Neural Processing Unit

Rémi Varenne

UNIVERSITÉ Grenoble Alpes

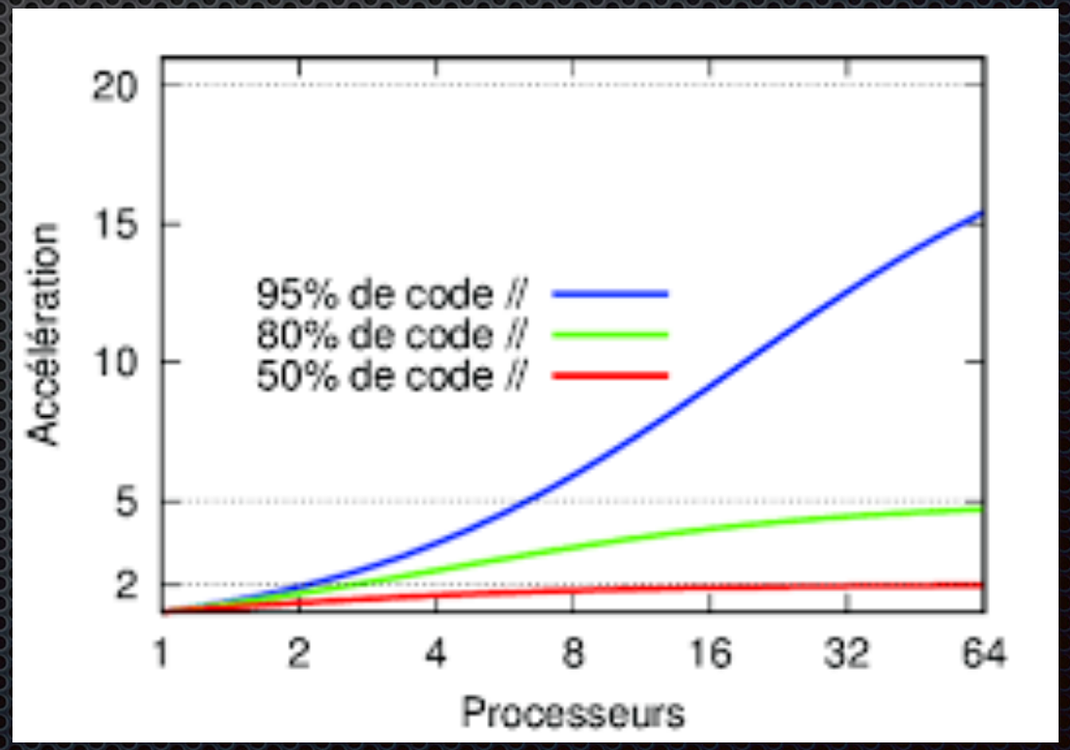
POLYTECH GRENOBLE

Qu’est ce que c’est ?

- Puce d’accélération de réseaux neuronaux
- Microprocesseur
- Son but ?

Limites de l'utilisation de l'IA

- Utilisation de CPU insuffisante
- Vitesse de calcul
- Loi d'Amdahl

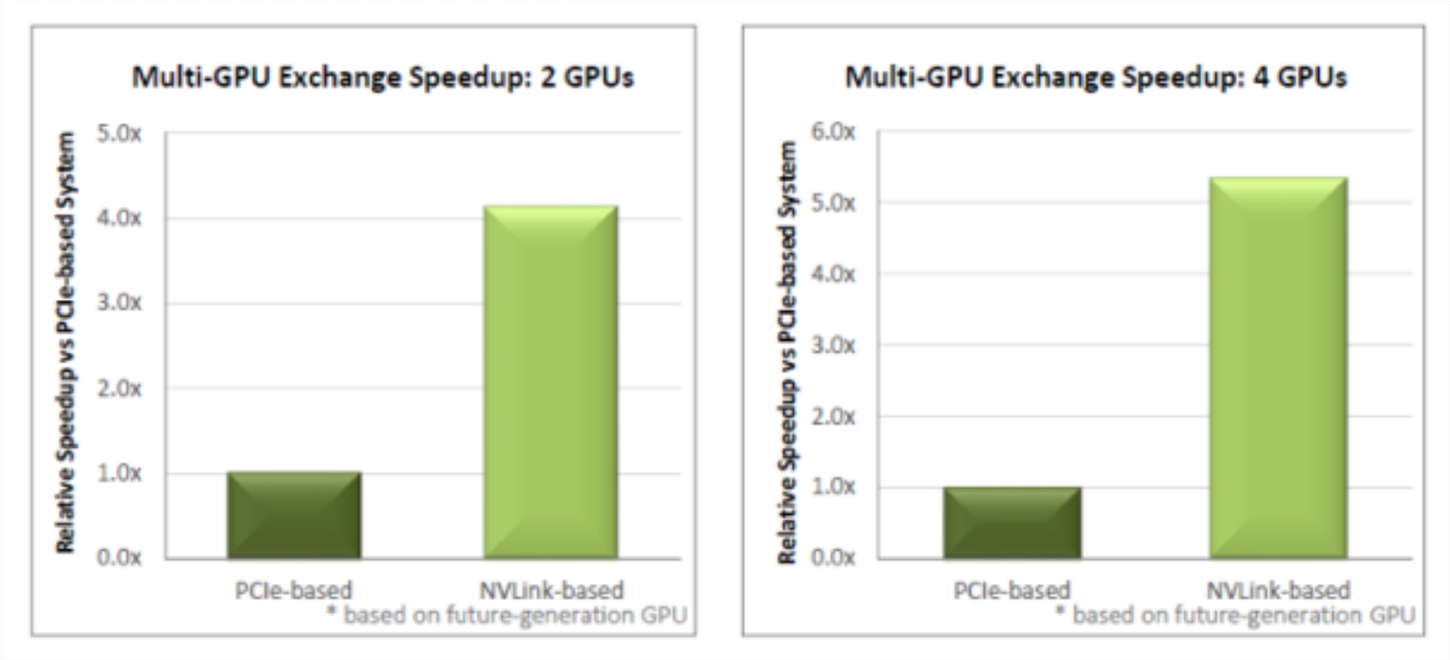


Solutions répondant à ces problématique

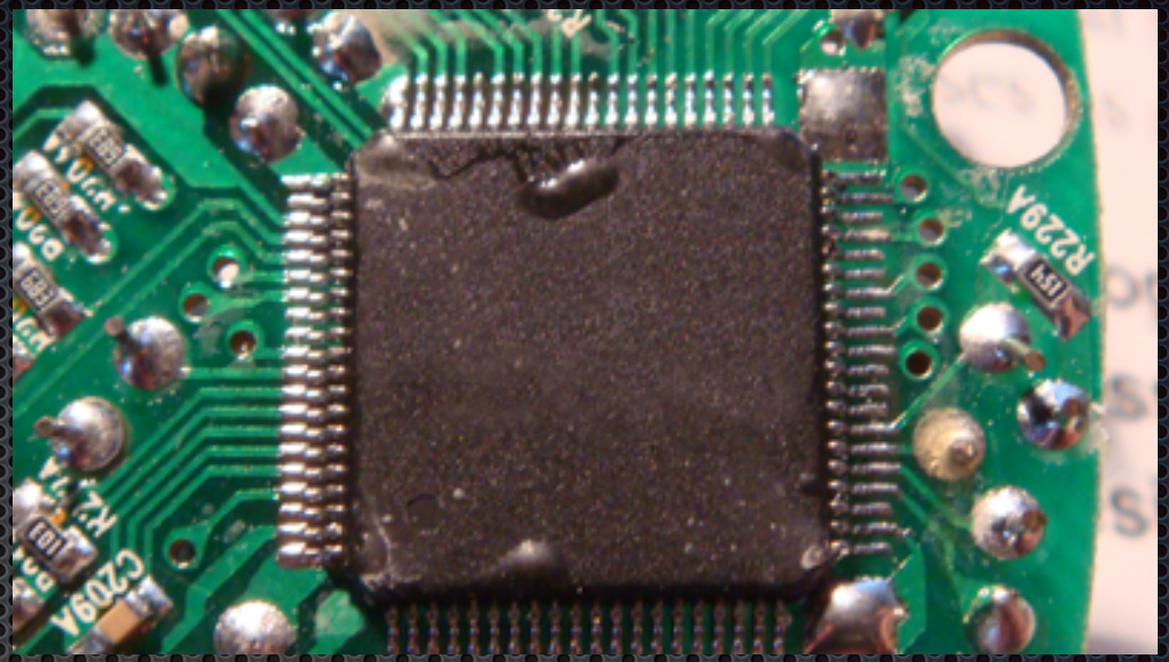
- Arithmétique basse précision
- Parallelisation
- Ajout de systèmes supplémentaires au CPU
  - GPU
  - ASIC
  - FGPA

Utilisation de GPU

- Manipulation d'images
- Obtention d'une mémoire plus importante
- Bande passante plus large
- Parallelisme interne accru
- Flexible



ASIC



- Application-specific integrated circuit
- Fait sur mesure
- Adapté pour la création d’un nombre important de produits

FPGA

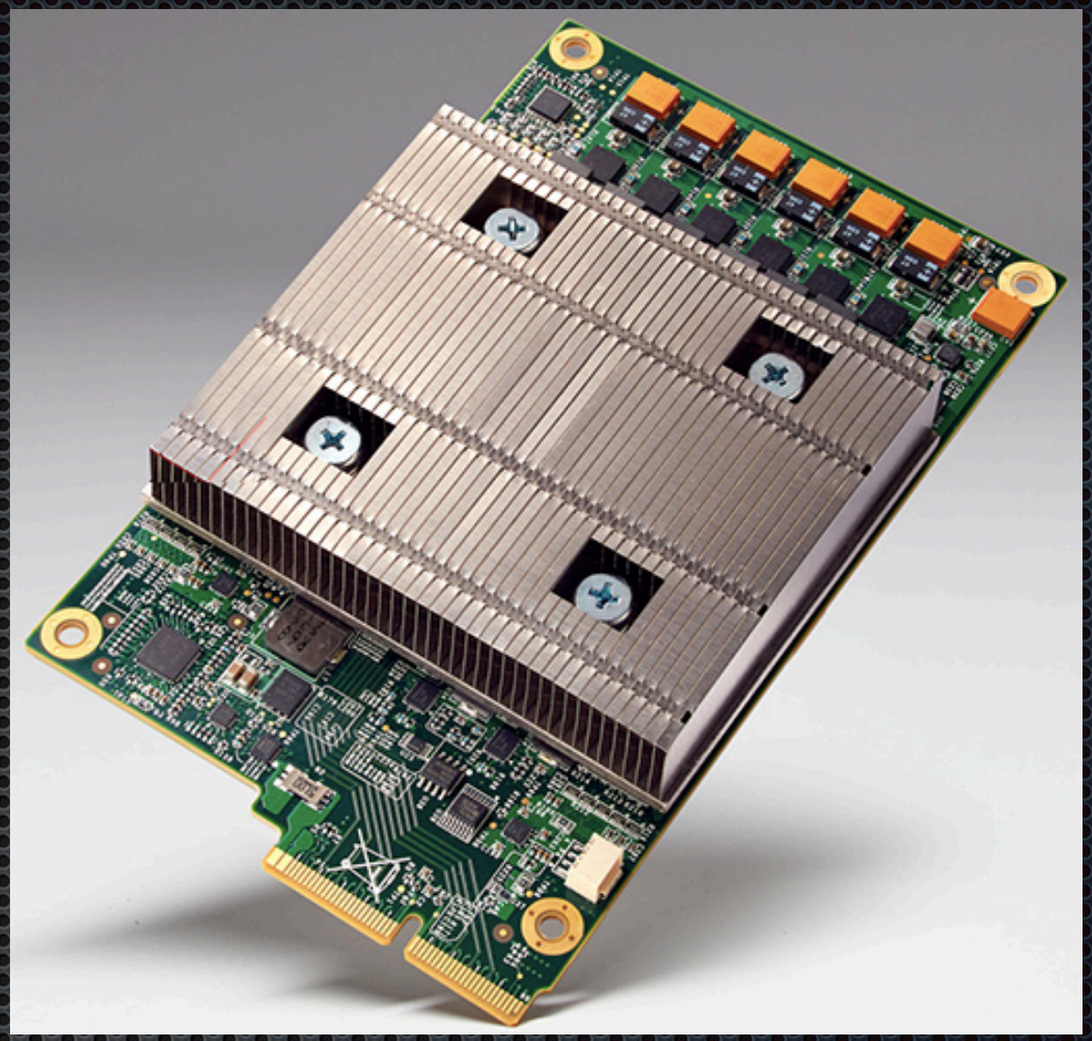
- Frameworks en constante évolution
- Field-programmable gate array
- Dispositif reconfigurable

FPGA	ASIC
Plus lents Plus chers à la pièce Plus coutant en terme d'énergie	Production moins rapide Coût plus important

Proposition de systèmes existants

- Google TPU
- Intel Nervana NNP
- Neural Engine / HiSilicon Kirin 970
- Intel Neural Compute Stick 2

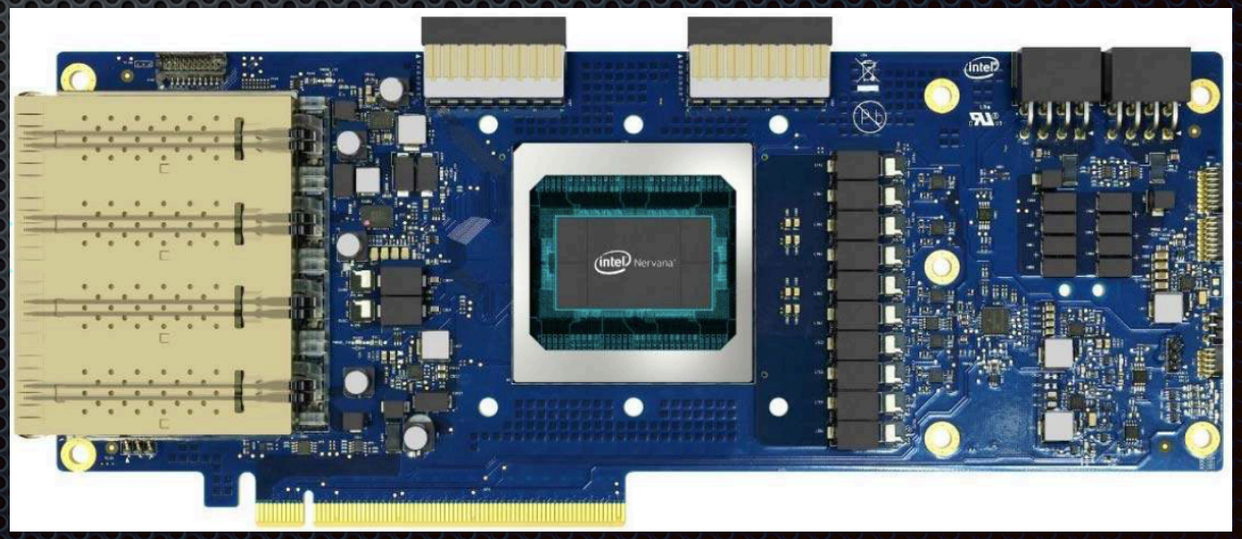
Google TPU



- ASIC
- 10 fois plus performant
- Adapté à Tensorflow
- Fonctionne avec l’arithmétique basse précision

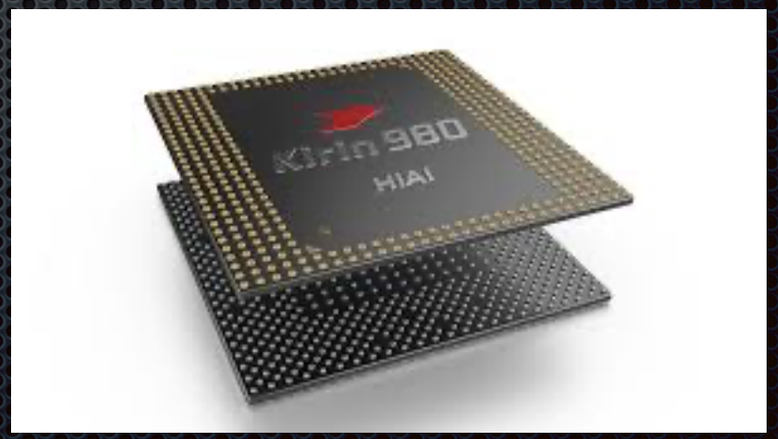
Intel Nervana NNP

- Intel Nervana Neural Network Processor
- Flexibilité
- Pas de hierarchie de cache standard
- Mémoire sur puce gérée directement sur le logiciel
- Interconnexion à haut débit
- Parallelisme



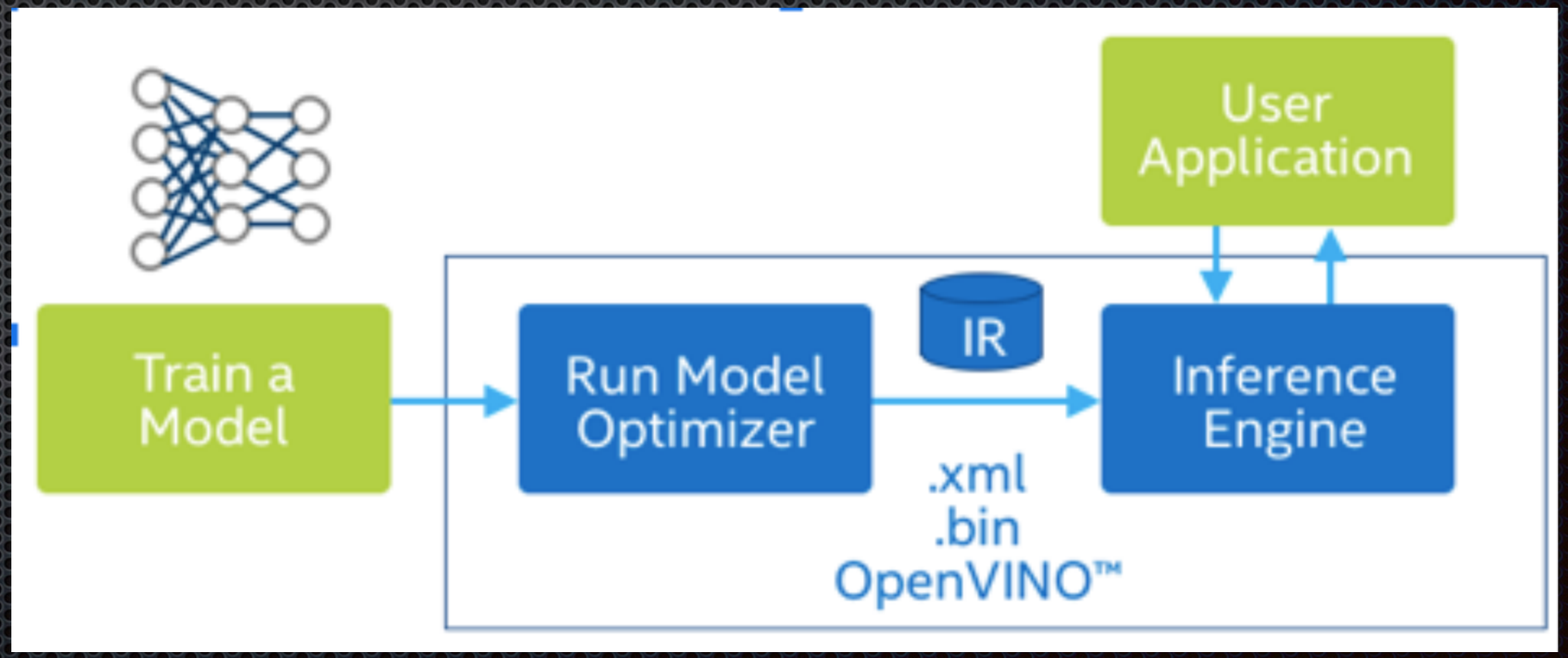
Neural Engine / HiSilicon Kirin 970

- Noyau d’accélérateur d’IA
- Dédié aux taches d’IA
- CPU de 8 coeurs et GPU de 12 cœurs
- 5,5 milliards de transistors
- 2000 images/sec



Intel NCS2

- Intel Neural Compute Stick 2
- Kit de développement plug and play pour l'IA
- Très flexible
- API commun
- C++ / Python



Passons à la démonstration

Je vous remercie de votre écoute

N’hésitez pas à poser des questions