - WP 5200 Tool updating V2

Level 1

Level 2

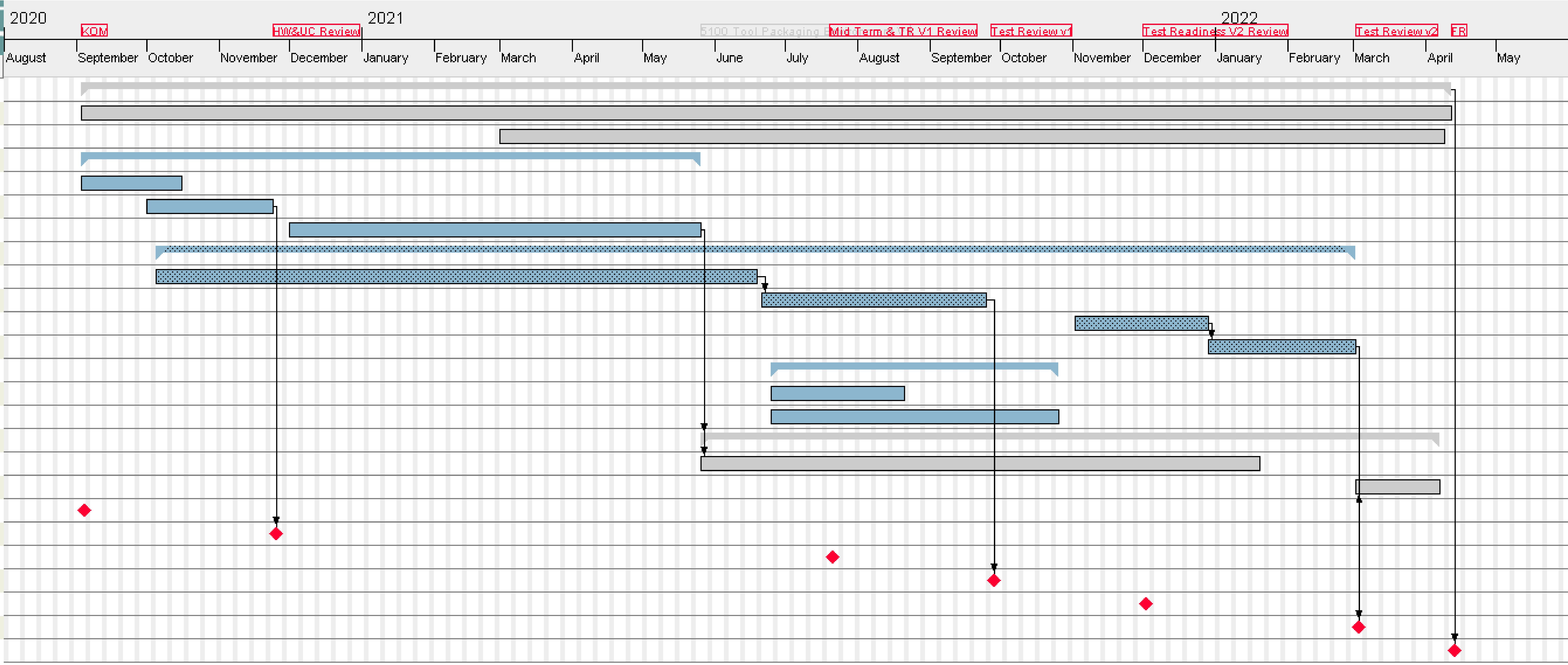
Final planning

- KO 07/07/2020
- PM1- 03/09 Starting activities
- PM2 – 13/10/2020: UC&devices analysis
- HW & UC Review merged at KO+2 : 24/11/2020 (~1 month late)
- …
- PM9 – 25/06/2021 : End WP 2300 Implementation V1 & WP 3100 HW Testing preparation + Test Plan delivery
- MTR – 20/07/2021 : Progress 3200 Test V1 & Activities planned for WP 4100 Simplification V2 & WP5100 Tool Packaging
- Test Review V1: End WP 3200 (Test HW V1) & Progress WP 4100 & 4200 & 5000
- End WP 4200 Implementation of UC V2+Last Tests HW V1.
- Test Readiness V2

+2 month against initial planning:
+1 UC&SWR +1TestReview V1

Final Review (MS-4) : 30 may 2022

Name	Begin date	End date
WP-1000 Management	9/3/20	4/11/22
1100 Project Mana...	9/3/20	4/11/22
1200 Commercialis...	3/1/21	4/8/22
WP-2 Framework for D...	9/3/20	5/25/21
2100 Device Selecti...	9/3/20	10/15/20
2200 Use-Cases Sel...	10/1/20	11/23/20
2300 Implementati...	12/1/20	5/25/21
WP-3 HW testing	10/5/20	3/1/22
3100 SW4HW prep...	10/5/20	6/18/21
3200 Test V1	6/21/21	9/24/21
3300 Improving Tes...	11/2/21	12/28/21
3400 Test V2	12/29/21	3/1/22
WP-4 Framework for D...	6/25/21	10/25/21
4100 Simplification...	6/25/21	8/20/21
4200 Implementaio...	6/25/21	10/25/21
WP-5 Tool packaging	5/26/21	4/6/22
5100 Tool Packagin...	5/26/21	1/19/22
5200 Tool Updating...	3/2/22	4/6/22
KOM	9/3/20	9/3/20
HW&UC Review	11/24/20	11/24/20
Mid Term & TR V1 Revi...	7/20/21	7/20/21
Test Review v1	9/27/21	9/27/21
Test Readiness V2 Revi...	12/1/21	12/1/21
Test Review v2	3/2/22	3/2/22
FR	4/12/22	4/12/22



Reviews & meetings

Milestone	**Title**	**Objectives**	**HW/SW Model associated to verif. or qualification**	**Decision Point(s)**
KO	Kick-off	Validate planning, project content, objectives and first activities.	KO presentation	Initiate the contract
HWR&UCR	HW & Use-cases Review	Select HW for testing (associated SW is also considered) and define use cases constraints (simplification requirements and constraints). Use cases to be implemented in the project in versions V1&V2 are defined and analyzed in WP 2100 (for HW aspects) and WP 2200 (for user/applications aspect). This review allows to discuss them with ESA to consolidate the priorities defined.	Hardware Analysis Report. Use cases report for V1&V2	Select HW for testing in V1 &V2 Validate Use cases to be managed in V1&V2 Initiate development of Use Cases for V1**(KO+2)**
MTR & TRR V1	Mid Term & Test Readiness Review V1	Assess the project progress and review Test Plan for testing V1 on HW.	Planning and progress reports & TP	Evaluate project progress and TP suitability (KO+9: July/2021 but TP June/2021)
TRV1	Test Review V1	Review of the testing results of the simplification Framework v1	SW V1 (Simplification Framework V1)& Test Report	Validate simplification capabilities tested and update target performances of V2(KO+12)
TRRV2	Test Readiness Review V2	Review Test Plan for testing V2 on HW	TP (output WP 3300)	Evaluate project progress and TP suitability (KO+15)
PDR/TR-V2	**Test Review V2/Preliminay Design Review**	**Assessment of the simplification capabilities provided by the developed simplification framework after testing on HW. Since the TRL of the service and simplification methods are considered as TRL 5, this review is considered as a PDR.**	**Simplification Framework V2 (WP 4200) and HW testing results (WP 3400)**	**Validate the simplification performances of the target product (KO+17)**
FR	Final Review	Presentation an overview of the activities carried out during the project.		Confirm that all outputs, deliverables and work have been successfully completed/delivered to the satisfaction of the Agency and as defined in the Contract. (KO+18)

Deliverables – Confidential – AGENIUM Space's property

Document Reference	Deliverable Document	Document content	WP	Milestone
UCR	Use Cases Report V1&V2	Use cases description for versions V1&V2 of the simplification SW	2100	Use Cases & HW Review
HWR	Hardware analysis report	Report including the analysis of the devices considered (spatial HW and alternatives to rad hard/tol devices) and rational of selected ones	2200	Use Cases & HW Review
SFW V1	Simplification V1	Description of simplification SW implemented	2300	End WP 2300
SFW V2	Simplification V2	Description of simplification SW implemented	4200	End WP 4200
TP V1	Test plan for Simplification Framework V1.	Test plan V1	3000 (3100)	Test Review v1. End WP 3200 – 24/06/2021
TR V1	Test report V1.	Test report including analysis of results, code conversion limitations, power budgets for different networks and hardware types, and recommendations on cubesats architectures for different use-cases, for V1	3000 (3200)	Test Review v1. End WP 3200
TP V2	Test plan V2	Test plan for Simplification Framework V2.	3000 (3300)	Test Review v2 (PDR).
TR V2	Test report V2	Test report for V2. It includes the Worst-Case Analysis.	3000 (3400)	Test Review v2 (PDR).
UM	User Manual	User Manual of the simplification software (V1 & V2). Updated after for each version after each test phase.	5000	End WP 3200 & End WP 3400
FDP	Final Data Package	Comprise a compilation of the most recent versions of all deliverable documents including Final Report and the Website Article Template. Include an index document with links to the different document files contained therein.	1000	Final Review
SRF	Software Re-used File	Analysis of existing software intended to be reused.	2200	End WP 2200 -> PM4 (14/01/2021) Updated MTR
PR-#	Progress reports	Description of activities performed during the period (detailed description is provided in Part 4)	1000	Monthly (Replaced by PM)

Training & Distillation
WP2000 & WP4000

Training & Distillation - Use cases tested (WP2200 & WP 4200)

Use case	DNN method	Input data	Project version
Cloud coverage	Segmentation or classification	Aerosol-RGB-NIR (Sentinel-2 10m, Landsat-8 30m)	**v1**
Deforestation	Change detection	RGB+NIR (Sentinel-2 10m)	**v1**
Vessel classification	Classification and bounding boxes detection	RGB, Airbus Dataset Kaggle.	**v1**
~~Fire detection~~	Segmentation	Multispectral (Landsat-8 30m)	v2
Snow vs clouds	Segmentation	RGB (Sentinel-2 10m, Landsat-8 30m)	v2

AGENIUM SPACE

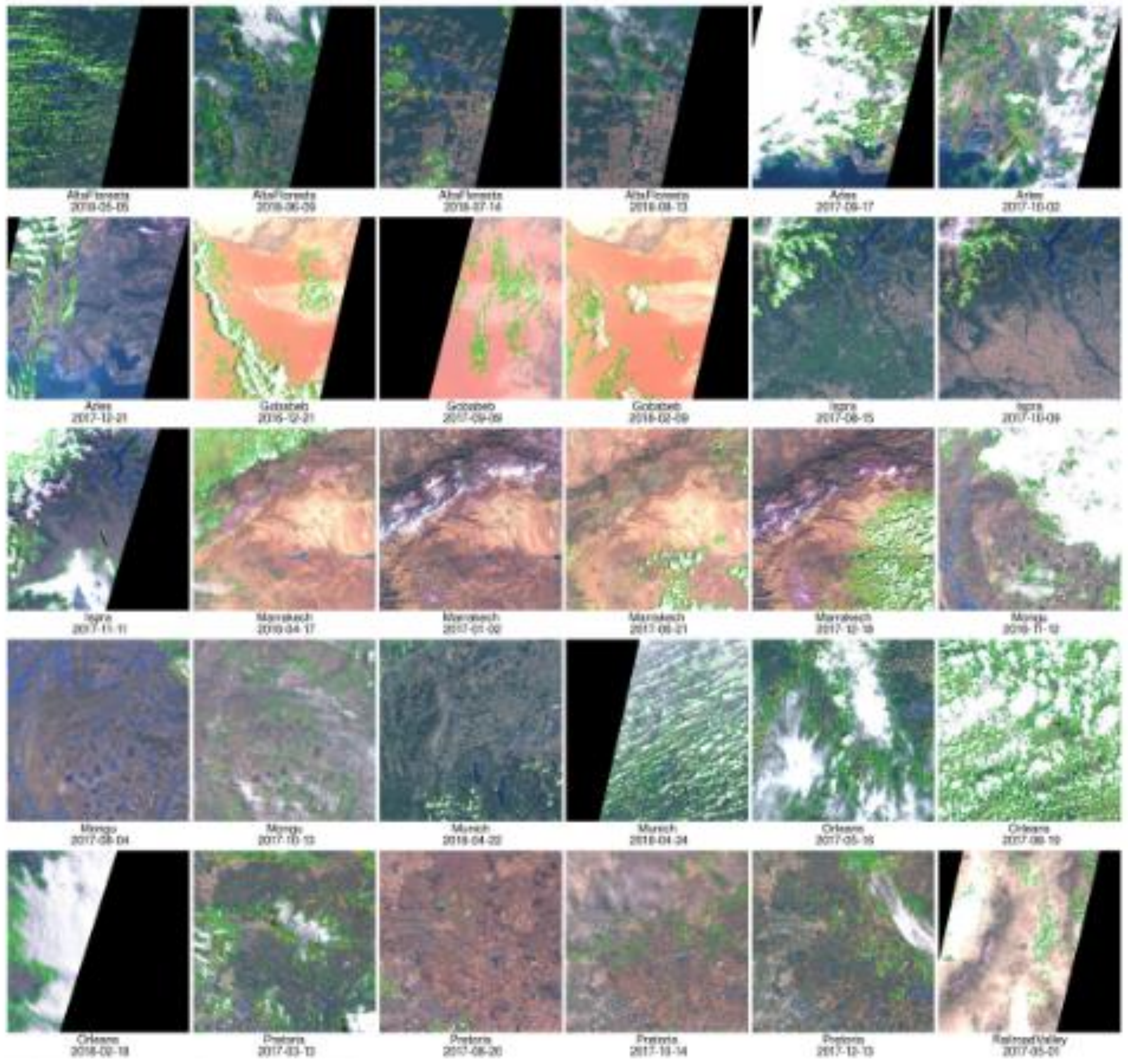
Training & Distillation – UC1: Cloud Segmentation (WP2200 & WP 4200)

Selected data base:

- 95-cloud : 95 Landsat-8 scenes (extension of 38-cloud DB)
- ALCD DB : 31 Sentinel-2 scenes
- Sentinel-2 Cloud Mask Catalogue : 513 1022-by-1022 pix. subscenes from Sentinel-2
- SPARCS : 80 1000-by-1000 pix. subscenes from Landsat-8

Selected radiometric bands:

- Aerosol band (433-453nm): very useful for cloud detection
- RGB bands: most conventional bands
- NIR band (784-900nm): help discrimination with water



AGENIUM SPACE

Training & Distillation – UC1: Cloud Segmentation (WP2200 & WP 4200)

Results :

- Distillations done with **1M parameters** for the ***students*** on different architectures *(PSPNet, Unet & FPN)*
- Architecture : Unet with encoder EfficientNet-B5 (**59 millions** of parameters) for the **master**
- Test of different variations of the distillation loss (table on the right)

DNN	F1 score	Precision	Recall
MASTER	**85.7**	**84.5**	**86.8**
UNet + VGG	78.2	88.3	70.1
UNet + EffNet	76.9	73.6	80.5
FPN + VGG	71.8	68	75.9
FPN + EffNet	69.2	79.4	61.4
PSP + VGG	76.7	81.5	72.3

Results on ALCD DB testset (all models have around 1M parameters)

DNN with Different Losses	F1 score	Precision	Recall
MASTER	**85.7**	**84.5**	**86.8**
Original loss (MSE on logits with T=10)	78.2	88.3	70.1
Variable weighting + MSE on logits with T=1	78.6	86.3	72.2
Variable weighting + MSE on logits with T=10	82.9	90	76.9
Variable weighting + MSE on softmax output with label smoothing	82.8	92.2	75

Results on ALCD DB testset with a UNet-vgg architecture

AGENIUM SPACE

Training & Distillation – UC2: Deforestation Detection (WP2200 & WP 4200)

<u>Selected data base:</u>

- 940 geo-localized disjoint patches spread across Slovenia (57% of forest)

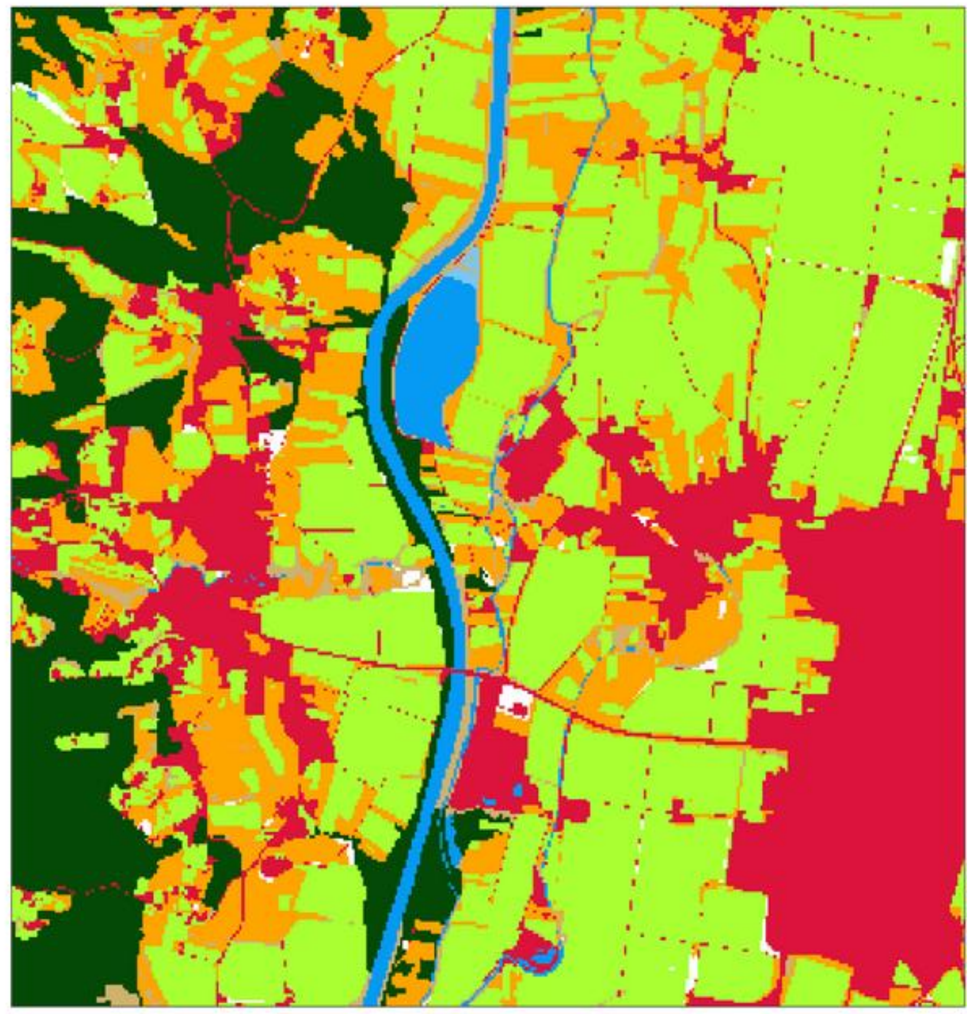
Ground Truth provided by : **Ministry of Agriculture, Forestry and Food – Republic of Slovenia** : https://www.gov.si/en/state-authorities/ministries/ministry-of-agriculture-forestry-and-food/

- For each patch : 96 Sentinel-2 images from 2019 (reflectance)
- Each image is 500 x 500 pixels large (10m resolution)

Data used for training:

- Almost 20 tiles per patch, 5 per season
- 4 channels : RGB+NIR

- Statistics will be computed from older ground truth of Slovenia (from 2006 to 2016) provided **by Ministry of Agriculture, Forestry and Food – Republic of Slovenia** : https://www.gov.si/en/state-authorities/ministries/ministry-of-agriculture-forestry-and-food/



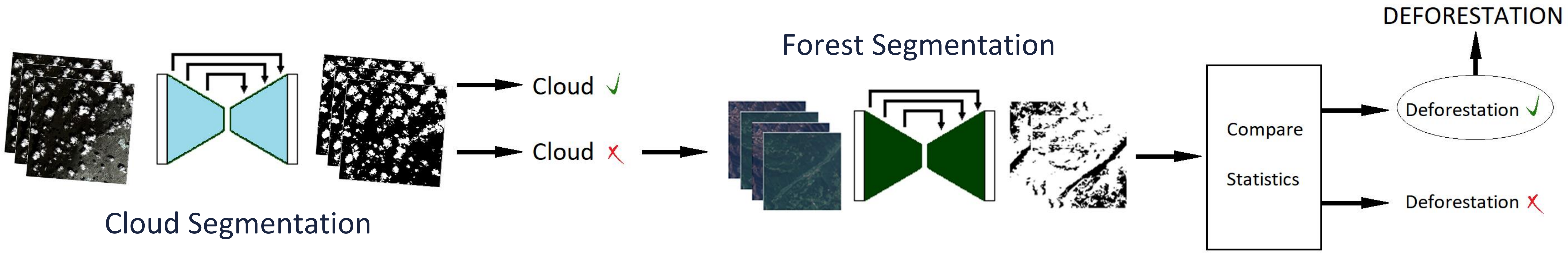
AGENIUM SPACE

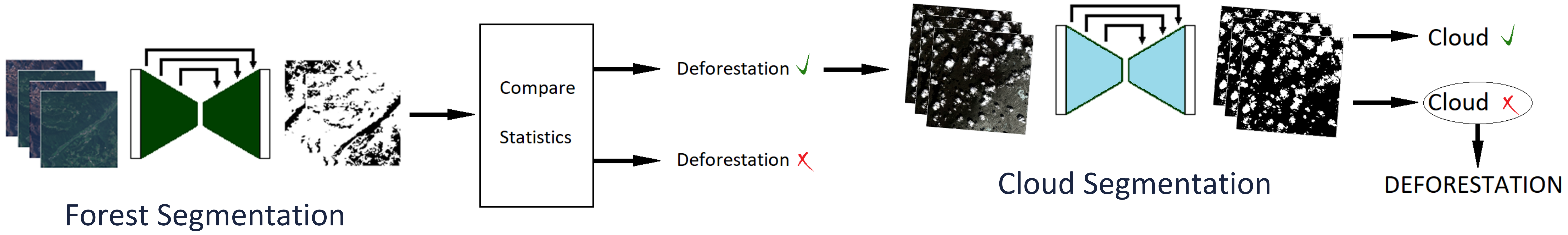
Training & Distillation – UC2: Deforestation Detection (WP2200 & WP 4200)

Different Possibilities :

- Two DNN

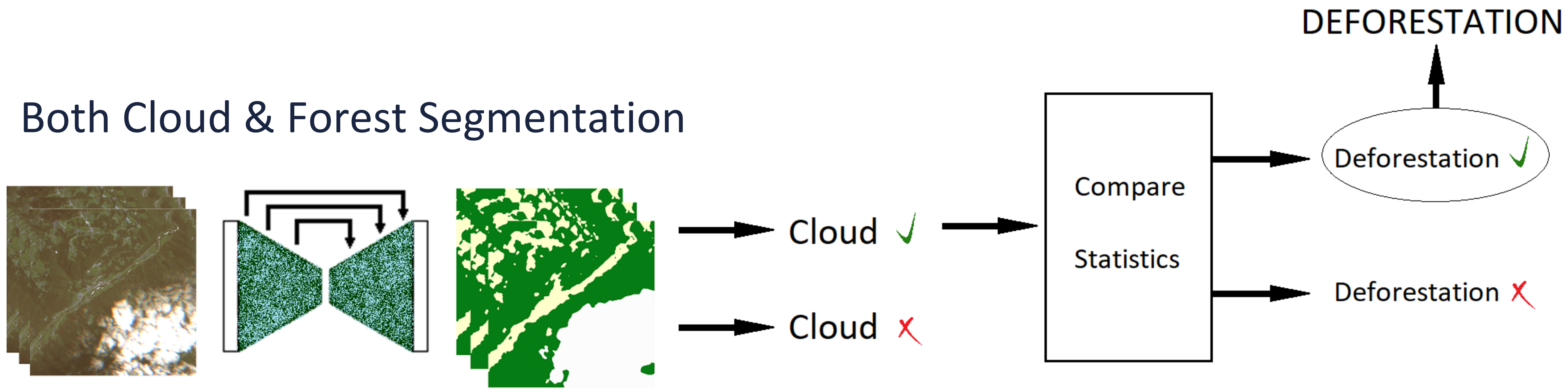
  trained and efficient





- One DNN trained

  but less efficient



Training & Distillation – UC2: Deforestation Detection (WP2200 & WP 4200)

Results :

- Distillations done with 100k parameters for the *students (~ 572 ko in float).*
- Architecture : Unets style (with EfficientNet B2 Encoder for the master)
- Good results : loss of **2-3% for the forest f1-score** with a network more than **150 times smaller.** (The master is 16M parameters ~ 64 Mo in float)

	MASTER (16M parameters)			STUDENT (100k parameters)		
	CLOUDS	FOREST	OTHER	CLOUDS	FOREST	OTHER
Precision	***0.91***	*0.87*	***0.91***	*0.83*	*0.84*	***0.87***
Recall	*0.90*	***0.95***	*0.80*	*0.82*	***0.92***	*0.74*
F1-Score	***0.91***	***0.91***	*0.85*	*0.82*	***0.88***	*0.80*

	MASTER (16M parameters)		STUDENT (100k parameters)	
	FOREST	OTHER	FOREST	OTHER
Precision	***0.93***	***0.93***	*0.88*	***0.92***
Recall	***0.96***	*0.88*	***0.96***	*0.81*
F1-Score	***0.94***	*0.90*	***0.92***	*0.86*

Training & Distillation – UC3: Boats Detection with BBOX (WP2200 & WP 4200)

<u>Selected data base:</u>

- Airbus Ship Detection, Challenge from Kaggle
- 768 x 768 pixels wide images (3 channels RGB)
- Label : class + bounding boxe coordinates : ***decided to detect only one class « Boat »***

Training & Distillation – UC3: Boats Detection with BBOX (WP2200 & WP 4200)

Results :

- Architecture : SSD300 network, the original with a VGG16 encoder for the master, and the number of kernels per layer was decreased for the students.

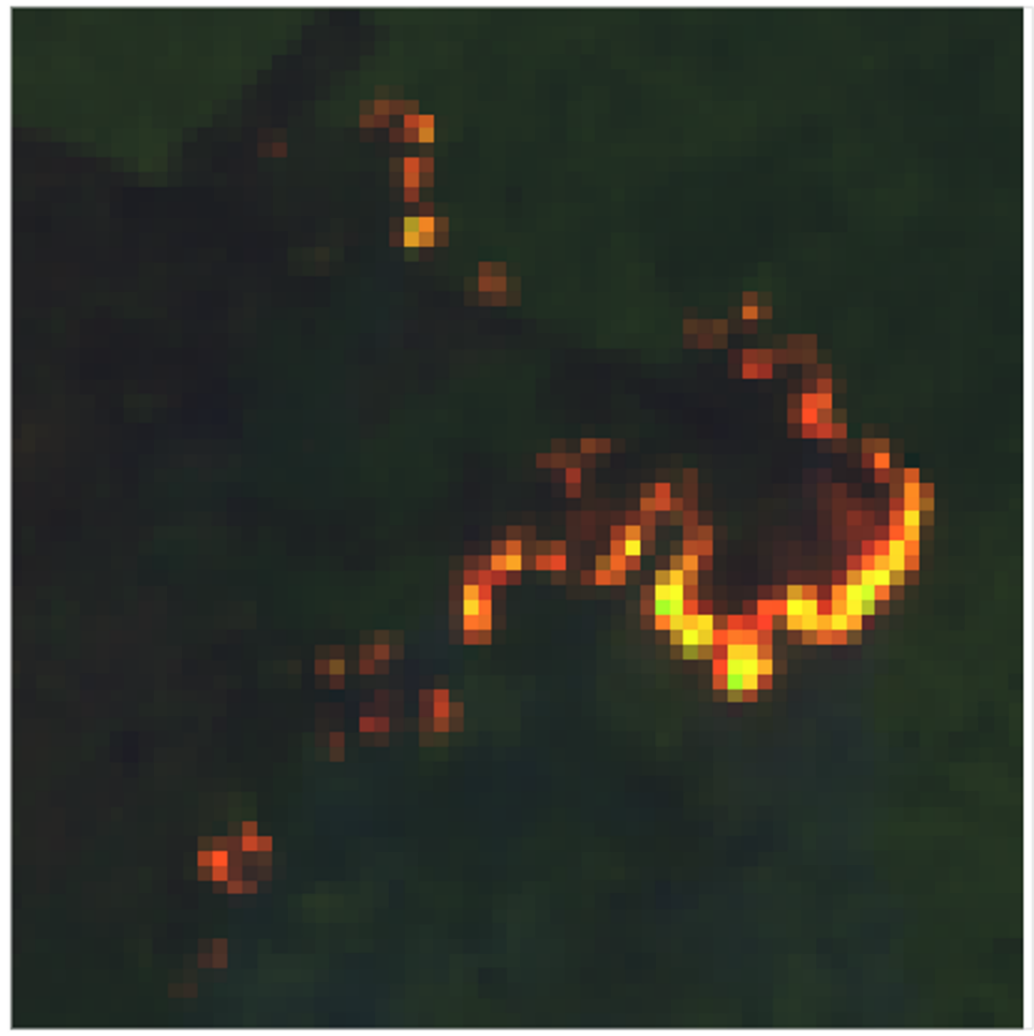
- Distillation with 1M parameters is a good candidate.

	MASTER (24M parameters)	STUDENTS		
		300k params	1M params	4M params
Precision	*84 %*	*83 %*	*86 %*	*85 %*
Recall	*81 %*	*73 %*	*76 %*	*76 %*
F1-Score	*83 %*	*77 %*	*81 %*	*80 %*

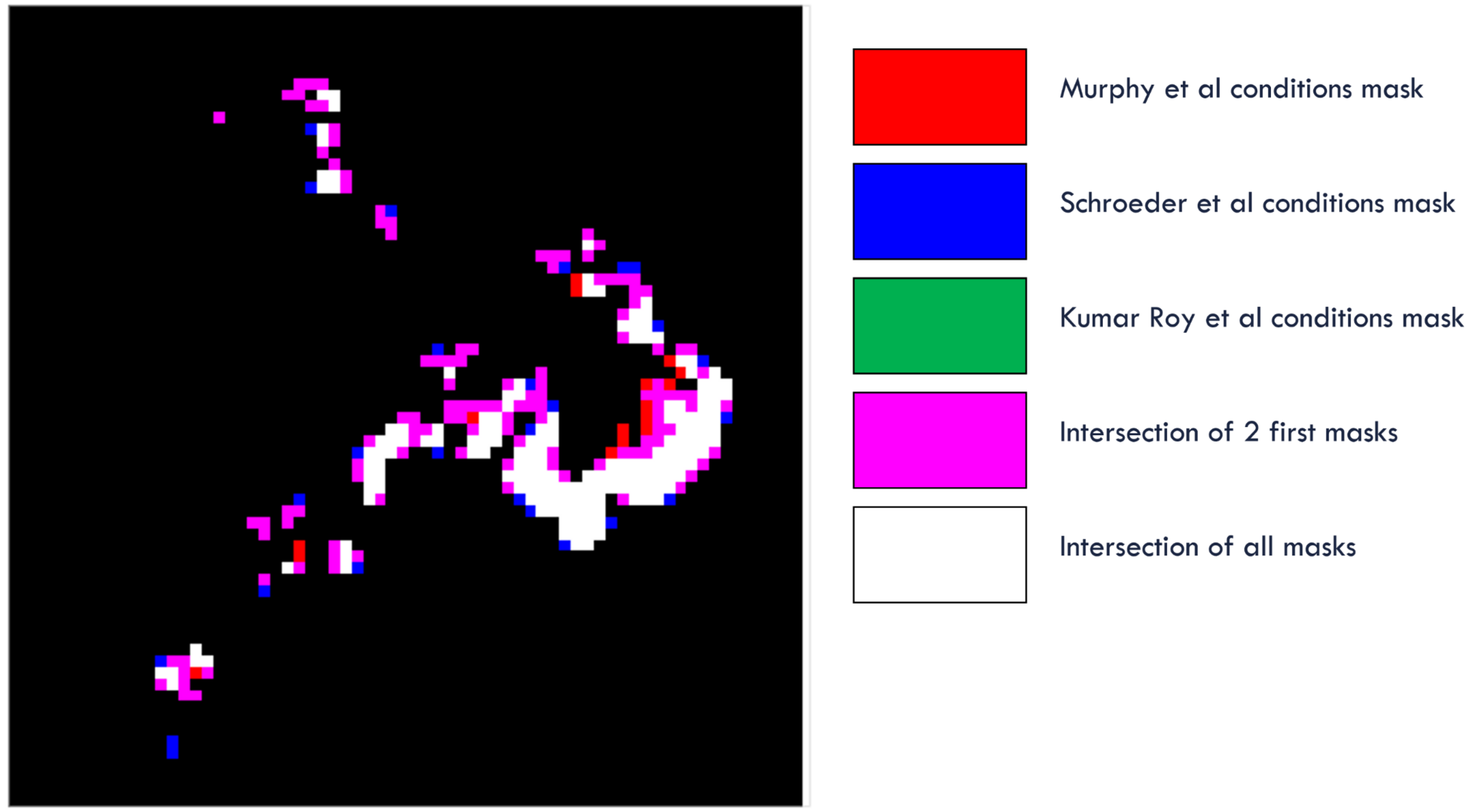
Training & Distillation – UC4: Fire Segmentation (WP2200 & WP 4200)

Selected data base:

- Ground truths generated using literature algorithm (Schroeder et al. conditions, Murphy et al. conditions, Kumar-Roy conditions)
- Dataset provided by Gabriel Henrique de Almeida Pereira, Andre Minoro Fusioka, Bogdan Tomoyuki Nassu, Rodrigo Minetto (UTFPR)
- Landsat-8 images (in reflectance) : Resolution of 30m for the first 8 bands, 100m (over-sampled) for the two last ones
- 256 x 256 (7680 x 7680 m²) geolocalized images spread across the whole world.
- 10 bands available : Aerosol, Blue, Green, Red, NIR, SWIR1, SWIR2, Cirrus, LWIR1, LWIR2



Sample from the dataset (false colours)



Some ground truths examples

AGENIUM SPACE

Training & Distillation – UC4: Fire Segmentation (WP2200 & WP 4200)

Conclusion:

- This problem seems ***too difficult*** without using SWIR bands, and ***too simple*** when using them.
- No real interest in increasing the number of trainable parameters in our architectures.
- Subjective Ground Truth which makes it difficult to train and evaluate the networks.

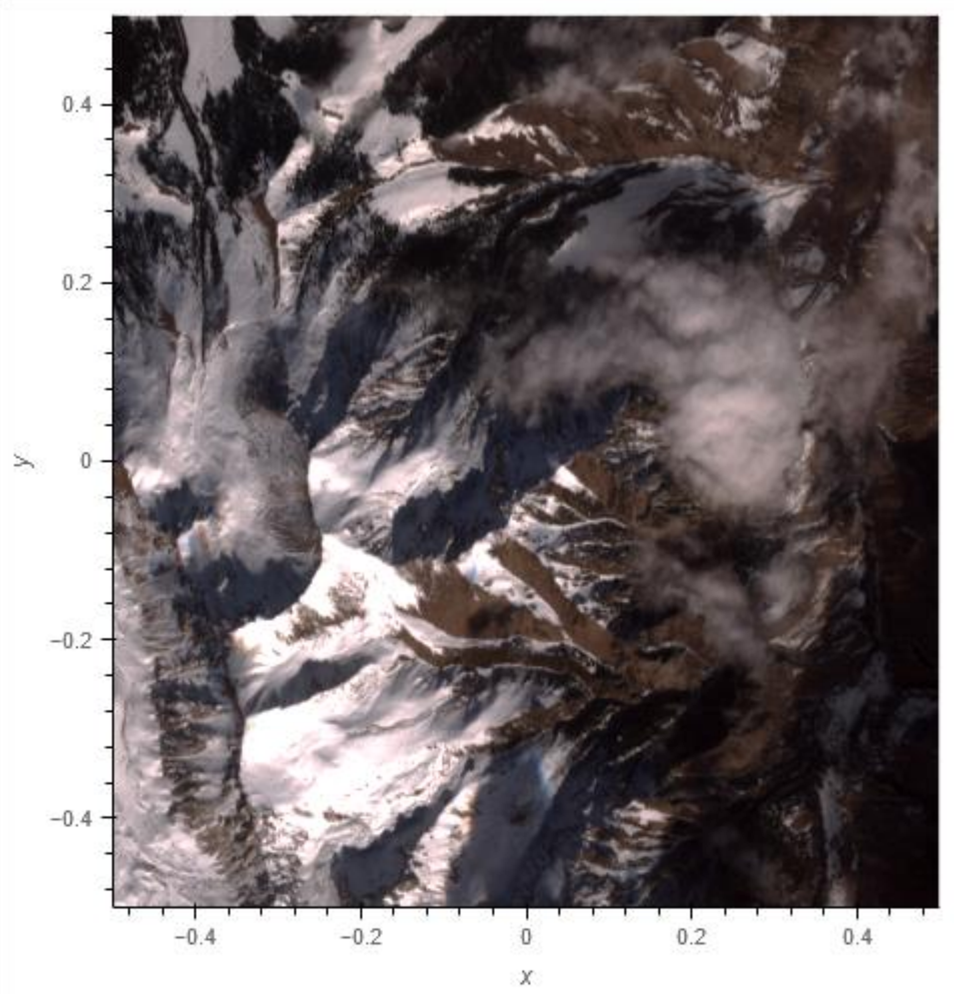
➢ The results ***do not*** seem to have enough advantages to ***justify the use of DL*** for this problem, ***the classical methods seems enough, simple and efficient***.

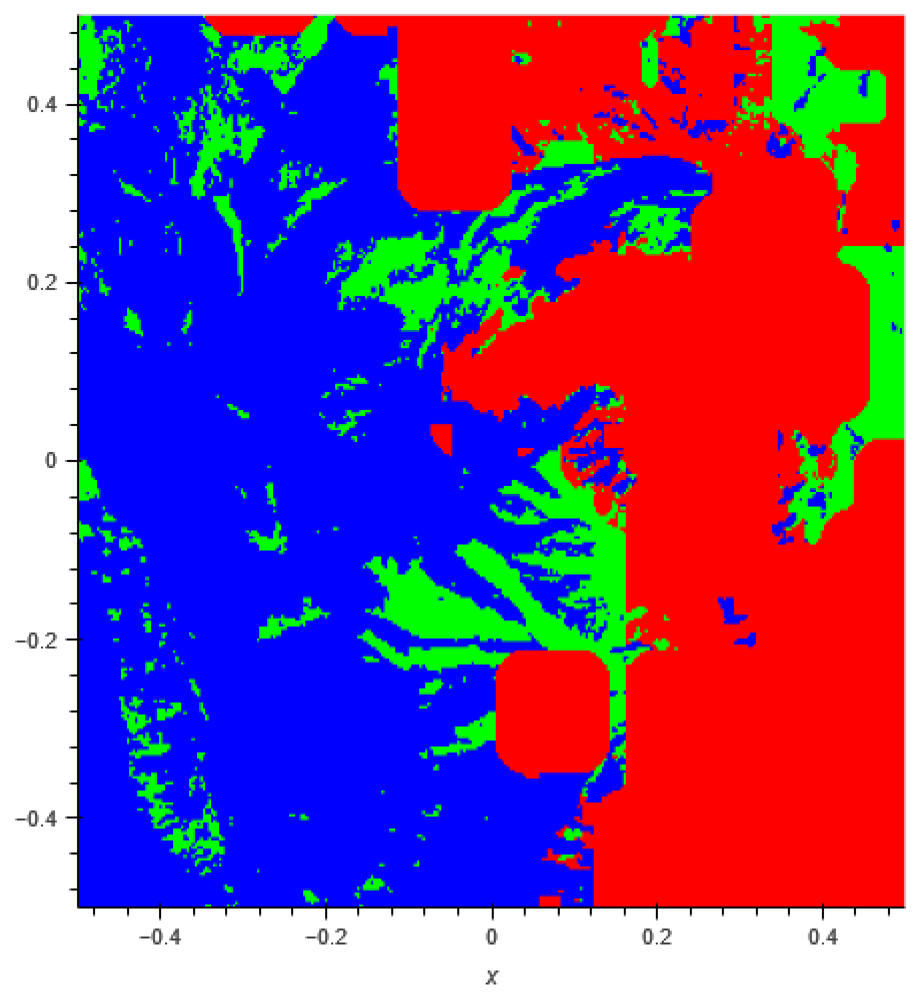
Training & Distillation – UC5: Cloud vs Snow Segmentation (WP2200 & WP 4200)

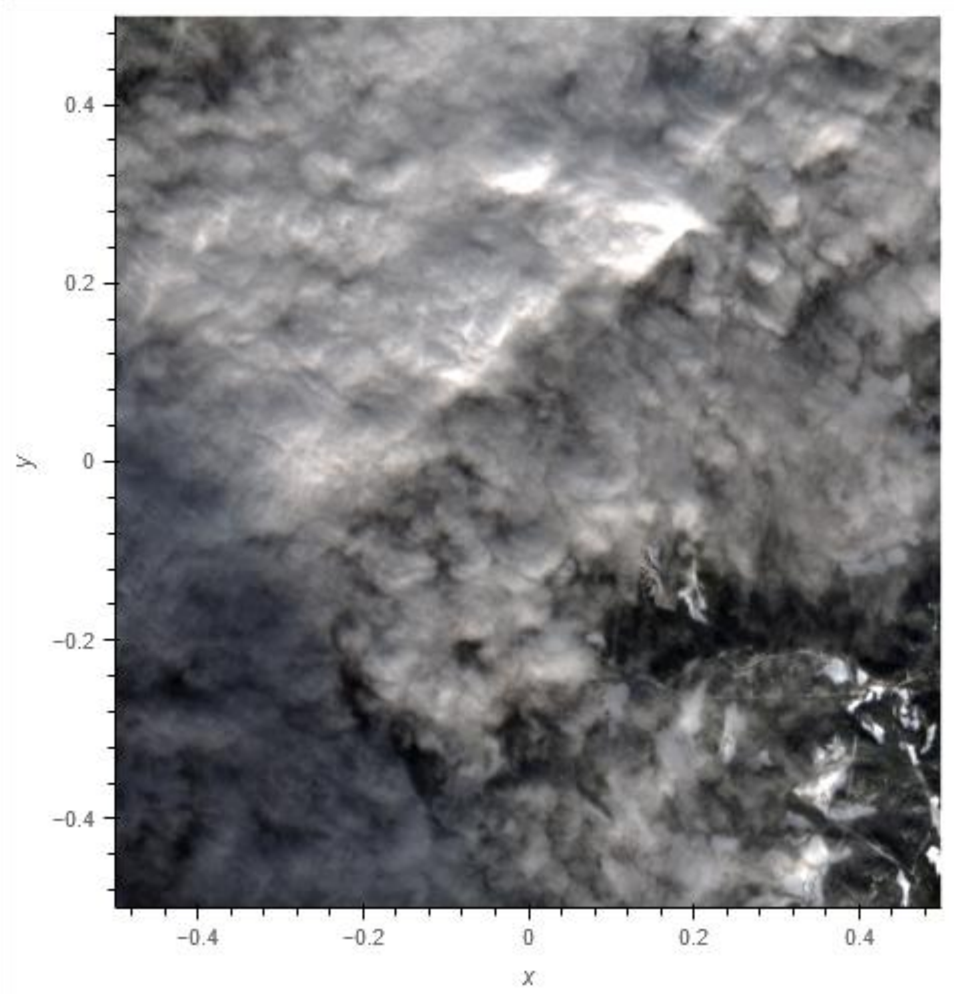
Selected data base:

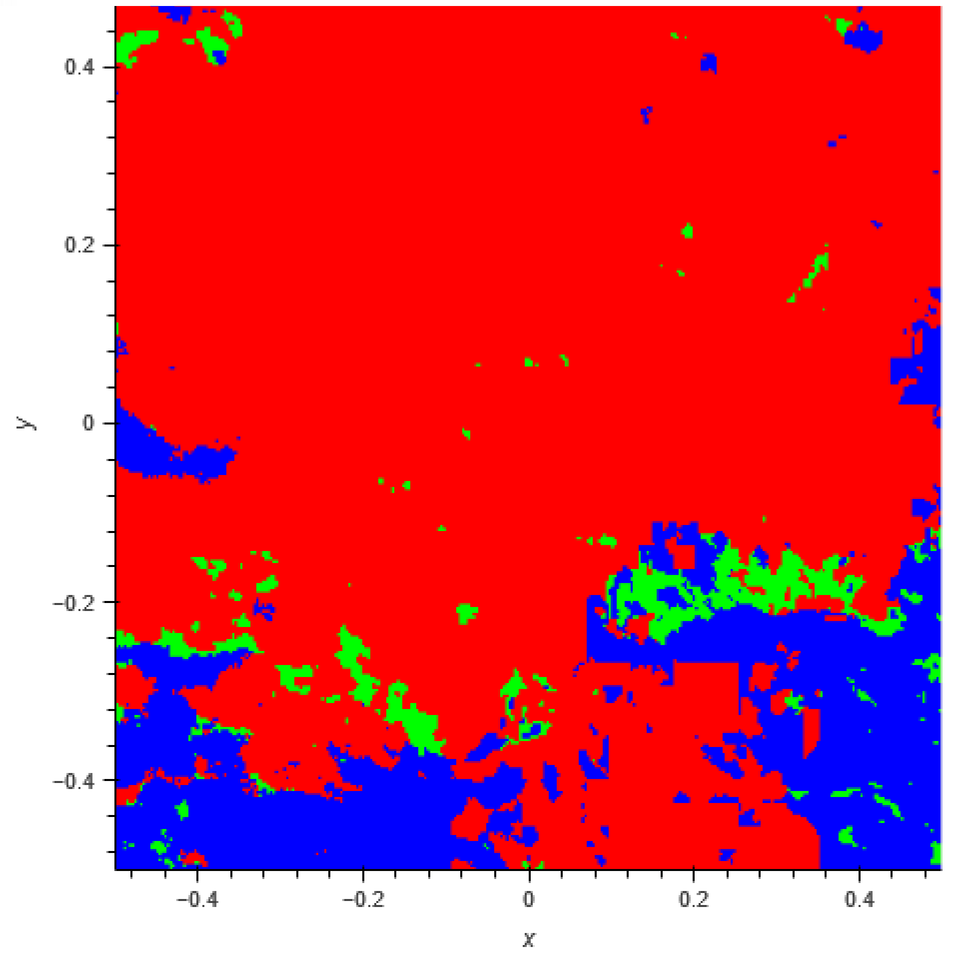
- Use of Theia Snow products
- Masks at 20m resolution with snow/cloud/other
- Cloud masks are conservative and coarse (based on L2A masks from MAJA at 240m resolution)
- Tests with RGB-NIR & RGB-NIR-SWIR1-2

Details in: https://essd.copernicus.org/articles/11/493/2019/









Training & Distillation – UC5: Cloud vs Snow Segmentation (WP2200 & WP 4200)

Results:

- Distillation of model with RGB+NIR bands (as performant as RGB+NIR+SWIR1-2)
- Training of small models from scratch (i.e. no distillation)
- Different operating modes:
  - Too few parameters => neither of them performs well
  - Many parameters => standard training gives results close to reference GT -- distillation training converge to master prediction
  - Reduced number of parameters => too few parameters to overfit reference or master -- distillation gives a better generalization with few parameters

MODELS	NB PARAMS	F1-score			Precision			Recall		
		Snow	Cloud	Others	Snow	Cloud	Others	Snow	Cloud	Others
U-Net-EffNet-B5	59M	*87.1%*	*80.3%*	*86.5%*	*83.4%*	*80.9%*	*89.3%*	*91%*	*79.7%*	*83.9%*
U-Net from scratch	100k	77%	69.6%	71.2%	64.7%	70.9%	**89.4%**	**95%**	68.3%	59.2%
U-Net distilled	100k	79.6%	59.9%	39%	70.4%	48%	86.3%	91.7%	**79.6%**	25.2%
U-Net from scratch	500k	80.5%	73%	**82.6%**	69.8%	**84.5%**	87.8%	**95.2%**	64.3%	**78%**
U-Net distilled	500k	**83.1%**	**74.5%**	**82.6%**	**76.5%**	**75.2%**	88.9%	91%	73.8%	77%
U-Net from scratch	1M	**83.9%**	**76.7%**	79.1%	**75.9%**	73.9%	**91.6%**	93.9%	**79.6%**	69.7%
U-Net distilled	1M	81.9%	73.9%	**82.7%**	75.6%	**75.2%**	88.1%	89.3%	72.6%	**77.9%**

AGENIUM SPACE

HW Tests
(WP3000)

AGENIUM
SPACE

HW Tests - Use cases tested (WP2200)

Use case	DNN method	Input data	Project version
Cloud coverage	Segmentation or classification	Aerosol-RGB-NIR (Sentinel-2 10m, Landsat-8 30m)	**v1**
Deforestation	Change detection	RGB+NIR (Sentinel-2 10m)	**v1**
Vessel classification	Classification and bounding boxes detection	RGB, Airbus Dataset Kaggle.	**v1**
~~**Fire detection**~~	Segmentation	Multispectral (Landsat-8 30m)	v2
Snow vs clouds	Segmentation	RGB (Sentinel-2 10m, Landsat-8 30m)	v2

AGENIUM SPACE

HW Tests - Devices Tested (WP2100)

Device	Type	Qualified/Mission	Missions	Project version
Xilinx Zynq UltraScale+	COTS FPGA	On-going (Leopard DPU)	Institutional & Small	v1
Xilinx Zynq 7000 Series	COTS FPGA	-	Small	v1
~~AMD R-Series~~ AMD G-Series	COTS GPU	-	Small	v1
Xilinx Kintex Ultrascale	RT-FPGA QML	YES	Institutional	v2
Intel Myriad 2	COTS VPU SoC	YES (on-going)	Small	v2

AGENIUM SPACE

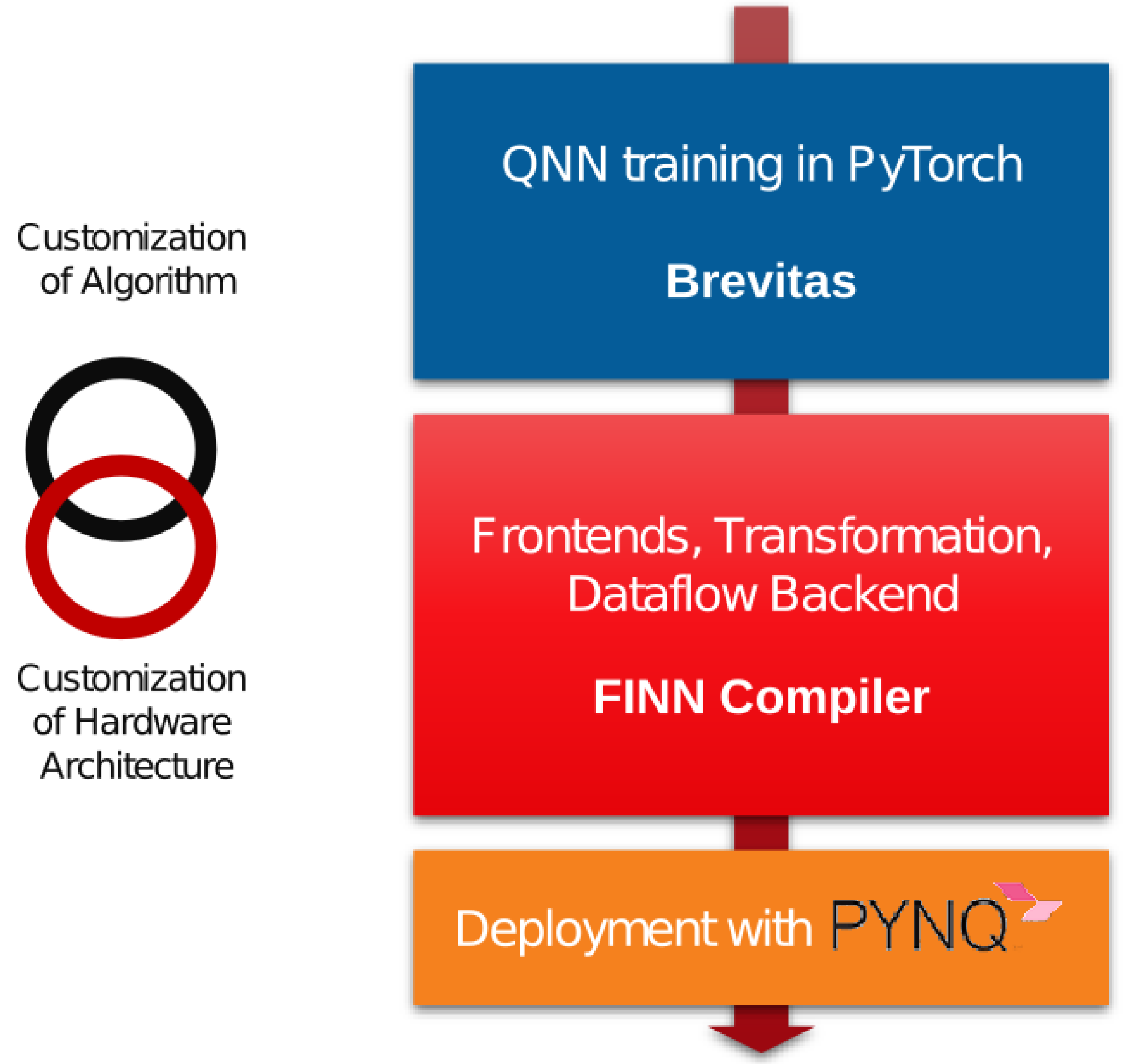
HW Tests – Xilinx : FINN VS VITIS AI (WP3000)

- Pynq : open-source Project from Xilinx
  - Python Lib
  - Using programmable logic with python
  - Compatible with all **Zynq** devices

- Brevitas : research project from Xilinx
  - Pytorch Lib for Quantization-aware training
  - Used with FINN to manipulate FINN-ONNX models

- FINN : Experimental framework from Xilinx
  - Generate a specific FPGA accelerator for a NN
    - Each layer independent on the chip

- **FINN was not ready to use**
  - **We fixed a lot of issue & edit their scripts**
  - **Inference too slow for INT8 : 60x slower than VAI**
  - **Could be interesting for <INT4, binary**



FINN	VITIS AI
Flexibility Architecture/Ressources	Fixed Architectures
Put weights on the on-chip memory: **Easy**	Put weights on the on-chip memory: **Difficult**
Custom data-types supported	Only 8-bits accelerators on Zynq US+

HW Tests - AMD Framework (WP3000)

Quantization tested	Pytorch	TFlite
INT8	Very Slow	Very Slow
FP16	Not Tested	Working fine Near zero loss

→ **AMD G-series has not AVX2 instructions**

→ **FP16 10 times faster than INT8 inference**

API	TF Python	TFlite Python	TFlite C++
EMBEDDED GPU	✗	✗	✓

PyTorch

TensorFlow Lite

AGENIUM SPACE

HW Tests – Requirements (WP3000)

Requirements	Description
RQ-1: Segmentation architectures scores	< 5% f1-score drop
RQ-2: Detection architectures scores	< 5% f1-score drop
RQ-3: Segmentation architectures FP ratio	< 10% **FP ratio** (False positive ratio) & **FAR** (False Alarm ratio) drop
RQ-4: FPGA throughput for Detection architectures	> 0.4M pixels/second/Watt (tol.:+/-5%)
RQ-5: FPGA throughput for Segmentation architectures	> 0.16M pixels/second/Watt (tol.:+/-5%)
RQ-6: AMD throughput for Detection architectures	> 27k pixels/second/watt (tol.:+/-5%)
RQ-7: AMD throughput for Segmentation architectures	> 11k pixels/second/watt (tol.:+/-5%)
RQ-8: Robustness across configurations	Bit perfect level of the output between different DPU architectures
RQ-9: Robustness across devices	Bit perfect level targeted, but tolerance is applied (cf. Test Plan)
RQ-10: Distillation efficiency	Distilled model efficiency x4 compared to master model
RQ-11: Myriad throughput for Detection architectures	> 200k pixels/second/Watt (tol.:+/-5%)
RQ-12: Myriad throughput for Segmentation architectures	> 80k pixels/second/Watt (tol.:+/-5%)

HW Tests - Board Configuration (WP3000)

Device	Board	Configuration	Framework
Xilinx Zynq UltraScale+	ZCU102 (ZU9EG)	• 3*B4096 • 2*B4096 • ~~1*B4096~~	Vitis AI 1.2
Xilinx Zynq 7000 Series	ZEDBOARD (Z7020)	1*B1152	Vitis AI 1.2
~~AMD R-Series~~ AMD G-series	Unibap IX5	UNIBAP SPACECLOUDS OS	TFlite (TF 2.4.3)
~~Xilinx Kintex Ultrascale KU060~~ Xilinx Kintex Ultrascale KU040	KCU105 (KU040)	1*B4096 + Microblaze	Vitis AI 1.3
Intel Myriad 2	Intel Compute Stick V2	N/A	Intel OpenVINO

AGENIUM SPACE

HW Tests – Combinatorial (WP3000)

Use-case	Xilinx ZCU 102	Xilinx 7020	AMD	Xilinx KU040	Intel Myriad 2
UC1 - Cloud coverage	V1	V1	V1	V2*	V2*
UC2 - Deforestation	V1	V1	V1	V2*	V2*
UC3 - Vessel classification	V1	V1	V1	V2	V2
UC4 - Fire detection	N/A	N/A	N/A	N/A	N/A
UC5 - Snow vs clouds	V2	V2	V2	V2	V2

- V1: Done in test V1
- N/A: no model to be tested, will not be tested
- V2: Done in V2
- V2*: optional done in V2

HW Tests - Tests Performed (WP3000)

Tests	Planned	Performed	Passed	Acceptable	Failed
Xilinx-ZCU102	117	100%	71	7	39
Xilinx-7020	29	100%	20	3	6
AMD-G-Series	32	100%	27	2	3
Xilinx-KCU105	29	100%	19	2	8
Intel Myriad 2 VPU	29	100%	24	4	1
Total	**236**	**100%**	**161**	**18**	**57**

- Failed Tests:
  - 35 : 1*B4096 configuration not working
  - 21 : Throughput requirement (UC1 & UC5 Tiled) – ZCU102 (4), Z7020 (6), KINTEX (8), AMD (3)
  - 1 : UC2_SEGM_CLOUD Master model faster than student – Intel Myriad 2 VPU

Scores Summary (WP3000)

F1-scores	Model parameters (M)	Distilled Models	Intel Myriad VPU 2	AMD G-Series (iX5)	Xilinx HW - FPGA & SoCs
Cloud Segmentation	1	Other – 0.96 Cloud – 0.84	Other – 0.96 Cloud – 0.84	Other – 0.96 Cloud – 0.84	Other – 0.96 Cloud – 0.79
Forest Segmentation	0.1	Other – 0.87 Forest – 0.92	Other – 0.87 Forest – 0.92	Other – 0.86 Forest – 0.92	Other – 0.82 Forest – 0.90
Forest & Cloud Segmentation	0.1	Other – 0.81 Forest – 0.88 Cloud – 0.84	Other – 0.81 Forest – 0.89 Cloud – 0.85	Other – 0.81 Forest – 0.88 Cloud – 0.84	Other – 0.75 Forest – 0.86 Cloud – 0.81
Snow VS Cloud Segmentation	0.5	Other – 0.86 Cloud – 0.75 Snow – 0.85	Other – 0.86 Cloud – 0.75 Snow – 0.85	Other – 0.86 Cloud – 0.75 Snow – 0.85	Other – 0.84 Cloud – 0.72 Snow – 0.85
Boat Detection	0.3 to 4	0.79 to 0.82	0.79 to 0.82	0.79 to 0.82	0.79 to 0.82

Embedded scores summary

AGENIUM SPACE

Throughputs summary (WP3000)

Throughputs	Model parameters (M)	AMD G-Series (iX5) (px/s/W)	Intel Myriad 2 VPU (px/s/W)	Xilinx ZU+ SoCs (ZU9EG) (px/s/W)	Xilinx Zynq SoCs (Z7020) (px/s/W)	Xilinx KU FPGA (KCU105) (px/s/W)
Segmentation	0.1 to 1	8k to 19k	170k to 350k	170k to 215k	70k to 120k	430 to 640
Detection	0.3 to 4	7k to 17k *	320k to 640k	350k to 600k	115k to 190k	3000

Throughput ranges summary

Efficiency Comparison (WP3000)

F1-scores	Model parameters (M)	Intel CPU Core i7-9700K (95W)	AMD G-Series (iX5) (10W)	Xilinx HW - FPGA & SoCs (~10W)	Intel Myriad VPU 2 (1W)	Xilinx HW - FPGA (<3.5W)
Cloud Segmentation	1	1.0	0.4	9.2	9.0	0.03
Forest Segmentation	0.1	1.0	0.6	6.5	9.8	0.03
Forest & Cloud Segmentation	0.1	1.0	0.5	5.2	9.6	0.025
Snow VS Cloud Segmentation	0.5	1.0	0.4	8.8	9.9	0.023
Boat Detection	0.3	1.0	0.2	5.6	6.0	0.16

Efficiency summary for the best results, computed as ratio of throughput over reference throughput

- Compared efficiency for the distilled models (best throughput): Workstation VS Embedded
- AMD less efficient than a workstation CPU → **Expect result**
- VPU & FPGA very efficient

→ **Adapt architectures to have better efficiency on the FPGA**

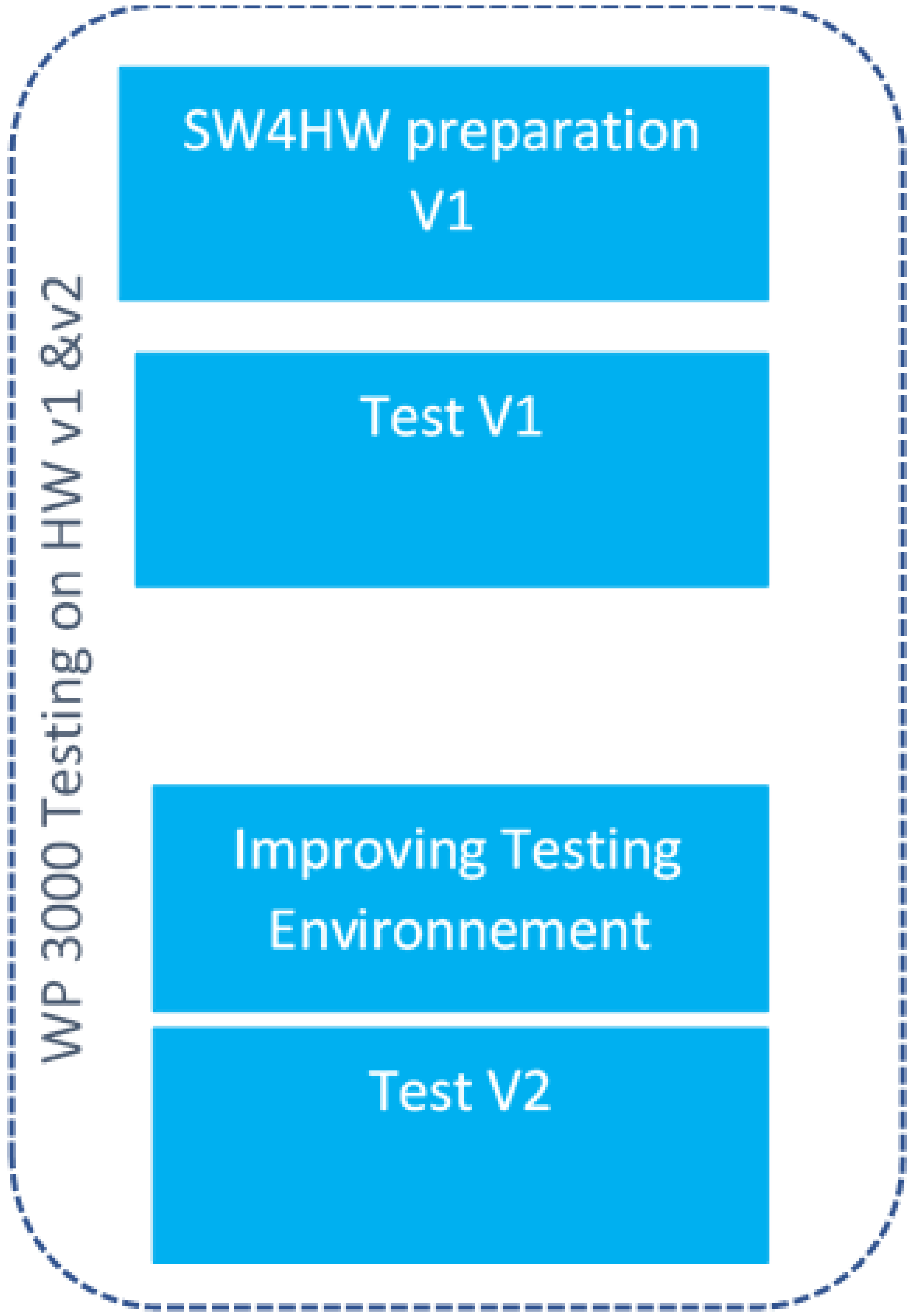
AGENIUM SPACE

HW Tests – Exploratory Tests (WP3000)

- Test variations of the current architectures :
  - Layers modifications
  - Reduce number of Operations to process an input
  - → Maximize the throughputs
  - → Maximize energy efficiency
  - → Identify best suited architectures for Xilinx DPU

- Steps
  1. Create models with random weights
  2. Quantize with Vitis AI and execute the models on Xilinx DPU
  3. Select the best models (throughputs, energy efficiency)
  4. Train the selected models to measure the predictions performances

- Results
  - Variation of Unet-Architectures have been tested
  - **Throughput (pix/s/W) improve by 2.5 for DEEPCUBE models with VAI 1.4**
  - **Power Consumption: +10% compared to DEEPCUBE MODELS with VAI 1.4**



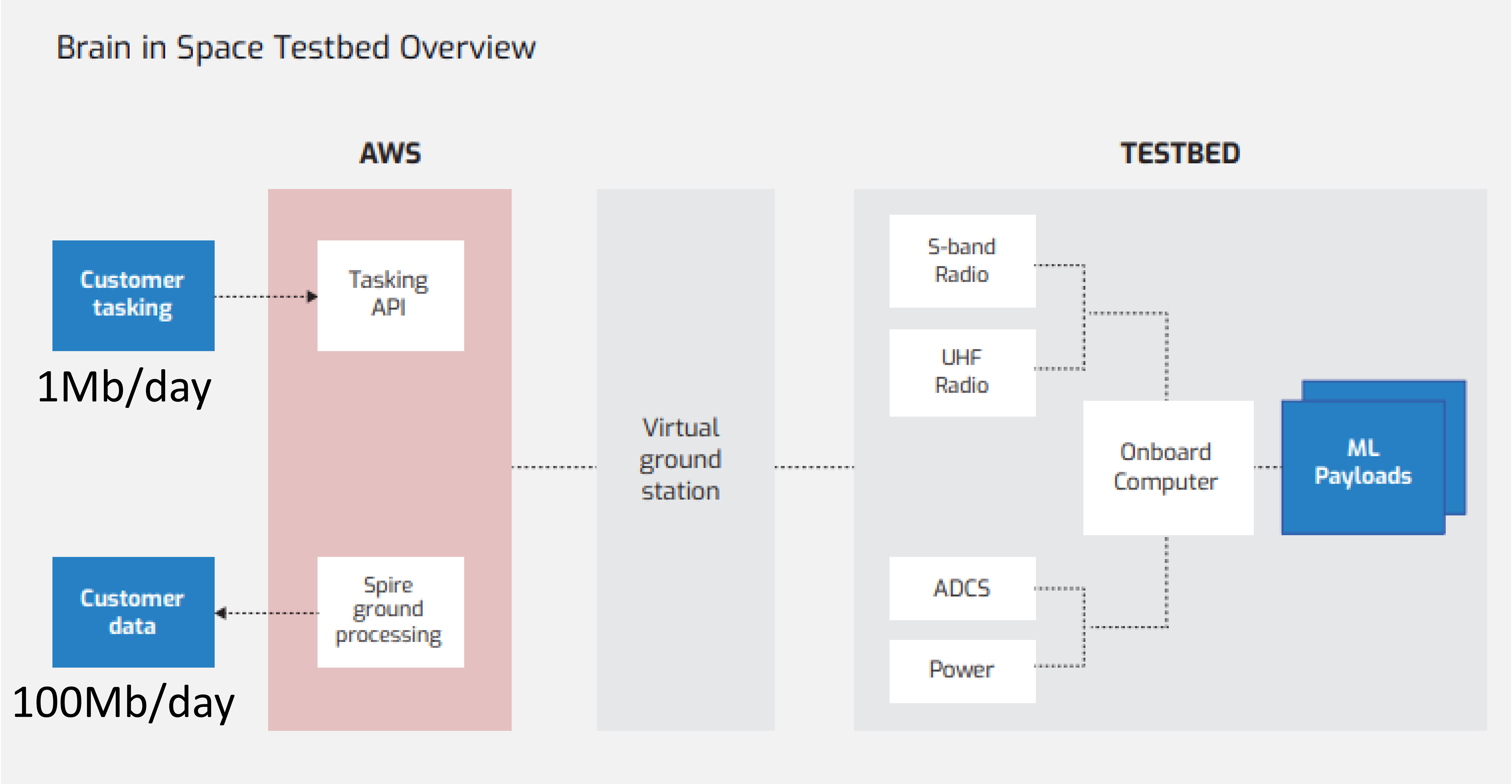
Brain in Space testbed (WP3000)

Interesting points

- Gateway to constellation of CubeSats constraints
- Several available payloads
- Easy to download output of experiments
- Interesting cross-compiler (not tested)
- Tech team relevant and convenient

Constraints maybe too restrictive for a testbed

- Logs of the execution of the script on the testbed not available.
- Large delay for uploading (1h for a script)
- Large delay for downloading (several hours no matter the size)
- Uploading libraries/images not supported yet



Tools and updating (WP5000)

- Base Tensorflow/Keras V1 code was ported to TFV2 to derisk compatibility issues with VITIS AI SW updates and support of PyTorch.

- Base code for distillation is now under PyTorch/PyTorchIgnite.

- Exploratory tests were made for distillation V3, current results are on par with SOTA on extremely small DNN (50K params) using PyTorch/PyTorchIgnite.
  => Still need to consolidate them to propose a new pipeline

Conclusion : achievements and perspectives

- Objective: develop a service as a product to simplify fitting DNN (Deep Learning Neural Networks) in on board HW to make better use of AI on space missions. The objective of the activity is to acquire wider experience on DNN simplification strategies to provide services to the future customers.

- To fulfil this objective, it is mandatory to develop the processing chain. This project allows us to enrich and update the chain
  - Framework change (move to Pytorch) and update of former framework
  - Several use cases have been tested : consolidation of previous results from another ESA OpenCall
  - Several HW for each of this use case have been tested

- All these steps have consolidated our experience and give us confidence in our ability to guarantee the level of flight performance

Perspectives

- Thanks to Unibap/ION : we are flight proven for CPU.
- What's next :
  - Flight proven FPGA : IOD project through Occitanie Region funding
  - Flight proven FPGA + our own IP : OPS SAT experimentation (ESA)
- Business development of our Deepcube service

AGENIUM SPACE

AGENIUM Space

Rosa.ruiloba@agenium.com

1, avenue de l'Europe
31400 TOULOUSE
FRANCE
t : +33 (0)5 61 41 03 98
m :+33 (0)6 46 78 63 34

agenium.group