Prospectives 2023

Portage de réseaux de neurones sur FPGAs

18/10/2023

Geoffrey GALBIT, groupe eDaq

FPGA vs GPU

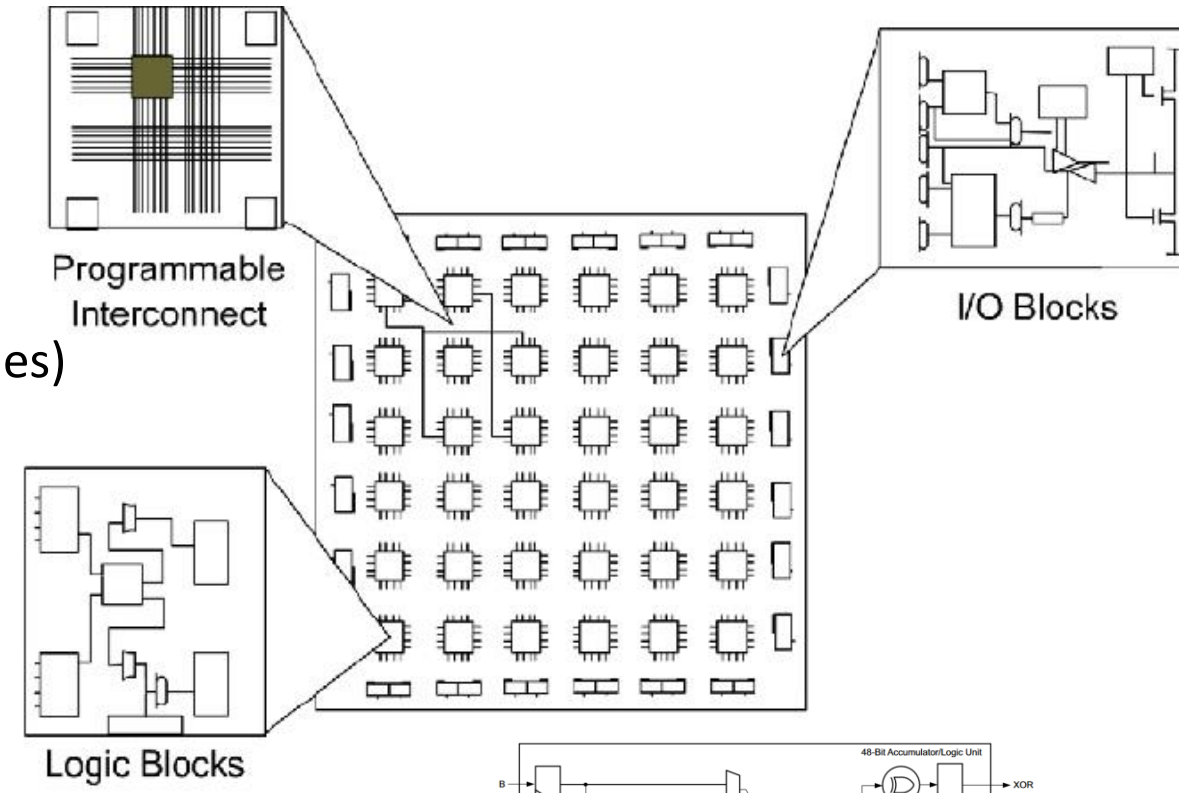
FPGA = Matrice d'éléments électroniques reconfigurables

Avantages:

- Faible latence (pas d'overhead software)
- Flexible et versatile
- Parallélisme total (dans la limite des ressources disponibles)
- Nombre d'I/O -> parallélisme en entrées/sorties
- Consommation limitée

Points faibles:

- Développement long et complexe
- Ressources (généralistes et spécialisées) limitées
- Faible fréquence de fonctionnement -> non-compétitif en séquentiel



DSP

Réseaux de neurones

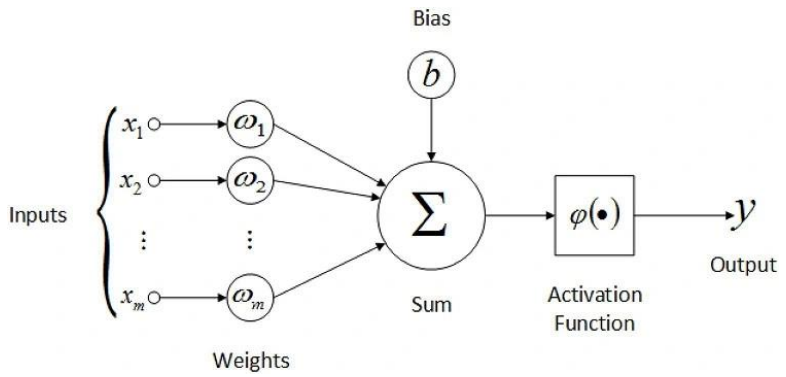
C’est un approximateur universel

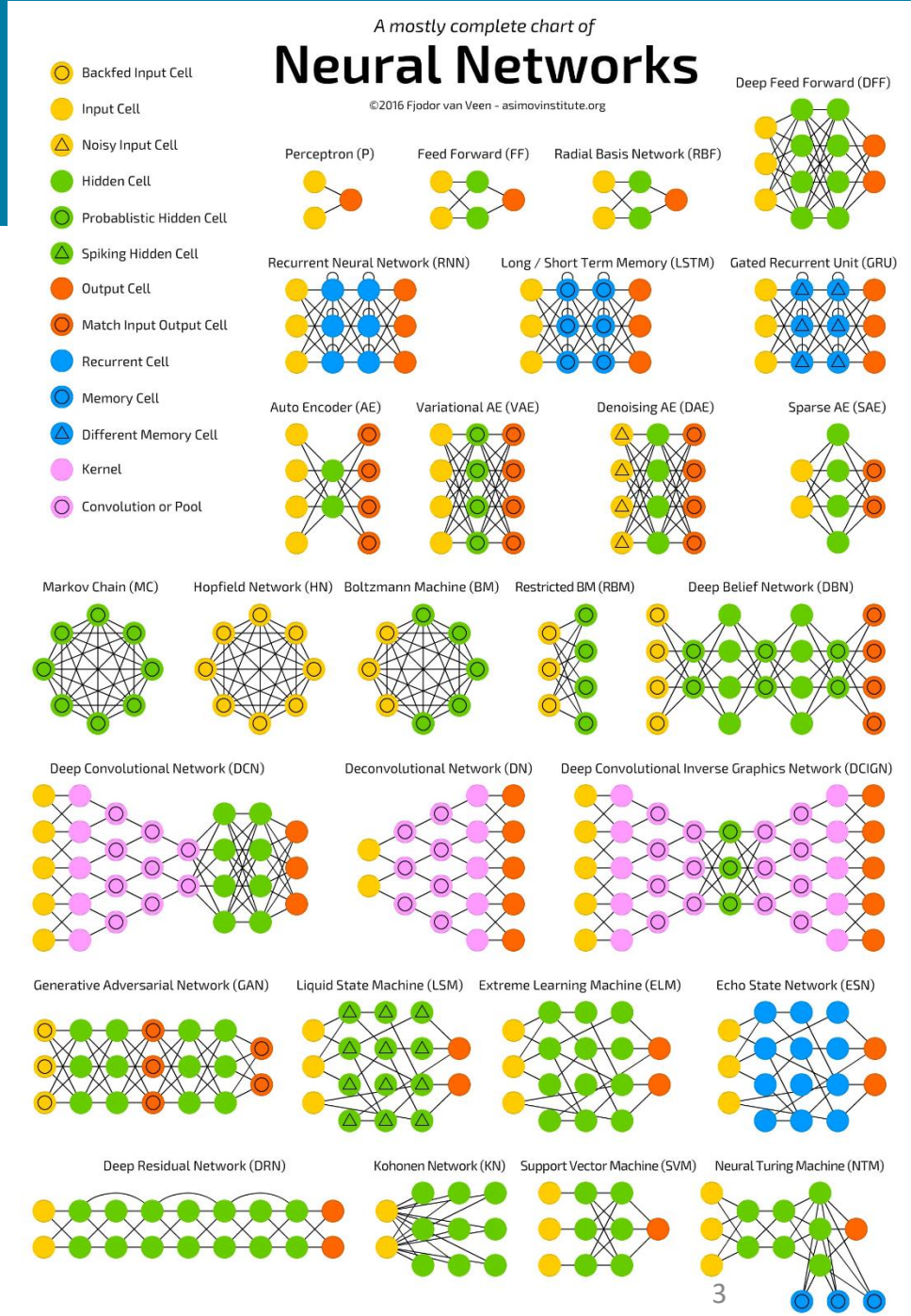
Création de l’espace latent:

- Choix de la topologie et du type de neurones utilisés
- Choix de la fonction d’activation
- Apprentissage -> définition de la partie variable du réseau

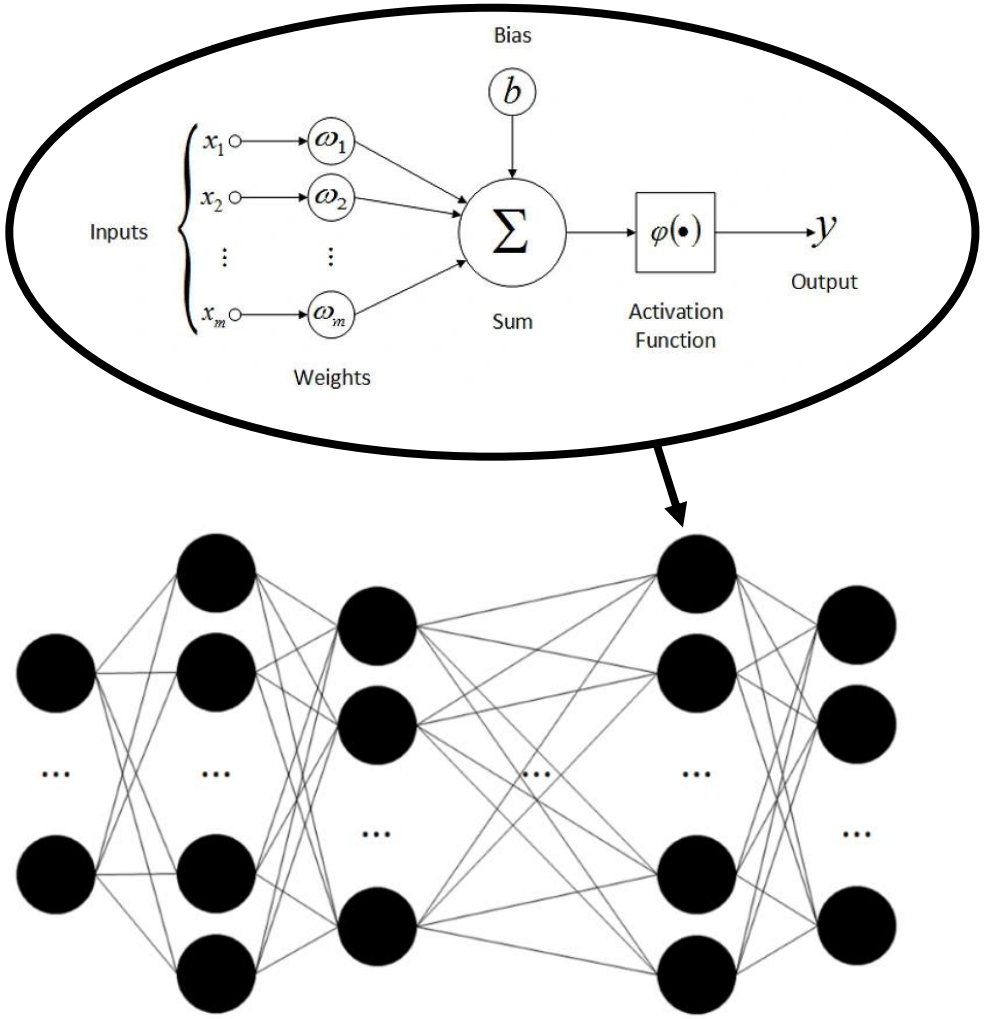
Evaluation:

- Projection entrée -> espace latent -> sortie





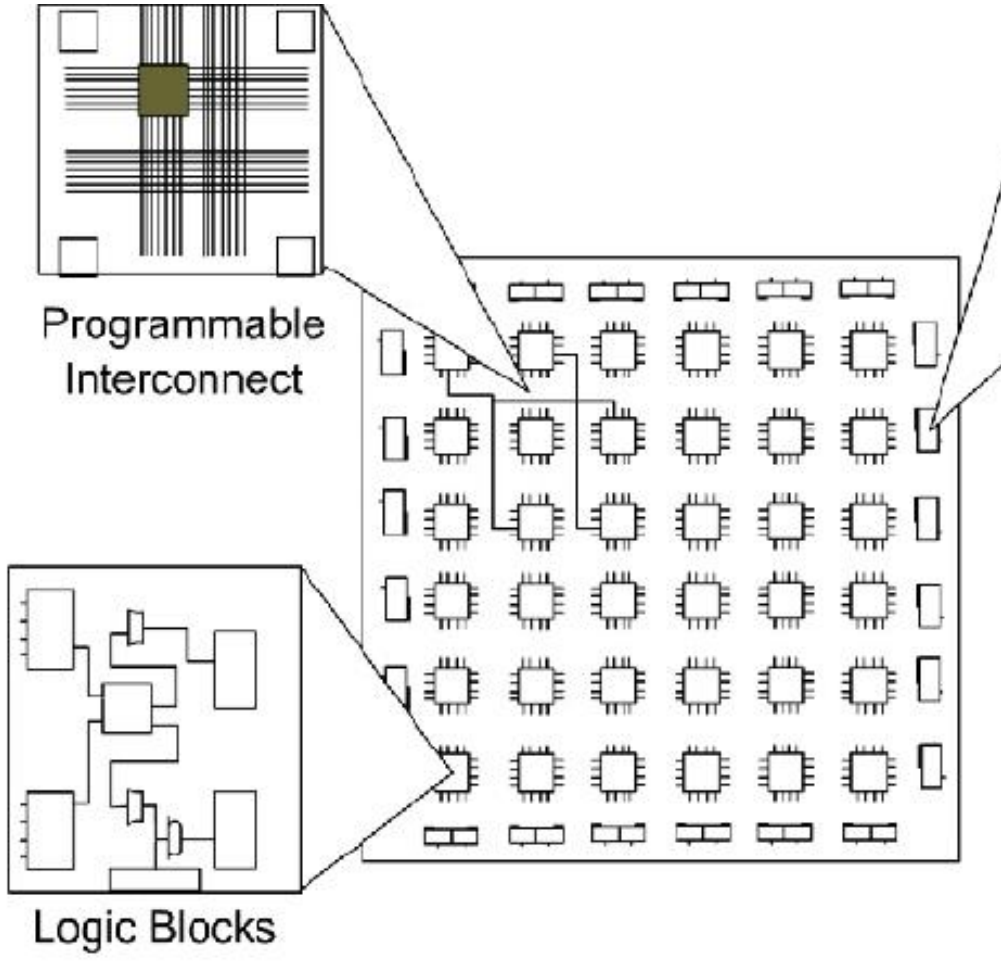
Portage de NN sur FPGA



Mapping direct peu efficace

???

La variabilité du NN doit s'adapter aux degrés de libertés du FPGA



Pour avancer, il faut se poser les bonnes questions

Approche utilisateur ou approche concepteur ?

Approche utilisateur:

- Adaptée pour l'intégration de petits réseaux avec une architecture standard
- Utilisation d'outils de portage spécialisés (hls4ml)

Exemple: Intégration d'un auto-encoder pour la compression de données

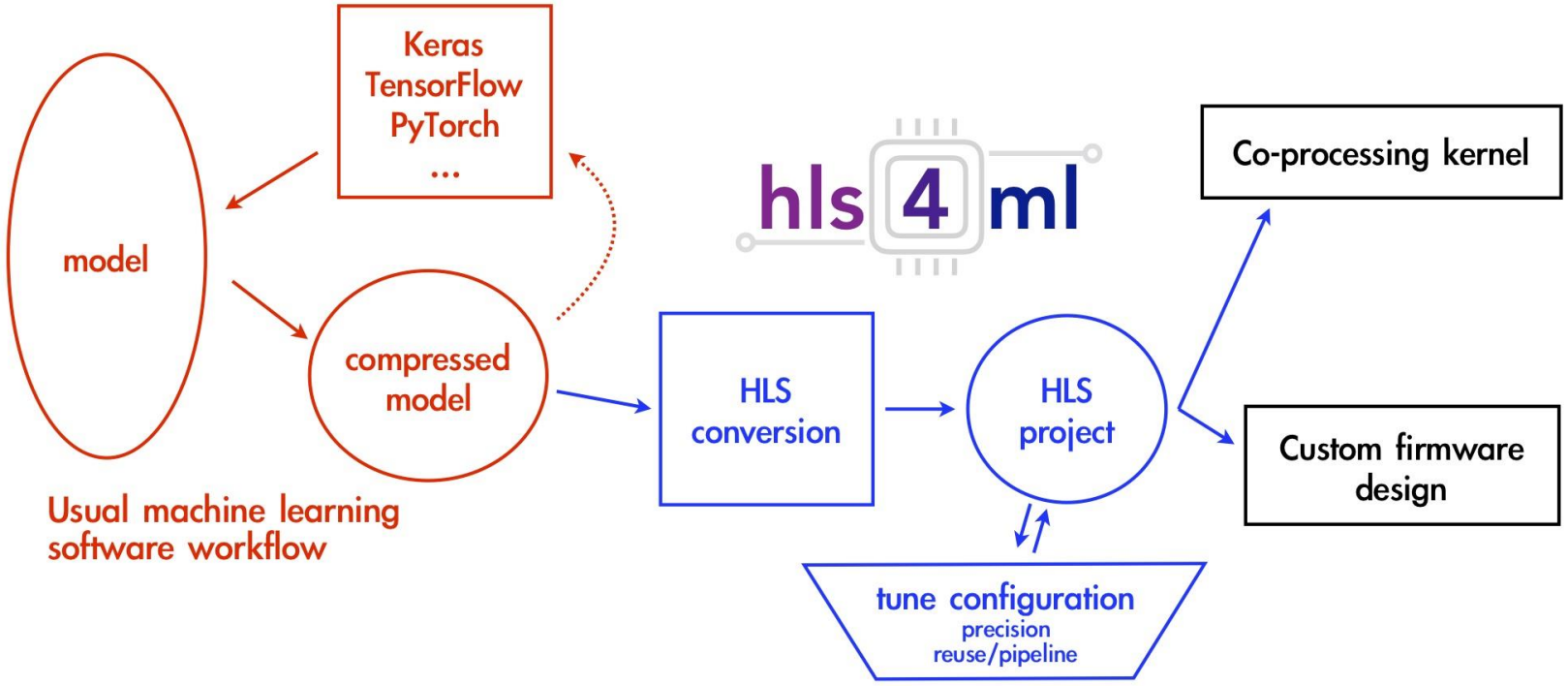
Approche concepteur:

- Adaptée pour une tâche complexe basée sur un réseau spécialisé
- Besoin d'une étude extensive sur l'adaptation entre espace latent et représentation interne du FPGA
- Besoin d'optimisation bas niveau

Exemple: Système de reconstruction de traces pour un trigger rapide

hls4ml

Outil pour un portage rapide des réseaux de neurones standards (topologie classique)



- Passerelle entre le réseau entrainé en software et l'adaptation sur FPGA
- Pruning, quantification (passage aux virgules fixes), gestion de parallèlisation/pipelining
- Gestion de la synthèse et des ressources du FPGA
- Offre une approche beaucoup plus "user-friendly" qu'un flow HLS standard
- Fournit une IP prête à être intégrée dans un projet

R&T Think IN2P3

Think : Benchmark de différentes approches d'IA embarqués (sur des cas tests)

- FPGA, puces neuro-morphiques, GPU
- Différents algorithmes et topologies de NN
- Différents flows de développement

Think phase 2: Application sur des cas physiques concrets dans différents projets

- Signal denoising, deconvolution, data clustering, …
- Mise en commun de certains développement (librairies, outils,…)
- Retours d'expérience sur l'implémentation dans différentes collaborations (Virgo, Atlas, CMS,…)

Responsable: F. Druillole (LP2I Bordeaux)

https://think.in2p3.fr/

Appel à discussions

Limites pour l'utilisation en recherche physique ?

- Très grande dépendance à la qualité de l'entrainement (méthodes et jeux de données)
- L'aspect boite noire rend difficile une approche analytique (gestion difficile des biais systématiques)

L'intégration des NN sur FPGAs est en pleine explosion, tout change très vite

⇒ Veille technologique

⇒ Réactivité

Il est impossible de concurrencer les géants de l'industrie sur des approches «mainstream»

⇒ Pour apporter quelque chose, il faut se rabattre sur les sujets de «niche»

Opportunités de développement :

- Réseaux et neurones adaptés nativement à l'architecture interne d'un FPGA (spike NN ?)
- Apprentissage embarqué sur cible => la plasticité d'un FPGA peut-être un grand avantage