







Centre de physique des particules de Marseille

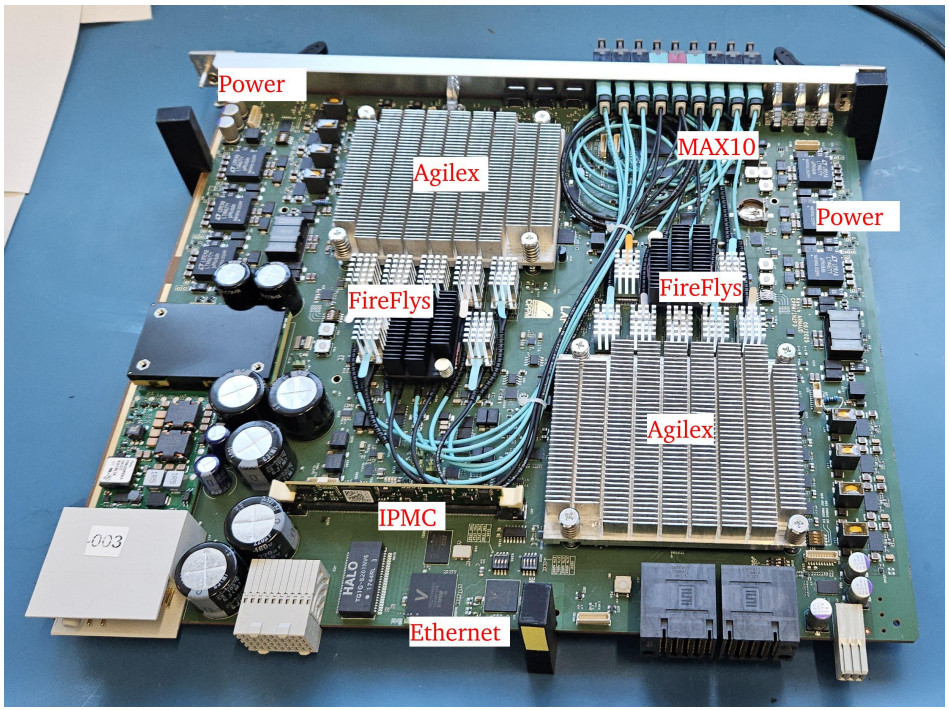
Implémentation de réseaux de neurones sur FPGA Agilex 7 pour le calorimètre à argon liquide de l'expérience ATLAS.

October 12, 2026

Etienne FORTIN,Tristan Panhelleux

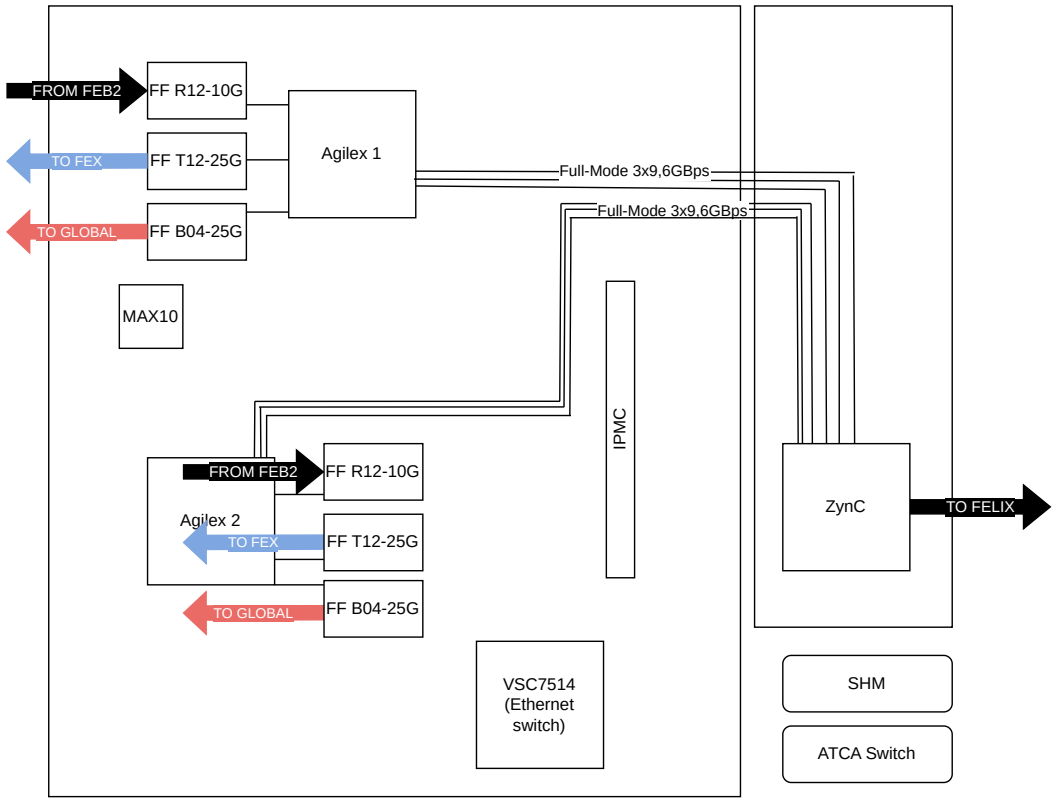
LASP

- Une nouvelle carte back-end en charge de la reconstruction de l'énergie.
- La carte LASP (CPPM) équipée de 2 FPGA Altera Agilex.
  - Carte démonstratrice équipée de FPGA Stratix.
- Des ressources suffisantes pour envisager l'implémentation de réseaux de neurones.

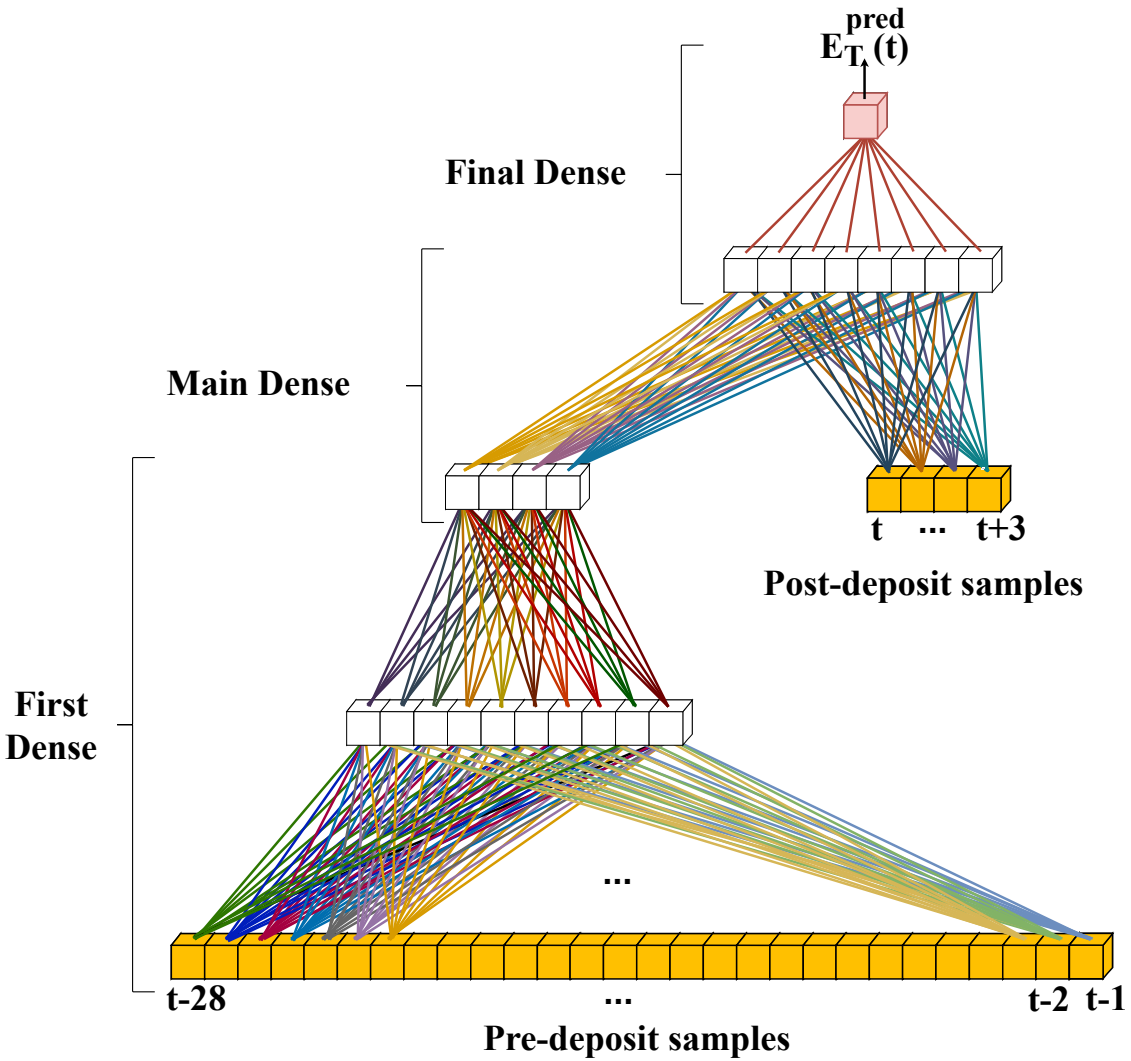


Flux de données

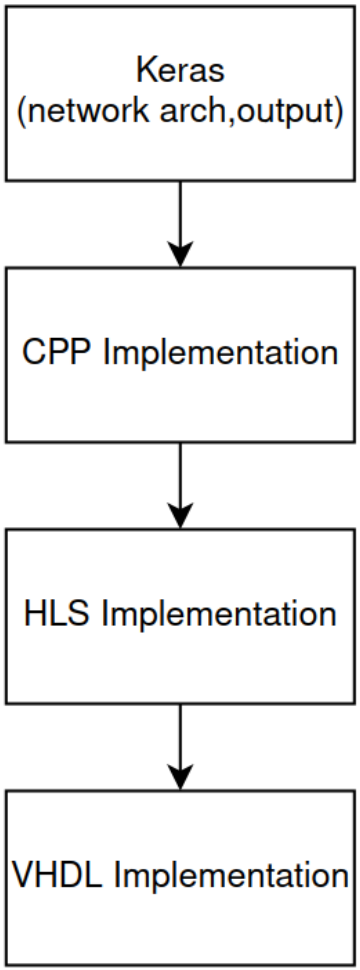
- Depuis les cartes Front-End
  - 3 x 22 x 10 Gbit/s par FPGA
- Vers la chaîne de déclenchement (Trigger)
  - Global : 3 x 25 Gbit/s par FPGA
  - fFEX : 24 x 25 Gbit/s par FPGA
- Vers la chaîne d'acquisition
  - 3 x 9,6 Gbit/s par FPGA



Dense network



Implementation flow



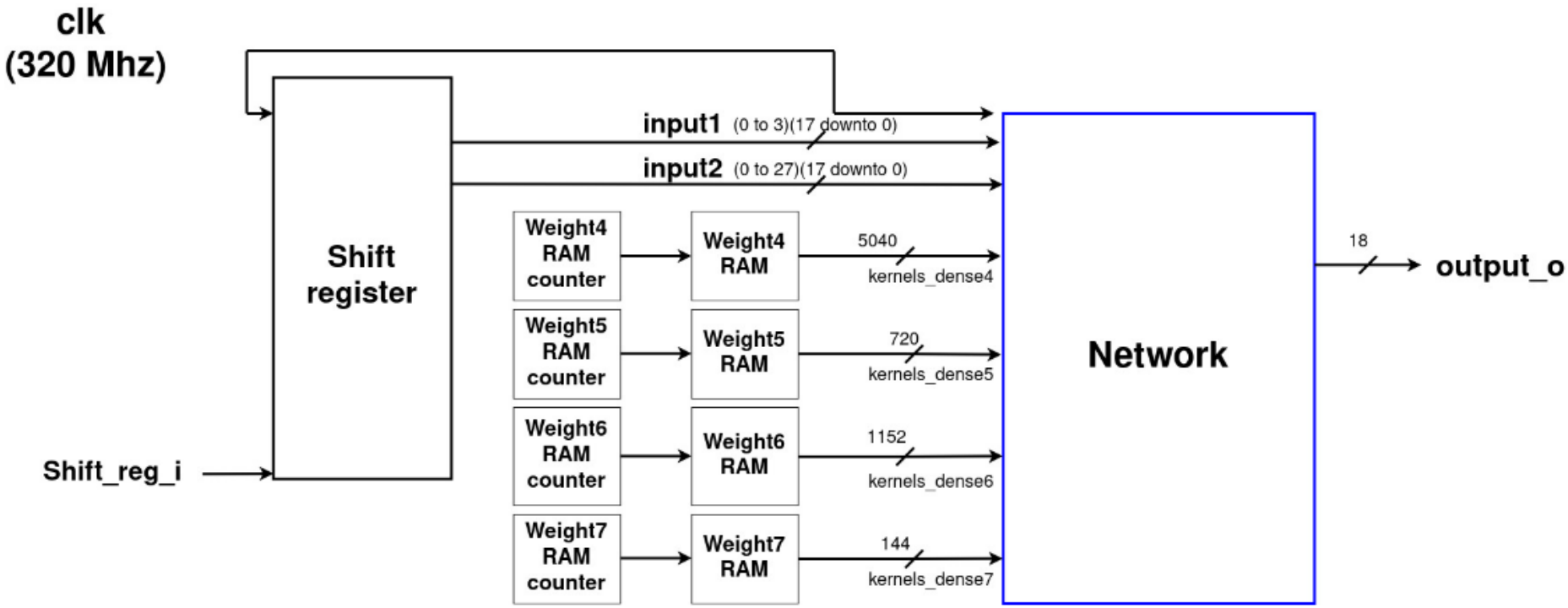
- Check that the output if the same as keras
- Validate the understanding of the network architecture

- Study quantization
- Architecture optimisation
- First insight on the ressource usage

- Ressource usage
- Fmax
- Placement optimisation

Generateur python qui cree tout les fichiers a partir du reseau python.

Architecture



- Multiplexage $\times 8$ (320 MHz) pour disposer d'un nombre suffisant de DSP.
  - Nécessite de changer les poids à chaque cycle d'horloge.

Générateur

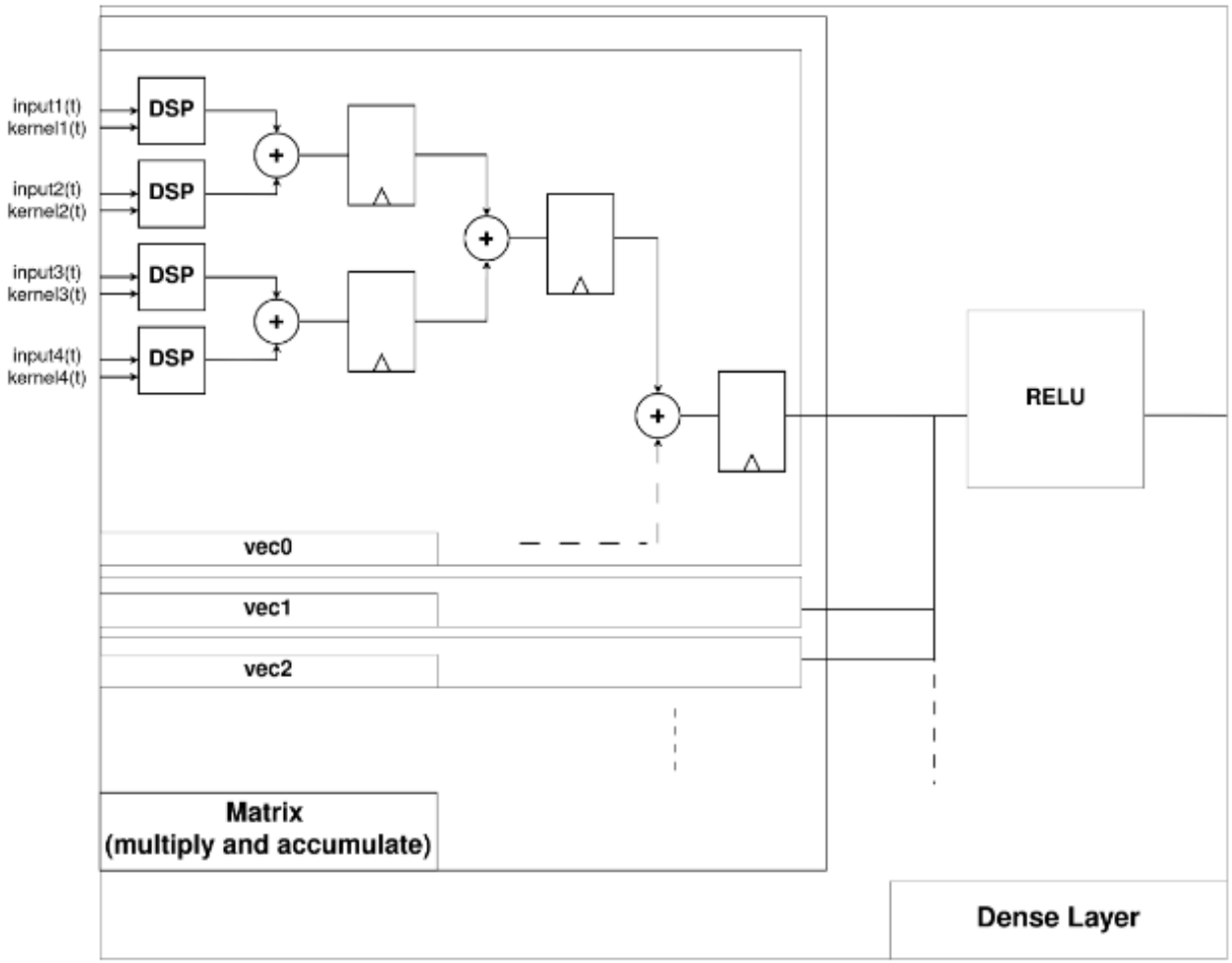
- Python + Jinja2.
- Modèle Keras/TensorFlow et configuration des options d'implémentation.
- Génération :
  - C++ en virgule flottante.
  - C++ en virgule fixe (approche similaire à la HLS).
  - Pour l'implémentation sur cible :
    - VHDL : blocs du réseau, RAM et registres à décalage.
    - Fichiers MIF pour l'initialisation des mémoires.
    - Projet Quartus.
    - Banc de test Cocotb.
  - Graphiques de performances (ROOT).

Utilisation ressources

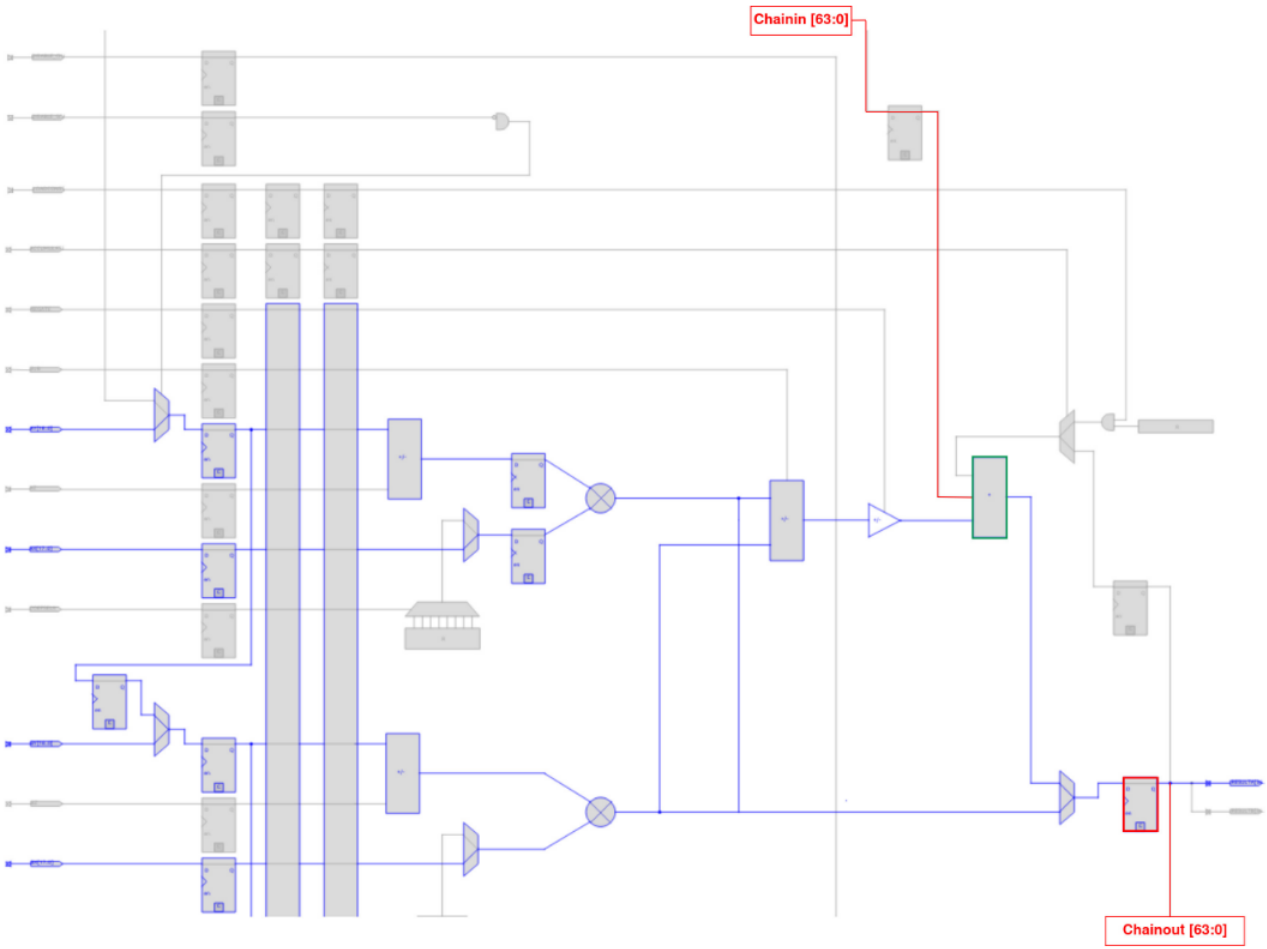
Ressource	Implementation de base
M20K	8496 (42.67 %)
DSP	9408 (73.55 %)
ALM	99768 (7.27 %)

- DSP OK , principalement utiliser par ce bloc
- ALM OK
- M20K: NOK doit etre utiliser par un autre bloc du firmware

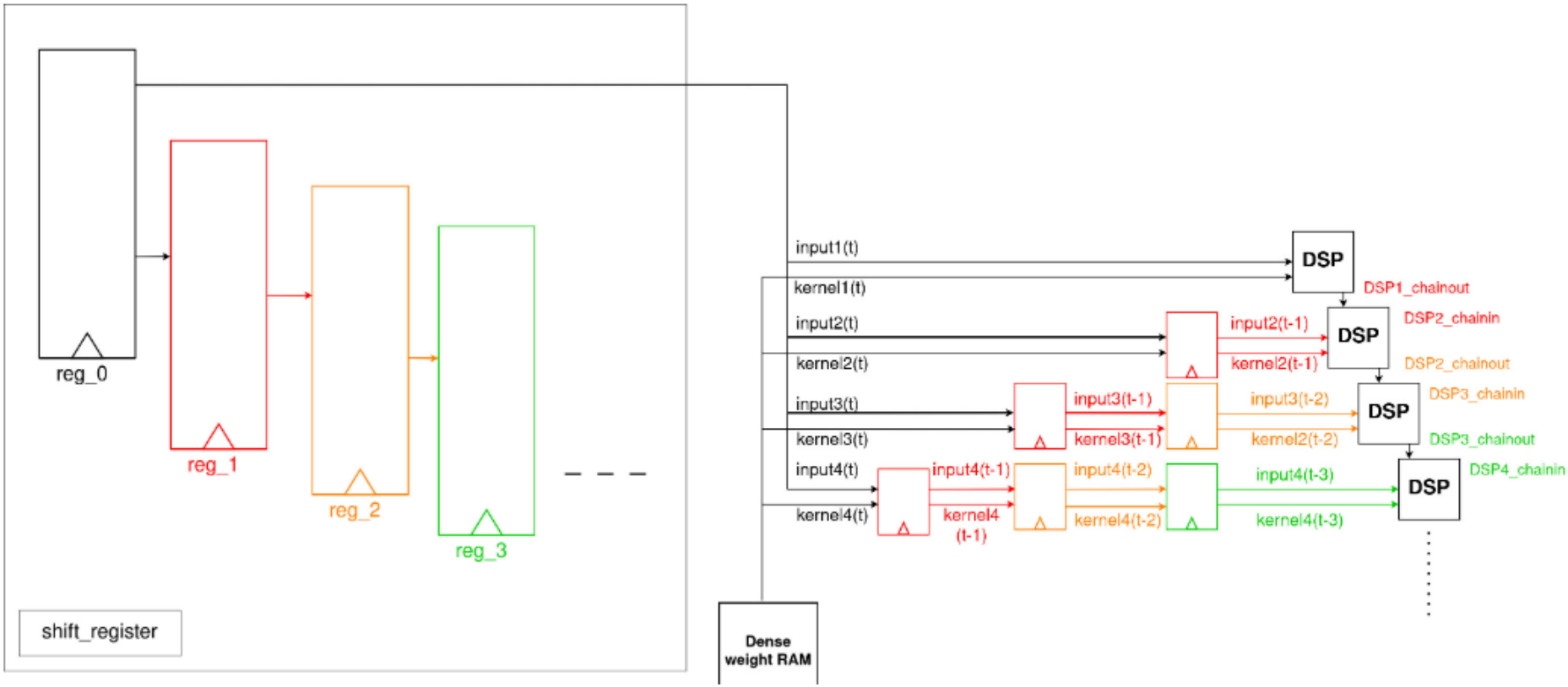
VHDL Optimisation: DSP chain



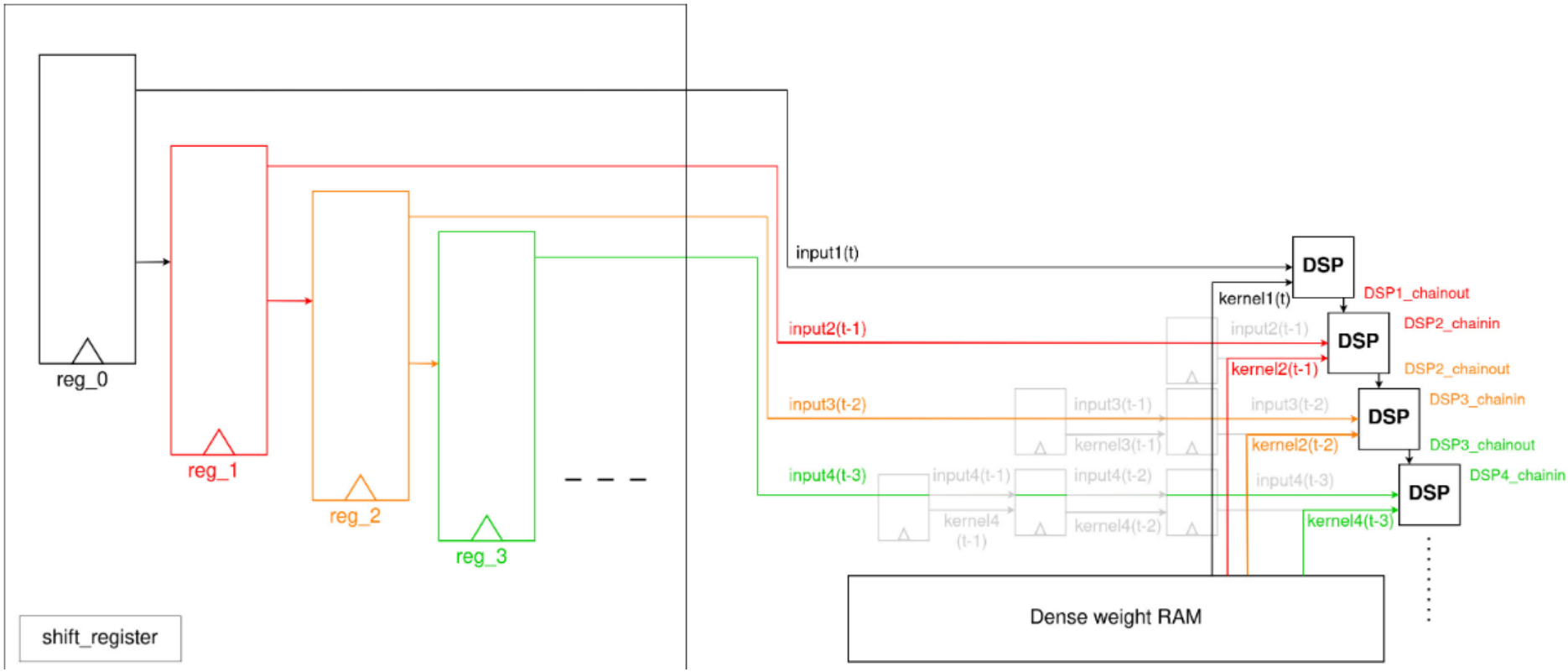
VHDL Optimisation: DSP chain



VHDL Optimisation: DSP chain



Optimisation VHDL : chaîne DSP



- Décalage interne de la mémoire contenant les poids.
- Chaînage possible sur les deux premières couches (les plus volumineuses).

Optimisation VHDL : chaîne DSP

Ressource	Implémentation de base	Avec chaînage
M20K	8496 (42.67 %)	8496 (42.67 %)
DSP	9408 (73.55 %)	9408 (73.55 %)
ALM	99768 (7.27 %)	53688 (3.91 %)

Optimisation VHDL : M20K

- 40 % du FPGA occupés par les RAM contenant les poids.
  - Environ 5000 bits par canal.
  - Mais seulement 0.3 % de la capacité de chaque bloc RAM est utilisée.
- Le limitant est la taille du bus de sortie des RAMs
- Test: Implementation des poids en MLAB (utilisation des LUT des ALM en tant que memoire)

VHDL Optimisation: M20K

Ressource	Implementation de base	Chainage	Chainage + MLAB
M20K	8496 (42.67 %)	8496 (42.67 %)	0
DSP	9408 (73.55 %)	9408 (73.55 %)	9408 (73.55 %)
ALM	99768 (7.27 %)	53688 (3.91%)	223032 (16.26%)

Conclusion et perspectives

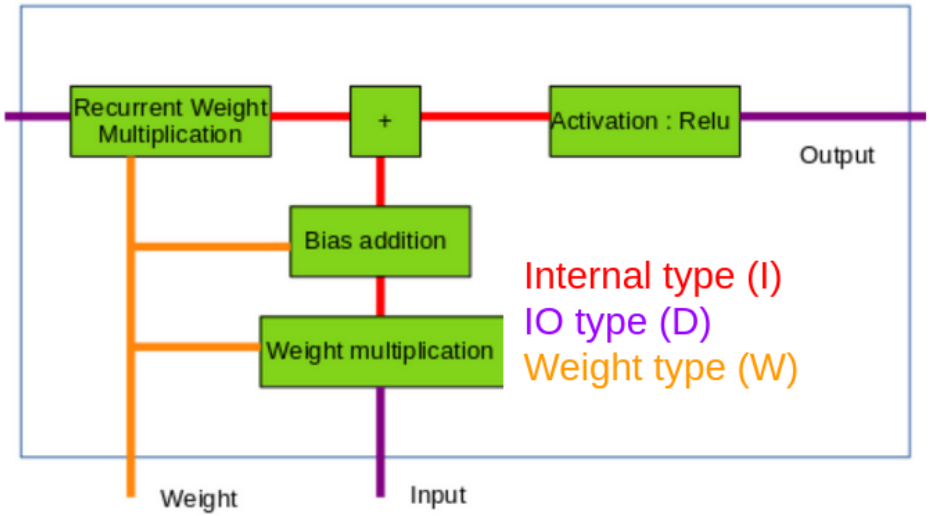
- Les calculs du réseau sont réalisés à 95% dans les DSP.
  - Utilisation d'ALM pour le registre à décalage.
- Le stockage des poids est la principale source de consommation des ressources.
- Étude en cours de l'impact de la réduction de la précision des poids sur les performances :
  - QAT (Quantization-Aware Training).
  - Précision des poids différente selon les couches.

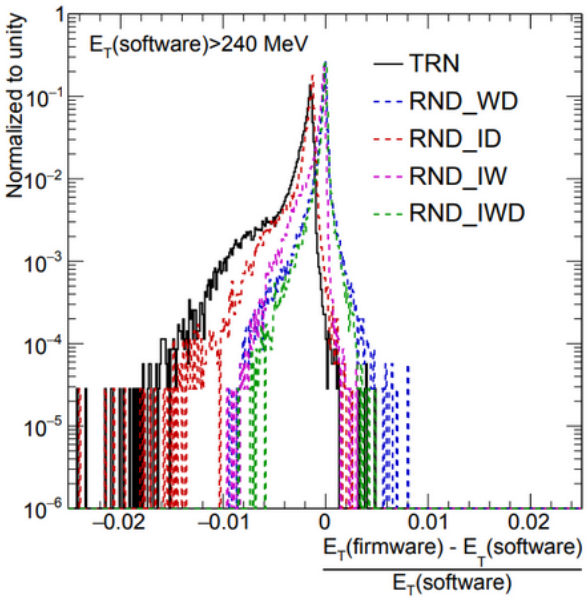
Questions ?

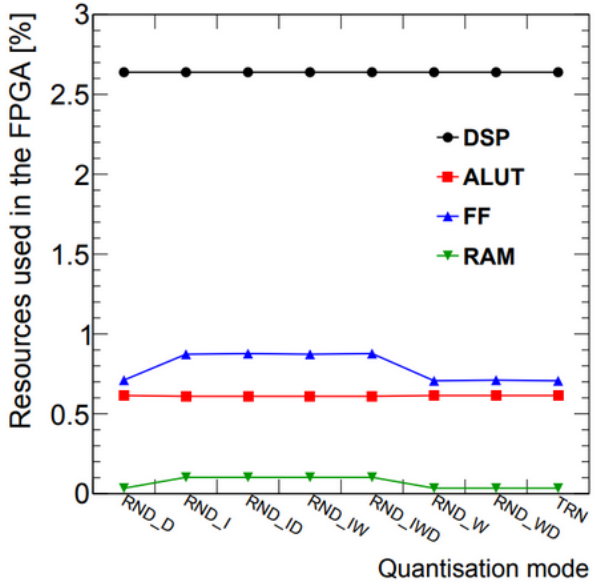
Backup slides

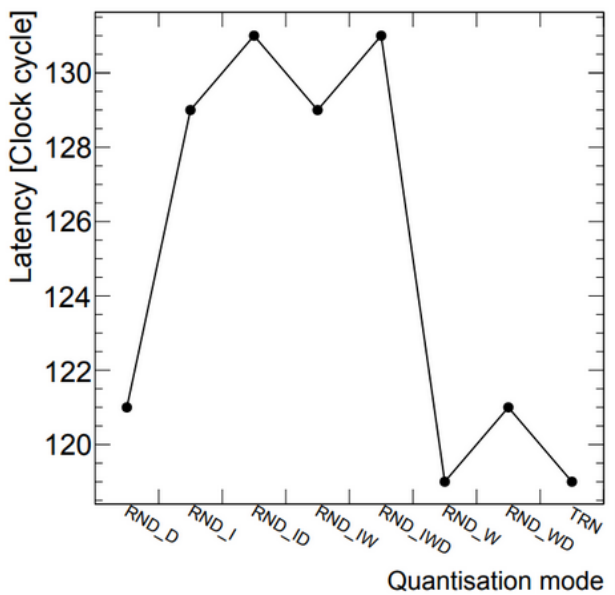
HLS optimisation: Truncation vs Rounding in detail

- Compromise between resolution and resource usage and latency
  - Truncation of I/O and Internal types leads to important reduction of latency with small impact on energy resolution
  - Weight type rounded in software: no impact on latency
- Truncation used in firmware



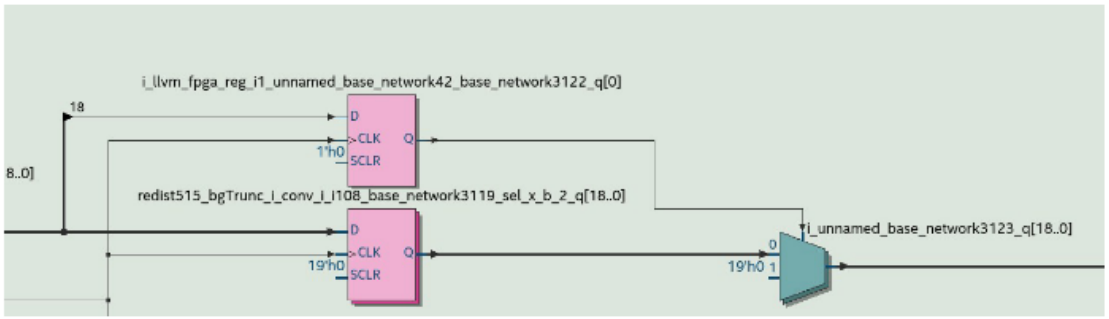






Optimisation: Activation function: RELU

- Activation functions are complex mathematical operation
- Not possible into firmare: use of Look up table
- But for each activation function we need:RAM for it....
- There is one exception the RELU: $f(x) = x$ if $x > 0$ else $f(x) = 0$



- Only logic element, no mathematical function as signed bit is used
- With a leaky relu $f(x) = x$ if $x > 0$ else $f(x) = \alpha x$ with α a power of 2 is the same