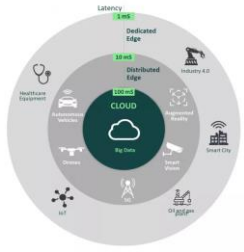
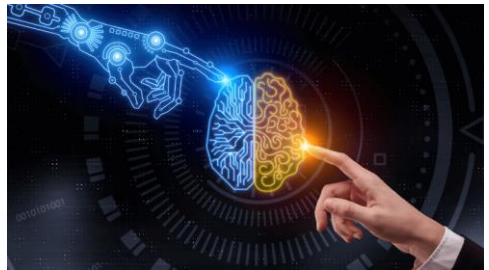
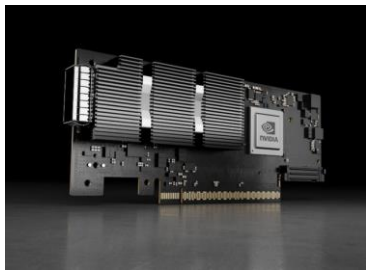
Frédéric Druillole LP2I Bordeaux IN2P3 frederic.druillole@lp2ib.in2p3.fr

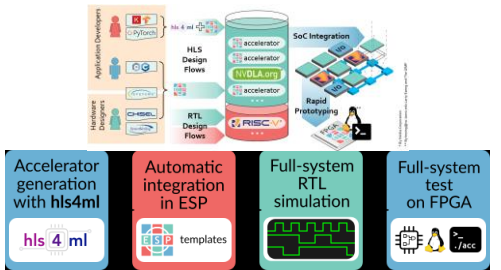
Embedded harware for deep learning neural network: an optimization process– Methodology – Software – Hardware – Firmware –

➔ Dominant Design and Challenges
➔ Stakeholders and Technologies
➔ Methodology and optimized instruments
➔ Futur to prepare

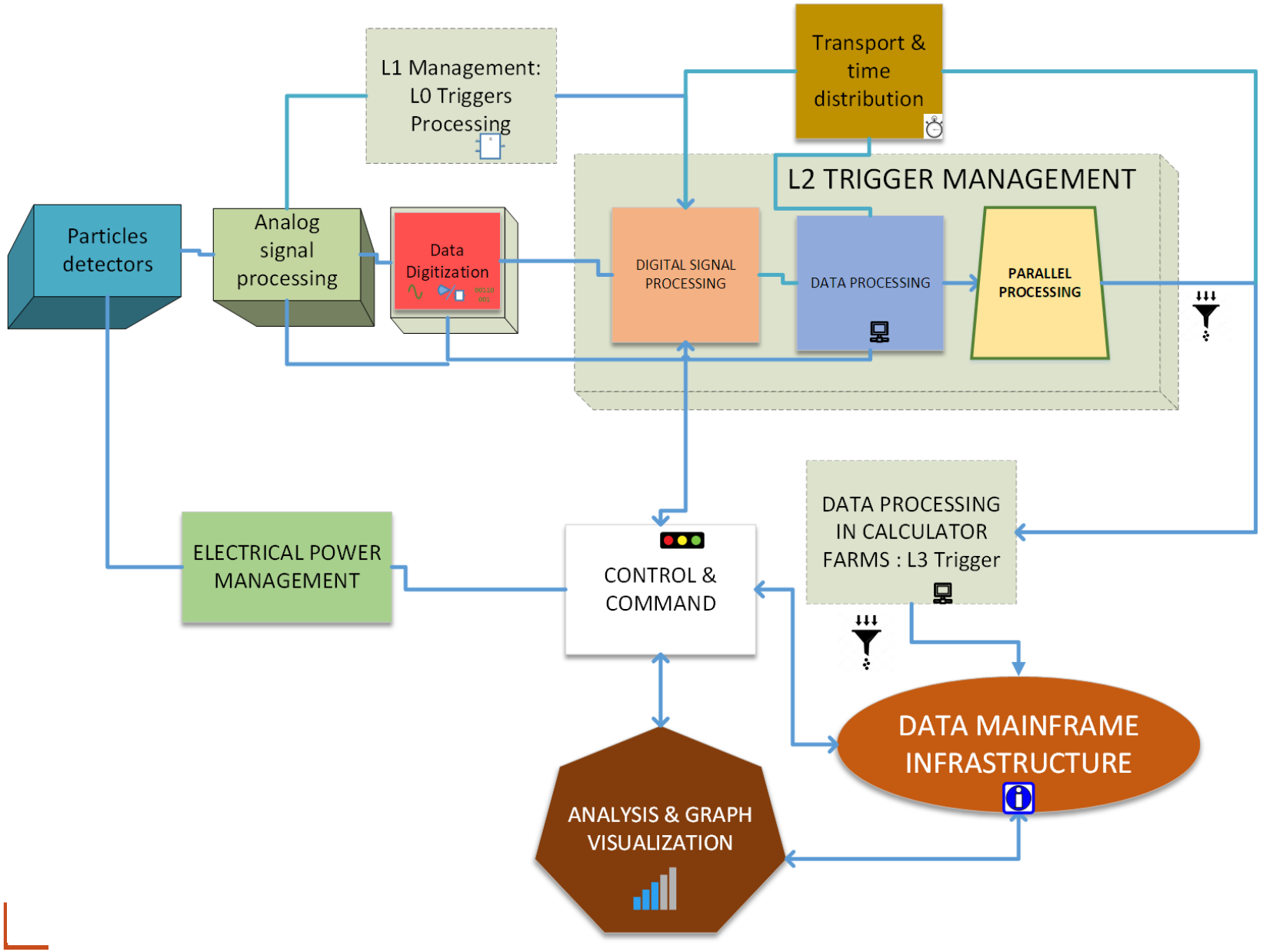




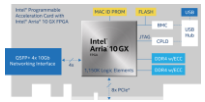


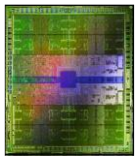


My research : Dominant Design in Instruments for research in fundamental physics



Intelligents algorithms





- **Reduced Data and Selection management:**
  - *L1: FPGA, ASIC, SNN*
  - *L2: FPGA, GPU, SNN*
  - *L3: GPU, MPPA, Accelerated Card*

Challenges in the field

LHCb – 2032 ~2000 Exabytes/year
ATLAS+CMS 2027 ~ 260 Exabytes/year
Square Kilometers Array – 2030 ~ 30000 EB/year

2021 global Ethernet Dataflow ~2800 EB/year

DataStream before storage
LHCb – 2032 ~500TB/s
ATLAS+CMS 2027 ~ 20-40 TB/s

Forecast cost of storing data to disk (Annual)
LHCb – 2032 ~2,5 Billions of €
ATLAS+CMS 2027 ~ 325 Millions of €

Challenge: Real-time data reduction to avoid disk storage (very expensive):

➔ Embedded algorithms in decision nodes

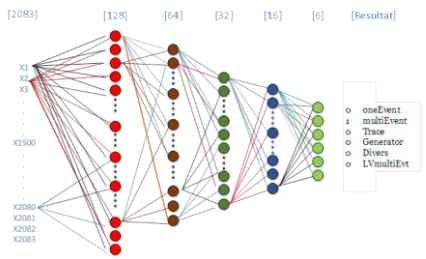
➔ Optimize processing (classifications, Prediction, Selection)

➔ Use a mixed GPU, MPPA, FPGA

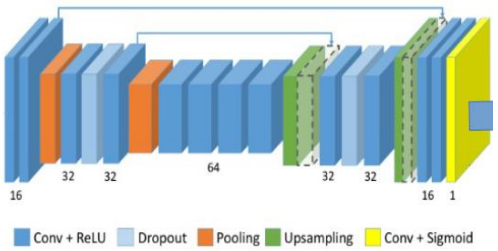
- **Use powerfull hardware component to compute ML Model**
- **And deploy them in ours instruments**

WHAT WE COULD DO WITH ML

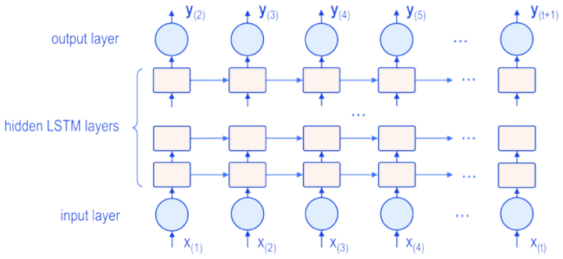
DEEP NEURAL NETWORK



CONVOLUTIONAL NEURAL NETWORK



RECURRENT NEURAL NETWORK



DECISION TREE

RANDOM FOREST

Off-line

- Signal generation
- Design Optimisation

On-line

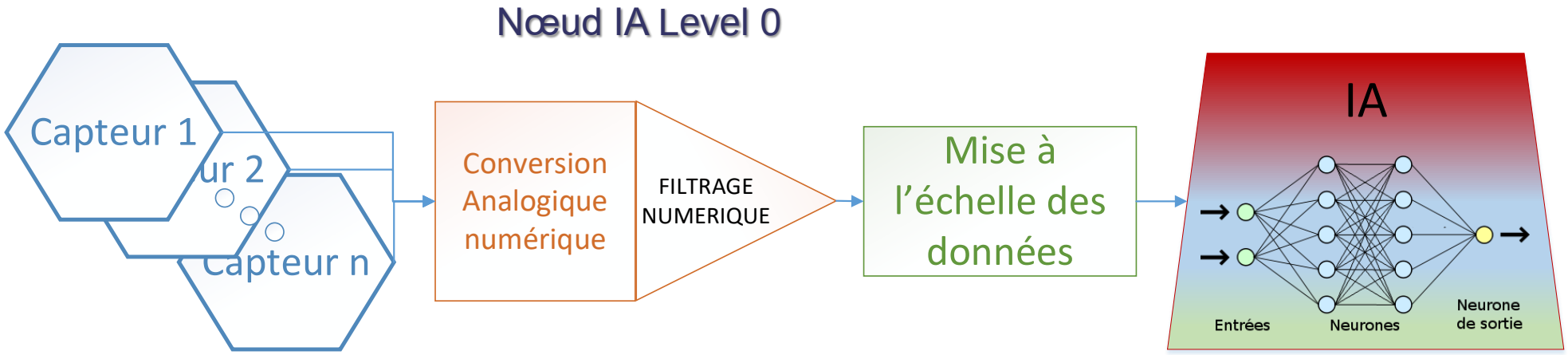
- Instrument Optimization
  - Signal recognition
  - Pile-Up recovery
  - Signal deconvolution
- Selection/ Classification/ Decision
  - Data selection
  - Data parameters prediction
  - Denoizing
- Data Compression
  - Reduced data format

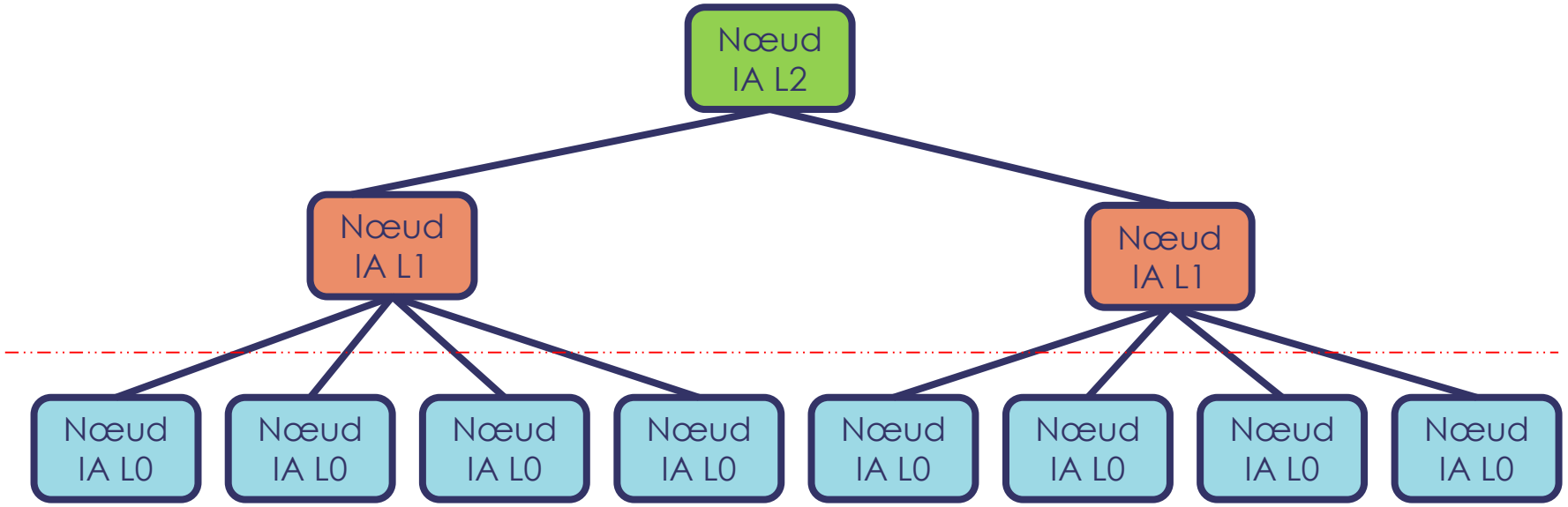
L1

L2

L3

My purpose


Nœud IA Level 0
Capteur 1
ur 2
Capteur n
Conversion Analogique numérique
FILTRAGE NUMERIQUE
Mise à l'échelle des données
IA
Entrées
Neurones
Neurone de sortie


Nœud IA L2
Nœud IA L1
Nœud IA L1
Nœud IA L0
Nœud IA L0
Nœud IA L0
Nœud IA L0
Nœud IA L0
Nœud IA L0
Nœud IA L0
Nœud IA L0

Les trois catégories d'apprentissage profond

DATA DRIVEN

Supervised Learning

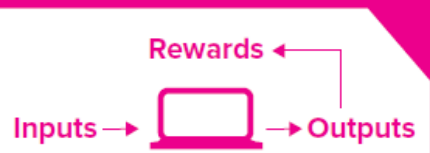
- Makes machine Learn explicitly
- Data with clearly defined output is given
- Direct feedback is given
- Predicts outcome/future
- Resolves classification and regression problems

Unsupervised Learning

- Machines understands the data (identifies patterns/ structures)
- Evolution is qualitative or indirect
- Does not predict/find anything specific

Reinforcement Learning

- An approach to AI
- Reward based learning
- Learning from +ve & +ve reinforcement
- Machine Learns how to act in a certain environment
- To maximize rewards



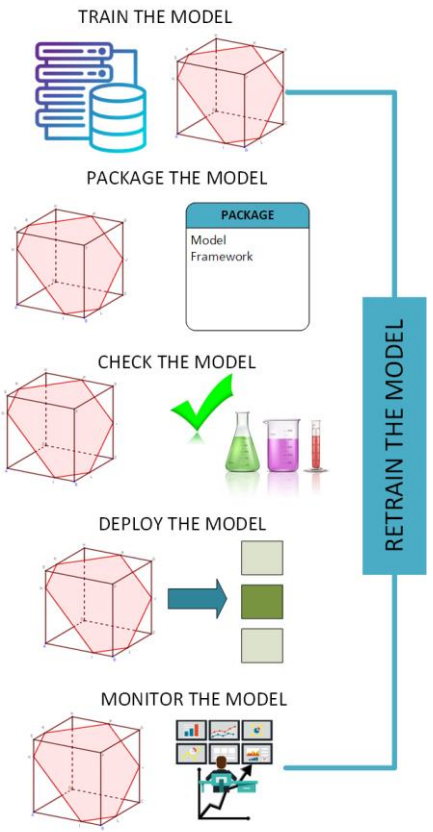
COMPUTING DRIVEN

Transfert learning

- Classical supervised technics
- Change part of the model with new inputs and outputs
- Add features to the preliminary model

SE

Stakesholders ➔ Responsive AI on the Edge



AI-Centric cloud

Big Data firms

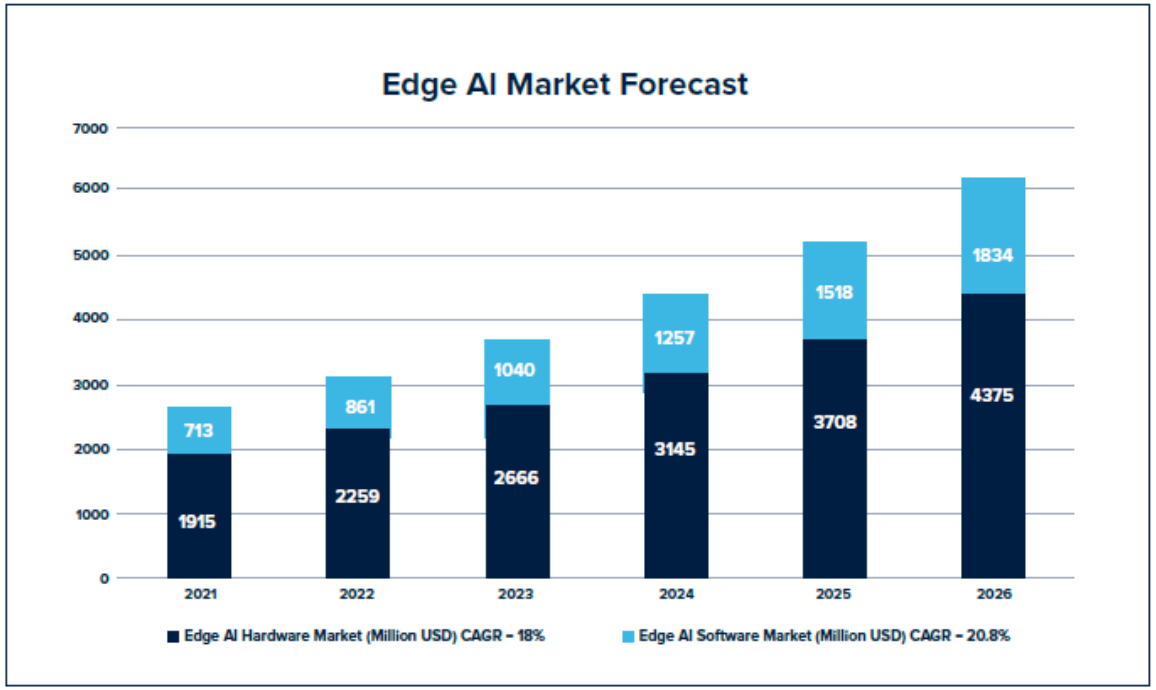
- Hungry in Electrical power
- Hungry in bandwidth
- Hungry in data quality

Migration

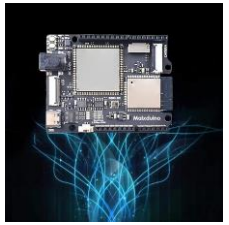
Embedded components network for edge AI

New Paradigme

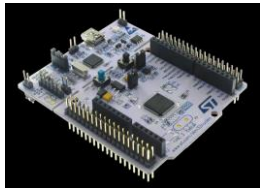
STMicroelectronics 2022



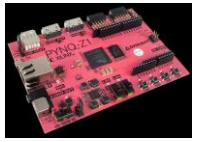
Kendryte K210



STM32 Cube AI



Digilent



AMD-Xilinx

Intel

challenges of ML : culture of work

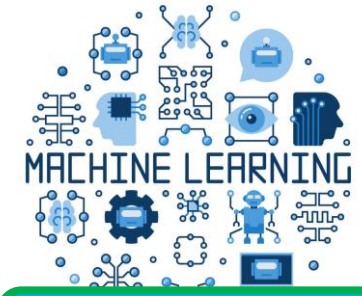
Roles & competencies

- *Data Physicist*
- *System Engineering team*
- *ML Engineer*
- *Software Engineer*
- *Hardware Engineer*
- *Infra & Security teams*

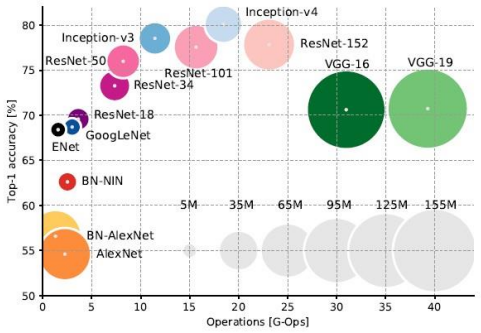
Tools

- *ML Tools:*
  - *TF-KERAS, PyTorch …*
- *HLS4ML (Xilinx…)*
- *HLS*
- *Brevitas & FiNN(Xilinx)*
- *CONIFER (LLR)*
- *N2D2 (CEA)*
- *VHDL*

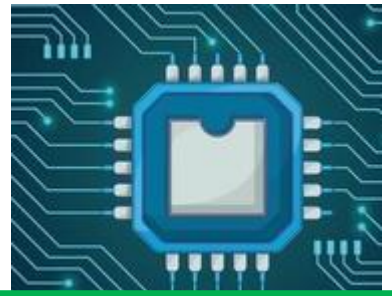
…



Artefacs & ML zoology



- *Model*
- *Code source…*



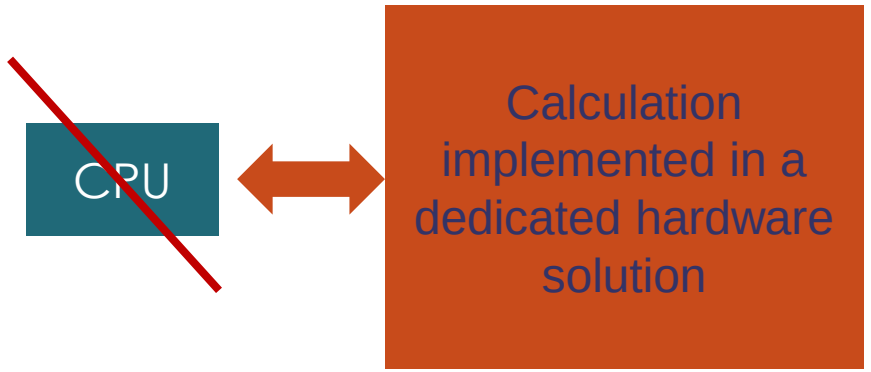
Digital hardware technologies

- *CPU*
- *FPGA SOM*
- *SNN*
- *MPPA*
- *GPU*
- …

Deployement & Operational AI

- *GitLab/Git*
- *Training Service skew*
- *Model Monitoring*
- *Responsible AI*
- …

Embedded AI: 2 technologies

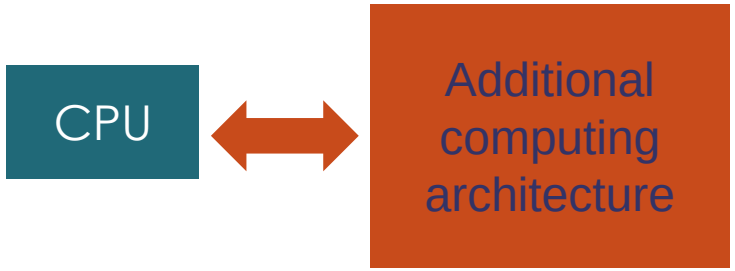


Spatial Accelerators = Fully Firmware

-- based on hardware functions dedicated to calculations without software (matrix calculation)

ASIC
Neuromorphic Circuit
FPGA

2 In Progress

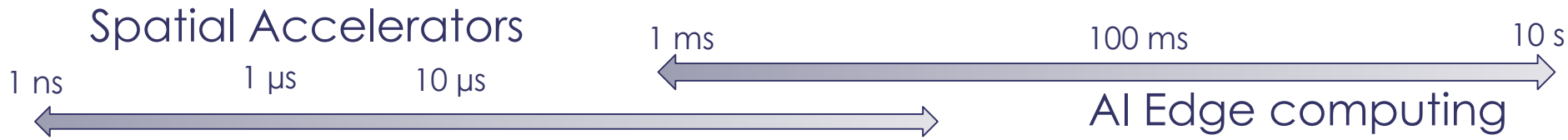


Edge AI Responsive

-- Based on software programmation

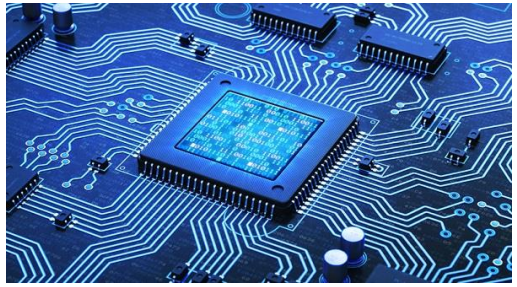
MMPA
GPU
FPGA – SOM-
PU designed for AI(TPU, KPU…)

1 Pilot current industrial developments

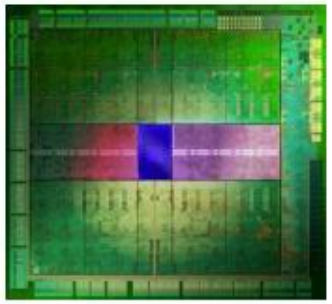


Hardware architectures vs digital hardware engineer

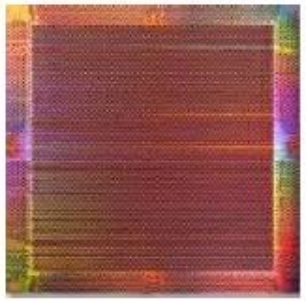
MPPA



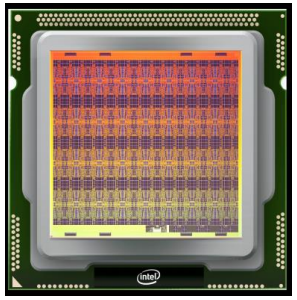
ASICs



GPUs



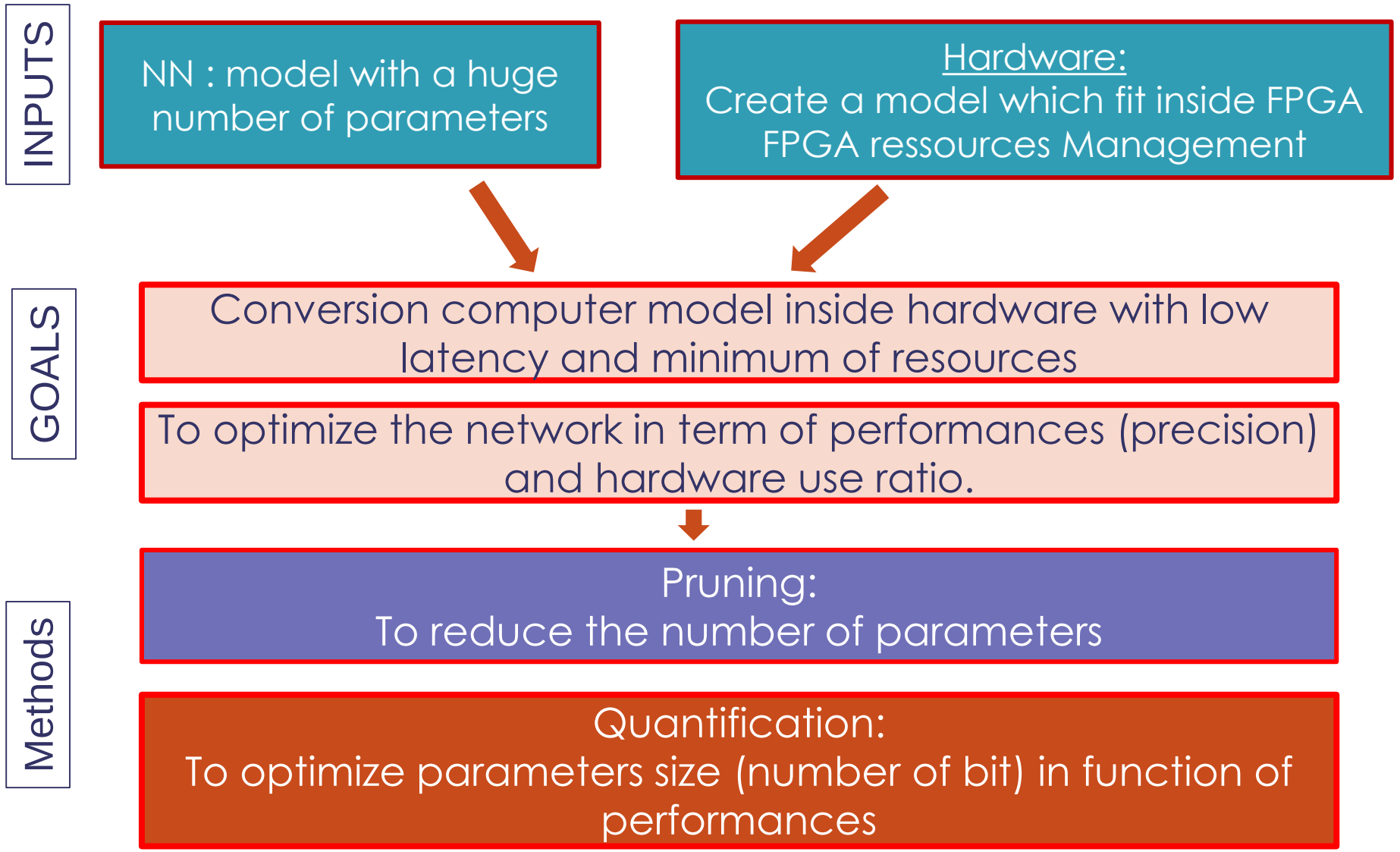
FPGAs



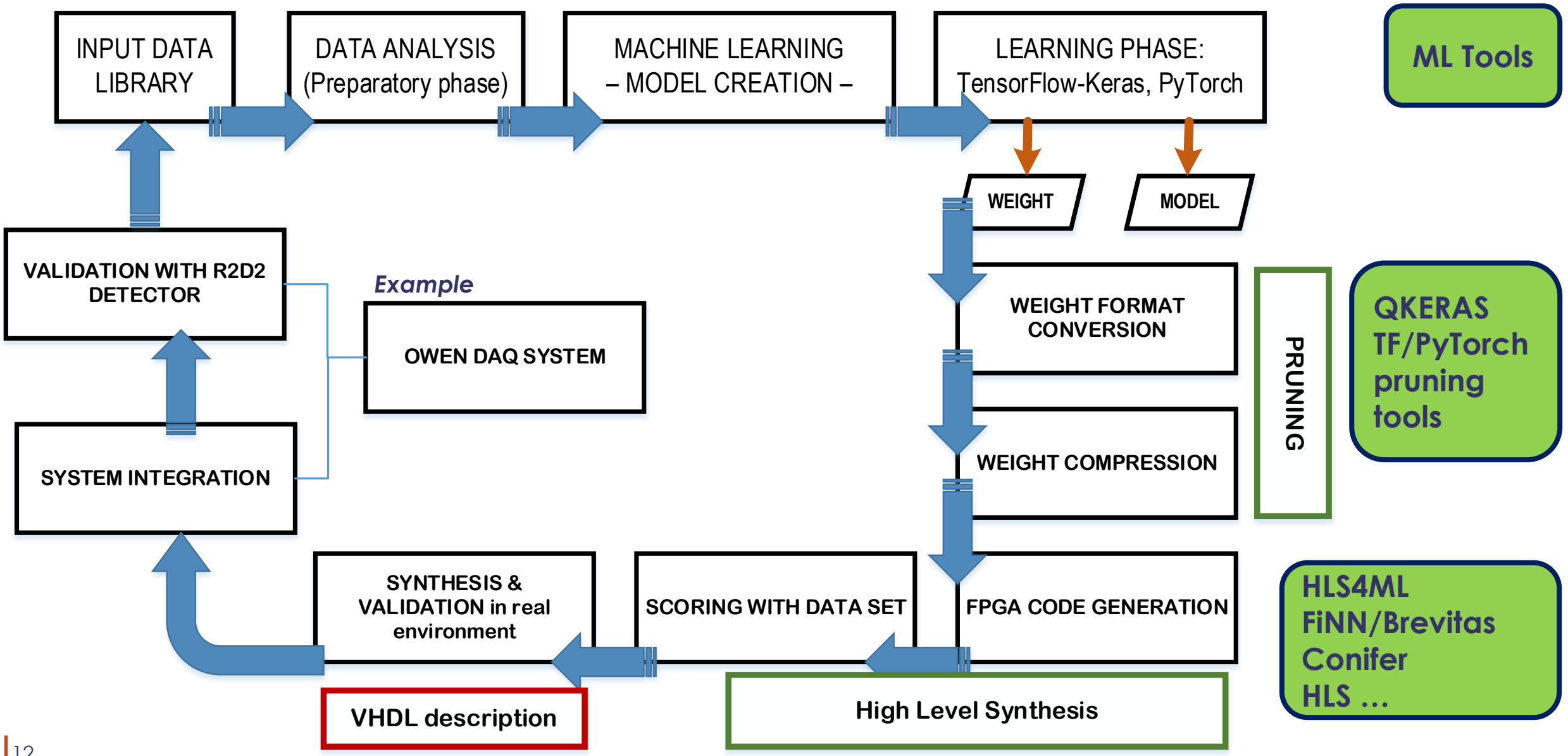
NMC

- **Designing embedded AI systems requires:**
  - *A knowledge of classic AI*
  - *An excellent understanding of hardware architectures*
  - *Specialization by hardware solution*
    - Resource Usage
    - Tools
    - Embedded functions
    - Use of network optimization tools

Embedded approach: a question of optimization



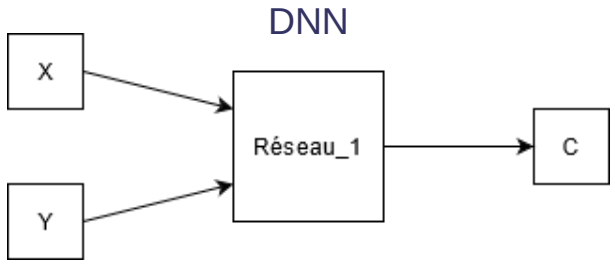
Methodology of design



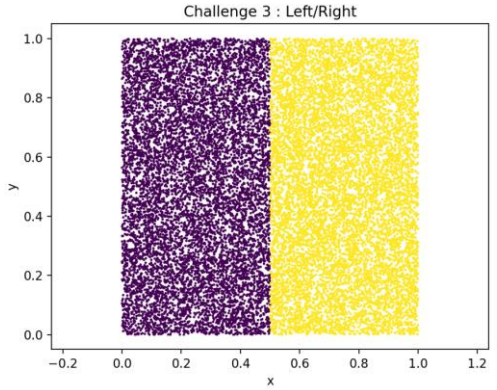
Zynq XcZ7020 & DNN: TF, Keras, QKeras, HLS4ML

XcZ7020: **70€ - 150€**

- Dual-core ARM Cortex-9
- Maximum frequency 667MHz-867MHz
- 2x AXI Master & 2 AXI Slave 32 bits
- 53,2 k LUT
- 220 DSP
- 106,4 k Flip-flop



2x200x100x50x1



Raw implementation <32,11>

```
================================================================
== Performance Estimates
================================================================
+ Timing:
    * Summary:
    +--------+---------+----------+------------+
    |  Clock |  Target | Estimated| Uncertainty|
    +--------+---------+----------+------------+
    |ap_clk  | 5.00 ns | 9.883 ns |   0.62 ns  |
    +--------+---------+----------+------------+

+ Latency:
    * Summary:
    +---------+---------+----------+----------+-----+-----+----------+
    |  Latency (cycles) |  Latency (absolute) |  Interval | Pipeline |
    |   min   |   max   |    min   |    max   | min | max |   Type   |
    +---------+---------+----------+----------+-----+-----+----------+
    |      382|      386| 3.775 us | 3.815 us |  100|  100| dataflow |
    +---------+---------+----------+----------+-----+-----+----------+
```

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+-----------------+---------+-------+--------+-------+-----+
|       Name      | BRAM_18K| DSP48E|   FF   |  LUT  | URAM|
+-----------------+---------+-------+--------+-------+-----+
|DSP              |        -|      -|       -|      -|    -|
|Expression       |        -|      -|       0|   2858|    -|
|FIFO             |        0|      -|    3515|  30732|    -|
|Instance         |      229|   1006|  133648|  95258|    -|
|Memory           |        -|      -|       -|      -|    -|
|Multiplexer      |        -|      -|       -|   6336|    -|
|Register         |        -|      -|     706|      -|    -|
+-----------------+---------+-------+--------+-------+-----+
|Total            |      229|   1006|  137869| 135184|    0|
+-----------------+---------+-------+--------+-------+-----+
|Available        |      280|    220|  106400|  53200|    0|
+-----------------+---------+-------+--------+-------+-----+
|Utilization (%)  |       81|    457|     129|    254|    0|
+-----------------+---------+-------+--------+-------+-----+
```

Model 1

HLS generated with HLS4ML

Model: Keras,
No Pruning:
tf.keras.optimizers,
Quantification:
HLS4ML

```
================================================================
== Performance Estimates
================================================================
+ Timing:
    * Summary:
    +--------+---------+----------+------------+
    |  Clock |  Target | Estimated| Uncertainty|
    +--------+---------+----------+------------+
    |ap_clk  | 5.00 ns | 9.143 ns |   0.62 ns  |
    +--------+---------+----------+------------+

+ Latency:
    * Summary:
    +---------+---------+----------+----------+-----+-----+----------+
    |  Latency (cycles) |  Latency (absolute) |  Interval | Pipeline |
    |   min   |   max   |    min   |    max   | min | max |   Type   |
    +---------+---------+----------+----------+-----+-----+----------+
    |      374|      378| 3.419 us | 3.456 us |  100|  100| dataflow |
    +---------+---------+----------+----------+-----+-----+----------+
```

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+-----------------+---------+-------+--------+-------+-----+
|       Name      | BRAM_18K| DSP48E|   FF   |  LUT  | URAM|
+-----------------+---------+-------+--------+-------+-----+
|DSP              |        -|      -|       -|      -|    -|
|Expression       |        -|      -|       0|   2858|    -|
|FIFO             |        0|      -|    3515|  18114|    -|
|Instance         |      101|    255|   36726|  82229|    -|
|Memory           |        -|      -|       -|      -|    -|
|Multiplexer      |        -|      -|       -|   6336|    -|
|Register         |        -|      -|     706|      -|    -|
+-----------------+---------+-------+--------+-------+-----+
|Total            |      101|    255|   40947| 109537|    0|
+-----------------+---------+-------+--------+-------+-----+
|Available        |      280|    220|  106400|  53200|    0|
+-----------------+---------+-------+--------+-------+-----+
|Utilization (%)  |       36|    115|      38|    205|    0|
+-----------------+---------+-------+--------+-------+-----+
```

Model:
Quantification:
QKeras,
No Pruning:
tf.keras.optimizers,

```
================================================================
== Performance Estimates
================================================================
+ Timing:
    * Summary:
    +--------+---------+----------+------------+
    |  Clock |  Target | Estimated| Uncertainty|
    +--------+---------+----------+------------+
    |ap_clk  | 5.00 ns | 9.143 ns |   0.62 ns  |
    +--------+---------+----------+------------+

+ Latency:
    * Summary:
    +---------+---------+----------+----------+-----+-----+----------+
    |  Latency (cycles) |  Latency (absolute) |  Interval | Pipeline |
    |   min   |   max   |    min   |    max   | min | max |   Type   |
    +---------+---------+----------+----------+-----+-----+----------+
    |      373|      377| 3.410 us | 3.447 us |  100|  100| dataflow |
    +---------+---------+----------+----------+-----+-----+----------+
```

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+-----------------+---------+-------+--------+-------+-----+
|       Name      | BRAM_18K| DSP48E|   FF   |  LUT  | URAM|
+-----------------+---------+-------+--------+-------+-----+
|DSP              |        -|      -|       -|      -|    -|
|Expression       |        -|      -|       0|   2850|    -|
|FIFO             |        0|      -|    3505|  18054|    -|
|Instance         |      100|    253|   36030|  81756|    -|
|Memory           |        -|      -|       -|      -|    -|
|Multiplexer      |        -|      -|       -|   6318|    -|
|Register         |        -|      -|     704|      -|    -|
+-----------------+---------+-------+--------+-------+-----+
|Total            |      100|    253|   40239| 108978|    0|
+-----------------+---------+-------+--------+-------+-----+
|Available        |      280|    220|  106400|  53200|    0|
+-----------------+---------+-------+--------+-------+-----+
|Utilization (%)  |       35|    115|      37|    204|    0|
+-----------------+---------+-------+--------+-------+-----+
```

Model 2

Model: Keras,
Pruning: tf.keras.optimizers,
Quantification: HLS4ML
Optim for resources

```
================================================================
== Performance Estimates
================================================================
+ Timing:
    * Summary:
    +--------+---------+----------+------------+
    |  Clock |  Target | Estimated| Uncertainty|
    +--------+---------+----------+------------+
    |ap_clk  | 5.00 ns | 9.143 ns |   0.62 ns  |
    +--------+---------+----------+------------+

+ Latency:
    * Summary:
    +---------+---------+-----------+-----------+------+------+----------+
    |  Latency (cycles) |   Latency (absolute)  |   Interval  | Pipeline |
    |   min   |   max   |    min    |    max    |  min |  max |   Type   |
    +---------+---------+-----------+-----------+------+------+----------+
    |     2476|     2480| 22.638 us | 22.675 us |  1000|  1000| dataflow |
    +---------+---------+-----------+-----------+------+------+----------+
```

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+-----------------+---------+-------+--------+-------+-----+
|       Name      | BRAM_18K| DSP48E|   FF   |  LUT  | URAM|
+-----------------+---------+-------+--------+-------+-----+
|DSP              |        -|      -|       -|      -|    -|
|Expression       |        -|      -|       0|   2858|    -|
|FIFO             |        0|      -|    3515|  18314|    -|
|Instance         |       25|     27|   29254|  48722|    -|
|Memory           |        -|      -|       -|      -|    -|
|Multiplexer      |        -|      -|       -|   6336|    -|
|Register         |        -|      -|     706|      -|    -|
+-----------------+---------+-------+--------+-------+-----+
|Total            |       25|     27|   33475|  76230|    0|
+-----------------+---------+-------+--------+-------+-----+
|Available        |      280|    220|  106400|  53200|    0|
+-----------------+---------+-------+--------+-------+-----+
|Utilization (%)  |        8|     12|      31|    143|    0|
+-----------------+---------+-------+--------+-------+-----+
```

HLS generated with HLS4ML

Model 3 - #PRAGMA ARRAY_PARTITION dim=3

```
================================================================
== Performance Estimates
================================================================
+ Timing:
    * Summary:
    +--------+---------+----------+------------+
    |  Clock |  Target | Estimated| Uncertainty|
    +--------+---------+----------+------------+
    |ap_clk  | 5.00 ns | 9.062 ns |   0.62 ns  |
    +--------+---------+----------+------------+

+ Latency:
    * Summary:
    +---------+---------+----------+----------+-----+-----+----------+
    |  Latency (cycles) |  Latency (absolute) |  Interval | Pipeline |
    |   min   |   max   |    min   |    max   | min | max |   Type   |
    +---------+---------+----------+----------+-----+-----+----------+
    |      199|      203| 1.803 us | 1.840 us |   50|   50| dataflow |
    +---------+---------+----------+----------+-----+-----+----------+
```

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+-----------------+---------+-------+--------+--------+-----+
|       Name      | BRAM_18K| DSP48E|   FF   |   LUT  | URAM|
+-----------------+---------+-------+--------+--------+-----+
|DSP              |        -|      -|       -|       -|    -|
|Expression       |        -|      -|       0|    2858|    -|
|FIFO             |        0|      -|    3515|   14157|    -|
|Instance         |      171|     10|   35019|  207910|    -|
|Memory           |        -|      -|       -|       -|    -|
|Multiplexer      |        -|      -|       -|    6336|    -|
|Register         |        -|      -|     706|       -|    -|
+-----------------+---------+-------+--------+--------+-----+
|Total            |      171|     10|   39240|  231261|    0|
+-----------------+---------+-------+--------+--------+-----+
|Available        |      280|    220|  106400|   53200|    0|
+-----------------+---------+-------+--------+--------+-----+
|Utilization (%)  |       61|      4|      36|     434|    0|
+-----------------+---------+-------+--------+--------+-----+
```

Model 4: REMOVE ALL #PRAGMA in HLS code

```
================================================================
== Performance Estimates
================================================================
+ Timing:
    * Summary:
    +--------+---------+----------+------------+
    |  Clock |  Target | Estimated| Uncertainty|
    +--------+---------+----------+------------+
    |ap_clk  | 5.00 ns | 7.144 ns |   0.62 ns  |
    +--------+---------+----------+------------+

+ Latency:
    * Summary:
    +---------+---------+----------+----------+-------+-------+----------+
    |  Latency (cycles) |  Latency (absolute) |    Interval   | Pipeline |
    |   min   |   max   |    min   |    max   |  min  |  max  |   Type   |
    +---------+---------+----------+----------+-------+-------+----------+
    |    91251|    91255| 0.652 ms | 0.652 ms |  50000|  50000| dataflow |
    +---------+---------+----------+----------+-------+-------+----------+
```

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+-----------------+---------+-------+--------+-------+-----+
|       Name      | BRAM_18K| DSP48E|   FF   |  LUT  | URAM|
+-----------------+---------+-------+--------+-------+-----+
|DSP              |        -|      -|       -|      -|    -|
|Expression       |        -|      -|       0|    238|    -|
|FIFO             |        0|      -|     255|   1326|    -|
|Instance         |       25|     27|   33689|  36552|    -|
|Memory           |        5|      -|       0|      0|    0|
|Multiplexer      |        -|      -|       -|    486|    -|
|Register         |        -|      -|      56|      -|    -|
+-----------------+---------+-------+--------+-------+-----+
|Total            |       30|     27|   34000|  38602|    0|
+-----------------+---------+-------+--------+-------+-----+
|Available        |      280|    220|  106400|  53200|    0|
+-----------------+---------+-------+--------+-------+-----+
|Utilization (%)  |       10|     12|      31|     72|    0|
+-----------------+---------+-------+--------+-------+-----+
```

At one point, we need to optimize HLS code
At one point, we could develop directively in VHDL

Discussion about FPGA: Spatial Accelerators

- **Zynq + HLS4ML: io_parallel / io_stream /Reusefactor / Resource / Latency**

A question of budget

	ARTY-Z7 CH1	ARTY Z7 CH3	ARTY Z7 CH3 Optimisé	ZCU102 CH3	CH4	ZCU102 CH4 Qbit<16,2>	ZCU102 CH5	ZCU102 CH7 Optim Ressource	ZCU102 CH7 Optim Latence	ARTY Z7 CH7 Optim HLS Stream
Nombre de cellules	**16**	**25801**	**25801**	**25801**	**658951**	**658951**	**156951**	**156951**	**156951**	**156951**
Horloge de reference	4,166ns	9,408ns	9,408ns	4,396ns		4,369ns	4,369ns	4,369ns	4,028ns	9,410ns
Temps de latence	70ns	19,804us	19,804us	2,510us		5,510us	5,510us	10,010us	10,015us	141us
BRAM	0%	41%	8%	6%		36%	18%	9%	15%	72%
DSP48E	21%	115%	11%	10%		59%	59%	12%	24%	6%
FF	2%	85%	28%	10%		28%	22%	32%	53%	167%
LUT	1%	280%	68%	46%		103%	113%	81%	104%	423%
URAM	0%	0%	0%	0%		0%	0%	0%	0%	0%

depend on VITIS-HLS #pragma
The way HLS handles vector/matrix before DSP

€ ⟷ €€€

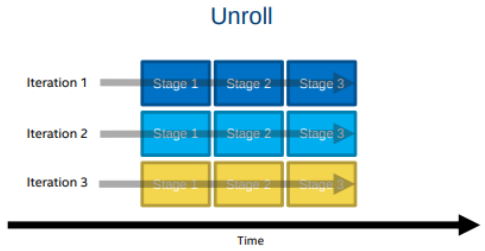
- **Intel Arria 10 + Intel HLS (expert HLS) : €€**

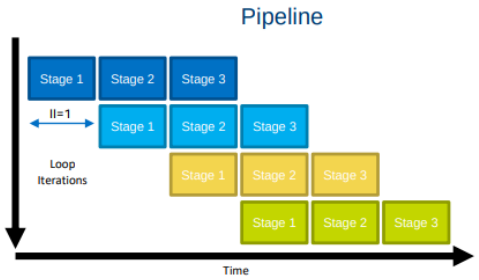
	ALUTs	FFs	RAMs	MLABs	DSPs
Ch1 vanilla	602 (0%)	547 (0%)	4 (0%)	2 (0%)	1.5 (0%)
Ch1 pipeline	610 (0%)	624 (0%)	4 (0%)	5 (0%)	1.5 (0%)
Ch1 unroll	515 (0%)	245 (0%)	4 (0%)	1 (0%)	0 (0%)
Ch1 u+p	515 (0%)	245 (0%)	4 (0%)	1 (0%)	0 (0%)

	ALUTs	FFs	RAMs	MLABs	DSPs
Ch3 vanilla	1 408 (0%)	1 809 (0%)	30 (1%)	12 (0%)	2.5 (0%)
Ch3 pipeline	2 093 (0%)	4 460 (0%)	32 (1%)	48 (0%)	2.5 (0%)
Ch3 unroll	118 041 (14%)	36 737 (2%)	5 (0%)	37 (0%)	0 (0%)
Ch3 u+p	27 524 (3%)	42 546 (2%)	1 855 (68%)	298 (1%)	75 (5%)

	ALUTs	FFs	RAMs	MLABs	DSPs
Ch4 vanilla	4 442 (0%)	6 415 (0%)	355 (1%)	20 (0%)	3.5 (0%)
Ch4 pipeline	5 555 (0%)	10 899 (0%)	362 (1%)	113 (0%)	3.5 (0%)
Ch4 unroll	Problème d'implémentation				
Ch4 u+p					

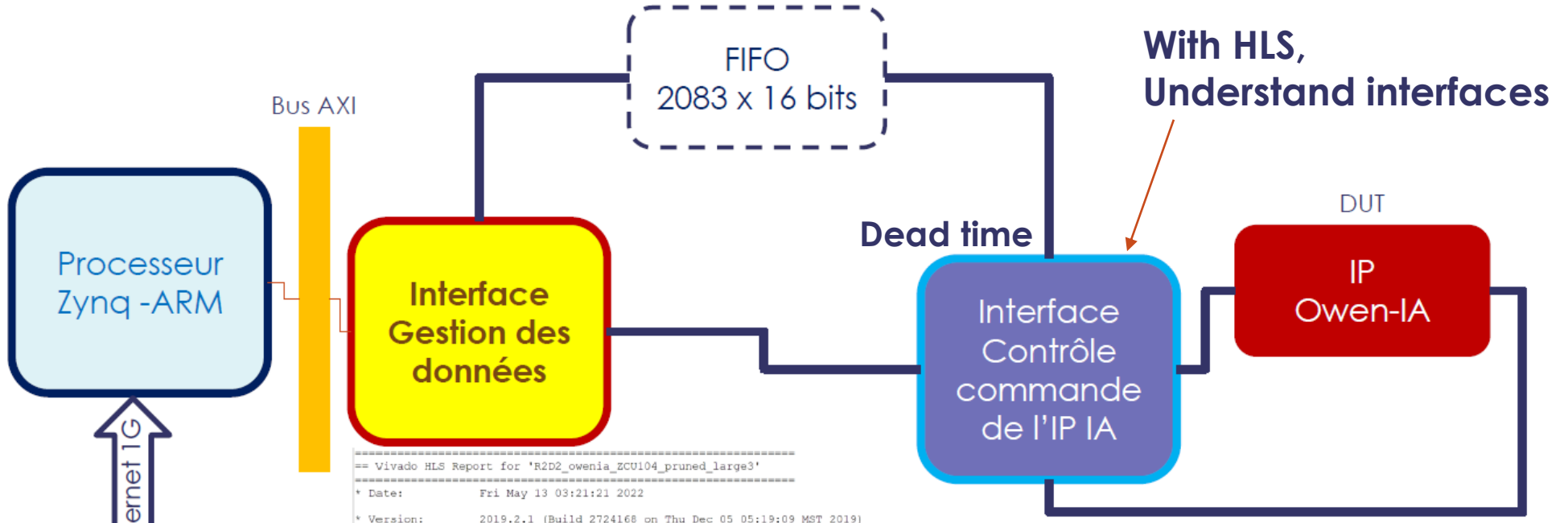
	ALUTs	FFs	RAMs	MLABs	DSPs
Ch5 vanilla	1 669 (0%)	2 193 (0%)	32 (1%)	16 (0%)	2.5 (0%)
Ch5 pipeline	2 617 (0%)	6 074 (0%)	35 (1%)	76 (0%)	2.5 (0%)
Ch5 unroll	569 259 (67%)	196 051 (11%)	6 (0%)	40 (0%)	0 (0%)
Ch5 u+p	17 560 (2%)	23 244 (1%)	507 (19%)	237 (1%)	50.5 (3%)



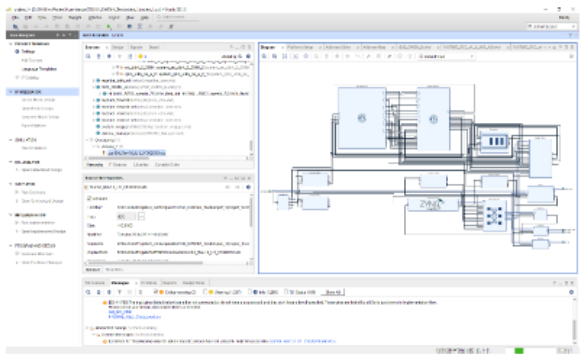


Question of HLS optimisation

Owen Test in real condition



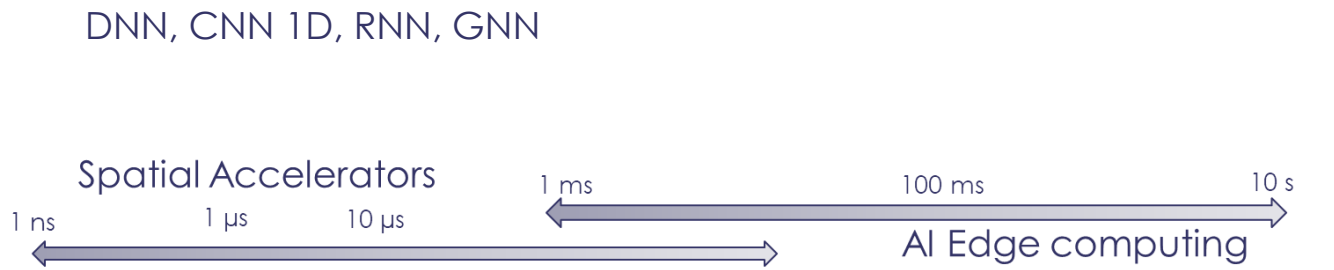

```
==========================================================
== Vivado HLS Report for 'R2D2_owenia_ZCU104_pruned_large3'
==========================================================
* Date:           Fri May 13 03:21:21 2022

* Version:        2019.2.1 (Build 2724168 on Thu Dec 05 05:19:09 MST 2019)
* Project:        R2D2_owenia_ZCU104_pruned_large3_prj
* Solution:       solution1
* Product family: zynquplus
* Target device:  xczu7ev-ffvc1156-2-e


==========================================================
== Performance Estimates
==========================================================
+ Timing:
    * Summary:
    +--------+---------+----------+------------+
    |  Clock |  Target | Estimated| Uncertainty|
    +--------+---------+----------+------------+
    |ap_clk  | 5.00 ns | 4.339 ns |   0.62 ns  |
    +--------+---------+----------+------------+

+ Latency:
    * Summary:
    +---------+---------+-----------+-----------+------+------+----------+
    | Latency (cycles)  |   Latency (absolute)  |   Interval  | Pipeline |
    |   min   |   max   |    min    |    max    |  min |  max |   Type   |
    +---------+---------+-----------+-----------+------+------+----------+
    |     2086|     2087| 10.430 us | 10.435 us |  2087|  2088| dataflow |
```



Features with others components

Microcontrollers
AI edge computing
RISC-V Architecture

Type	Company	Components	Dev Boards	Pictures	Soft Tools
AI Edge computing	AMD Xilinx	Zynq	ARTY Z7 PYNQ ZCU104 SCU102		VITIS-AI
Spatial acceleration					Tensorflow/Keras + HLS4ML PyTorch + FiNN+Brevitas
					
AI Edge computing	AMD Xilinx	VERSAL	VK290		HLS+VITIS-AI
AI Edge computing	nVidia	Jetson	Nano TX2 Xavier NX AGX Xavier Orin		Jetpack SDK + DeepStream SDK + TensorRT
					
Spatial acceleration	Intel	AGILEX ARRIA	DE10-AGILEX ARRIA 10 GX		Intel HLS
AI Edge computing					FPGA AI Suite OpenVino
Spatial acceleration	Brainchip	AKD1000	Akida PCIe		Akida MetaTF ML framework
AI Edge computing	SiPEED	Kendryte210 RISC-V	MAIXDUiNO		SiPEED SDK + micropython
AI Edge computing	ST Microelectronic	STM32	Nucleo64F411		CubeMX + Cube.AI
AI Edge computing	Analog Devices	MAX78000	MAX78000EVKIT		Maxim Micros SDK



One component = one sdk Tools

Work package

- **Designing embedded AI systems requires:**
  - *Need a budget to dev kit purchase (devkit from ~200€ to 20k€)*
  - *Understand where to put AI model in the system*
  - *Understand the performances we need in term of*
    - Resources
    - Latences
    - Architectures
  - Constraint AI model

- Incorporated AI near data generation is a full time job on the project
- Model need to be constraint depending where to use it