CNRS - INP - UT3 - UT1 - UT2J

Institut de Recherche en Informatique de Toulouse

IA :
Qu'y a-t-il derrière le prompt ?

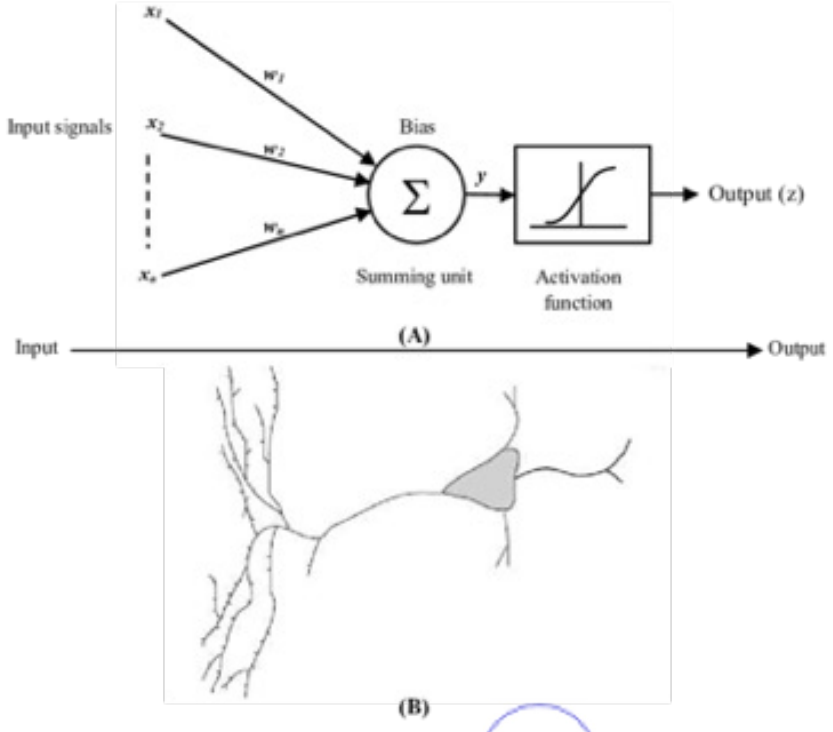
H. Luga IRIT, INSPE Toulouse, Université Toulouse 2.

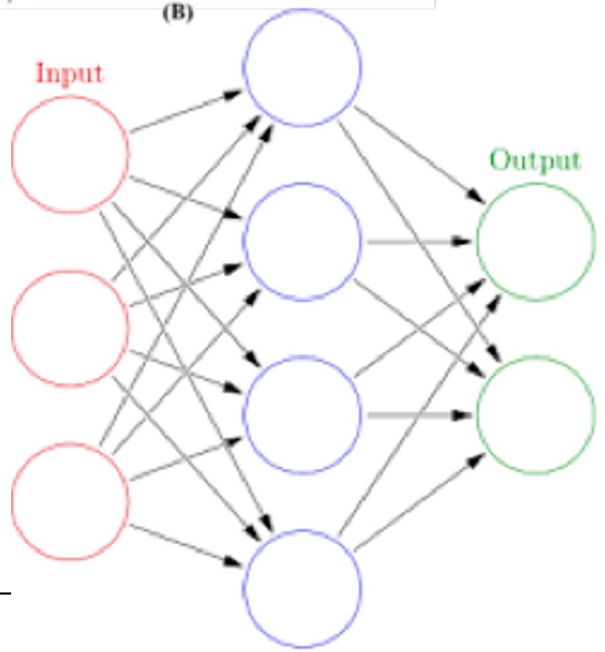
CNRS - INP - UT3 - UT1 - UT2J

Institut de Recherche en Informatique de Toulouse

Réseaux de neurones artificiels et DNN, les modèles de langage

Technique ancienne (années 1960)





- Bio inspiration pour garder les propriétés du vivant :
  - Apprentissage,
  - Généralisation
  - Résilience
- Réseaux de neurones artificiels
  - Inspirés du fonctionnement du cerveau
  - Modèles évoluant depuis les années 1950
- A la base: un neurone artificiel
  - Somme les entrées
  - Fonction de sortie (filtre, sigmoïde, …)

  ➔Fonction d'intégration
- Modélisation d'un réseau interconnectant les neurones
  - Apprentissage par modification des « poids » des liens ⇔ relations, dépendance entre neurones
  - Entrées et sorties connectées aux problème à résoudre

➔ La complexité vient de la topologie du réseau et pas d'un seul neurone

Modèle d'origine multi-couches complètement connectés: perceptron (Rosenblatt 1958)

L’entrainement

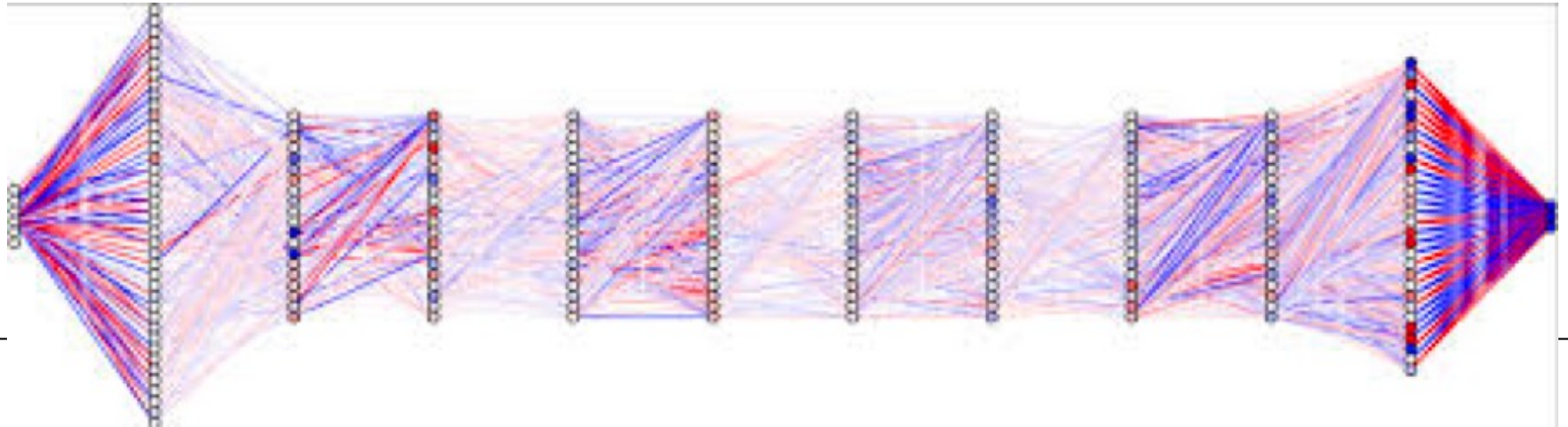
- L’entrainement d’un modèle consiste à lui fournir des milliers/millions valeurs d’entrée/sortie et de l’adapter afin qu’il réponde correctement
  - Le système ne va générer des informations que prises dans un espace correspondant aux entrées apprises
  - Les sorties du système reproduiront ce que l’on trouve dans ses données d’entrée ce qui est générateur de biais
  - Les modèles de langage ont appris sur le web et ont pour la plupart été supervisés par des humains pour éviter des réponses critiques

- **L’entrainement initial d’un (gros) modèle est coûteux et prend des jours sur un supercalculateur**

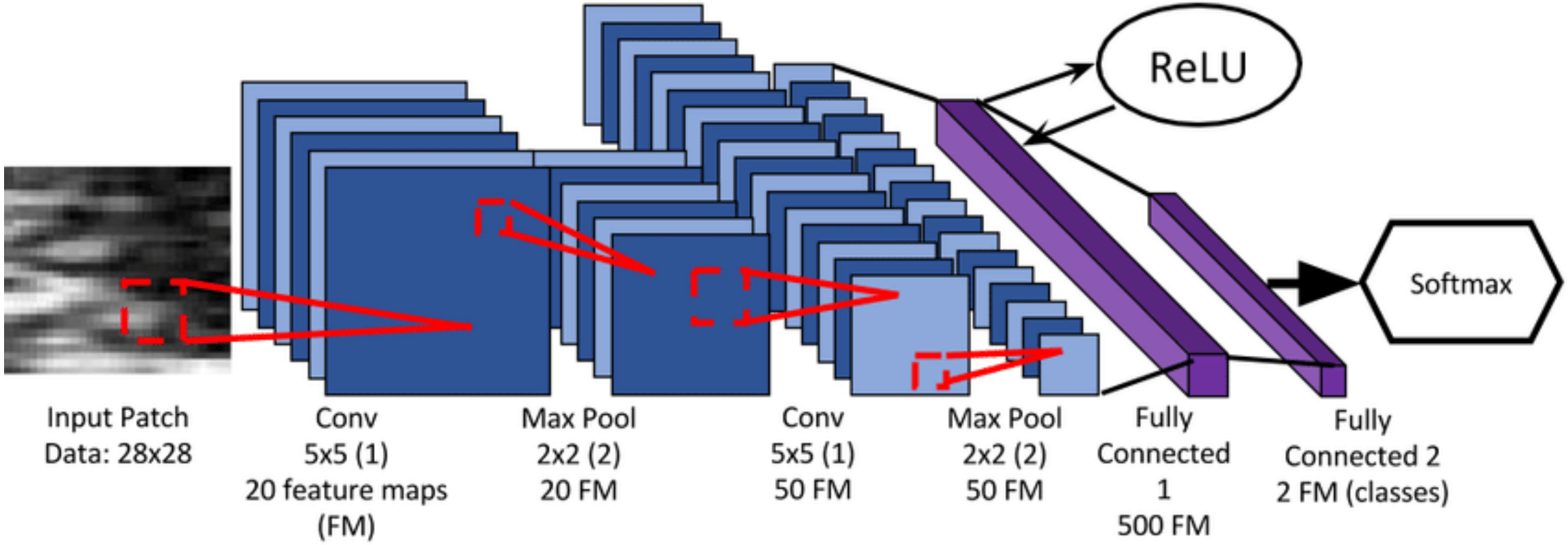
Les réseaux de neurones profonds

- S’appuient sur
  - Nouvelles techniques pour la simulation des neurones
  - Assemblage de nombreuses « couches » avec des propriétés différentes (maximisation, auto-encodeurs, complètement connectées)
  - Topologies « faites main »

- L’apprentissage
  - Couteux en calcul → besoin de puissance et donc d’électricité
  - Peut comporter des millions de paramètres → Difficulté d’interprétation du résultat



Exemples de structure pour la classification: Google LENet

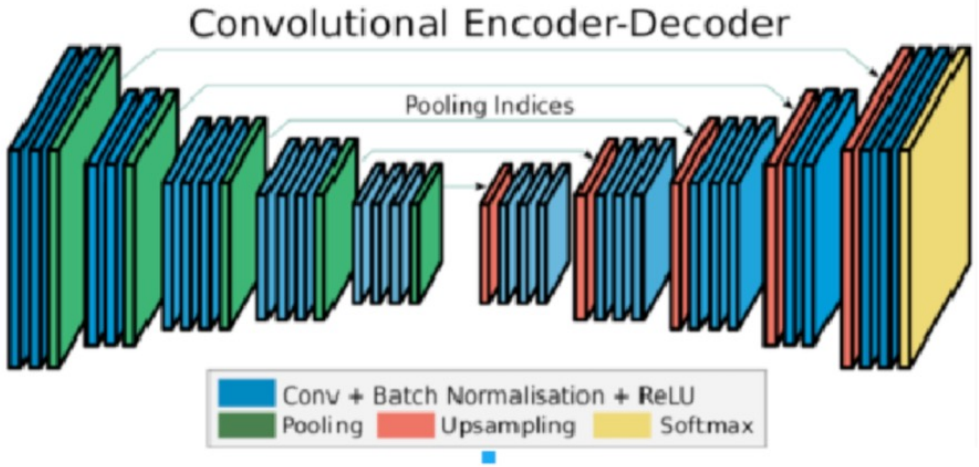


Deep learning pour la génération
Systèmes générateurs

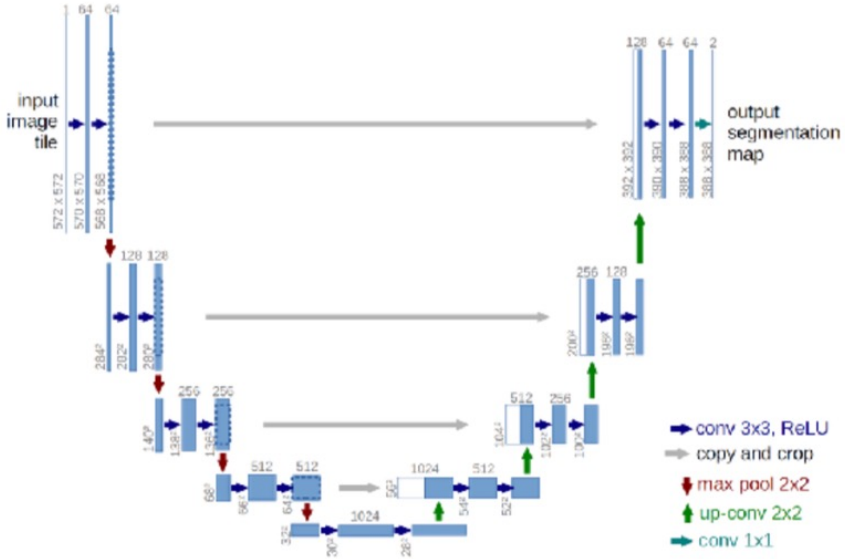
- Données d'entrée vers une représentation interne (classification)

➔ vers la sortie par recomposition

- **Segnet** A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [Badrinarayanan, Kendall, Cipolla, IEEE 2016]



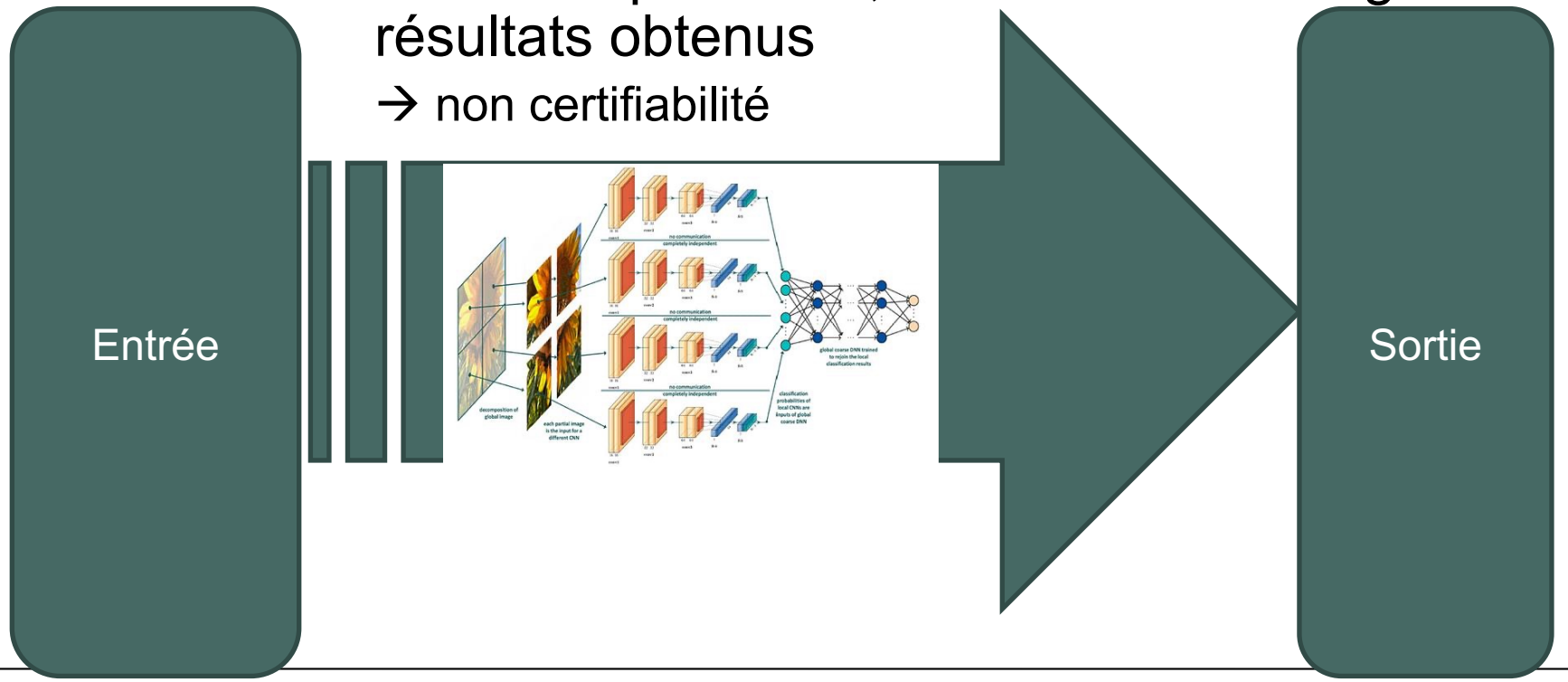
- **UNET** Convolutional Networks for Biomedical Image Segmentation [Olaf Ronneberger, Philipp Fischer, 2015]



L’aspect « boite noire »

- L’apprentissage profond soulève deux questions:
  - La partialité des résultats obtenu, selon la neutralité des jeux de données d’entraînement
  - La non explicabilité, voire le non bornage des résultats obtenus

    → non certifiabilité



CNRS - INP - UT3 - UT1 - UT2J

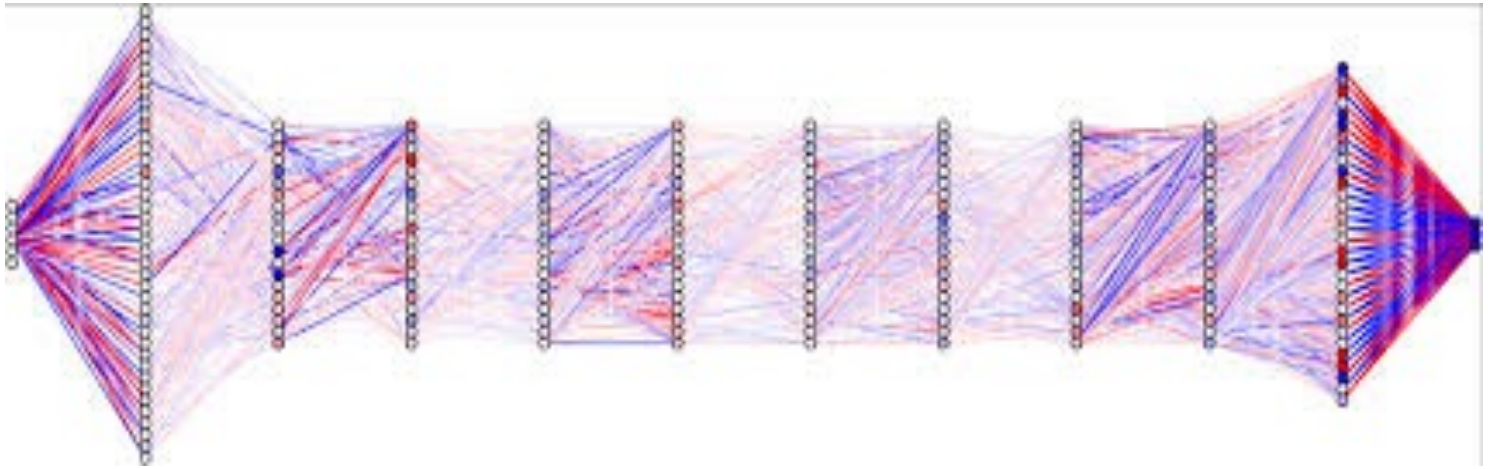
Institut de Recherche en Informatique de Toulouse

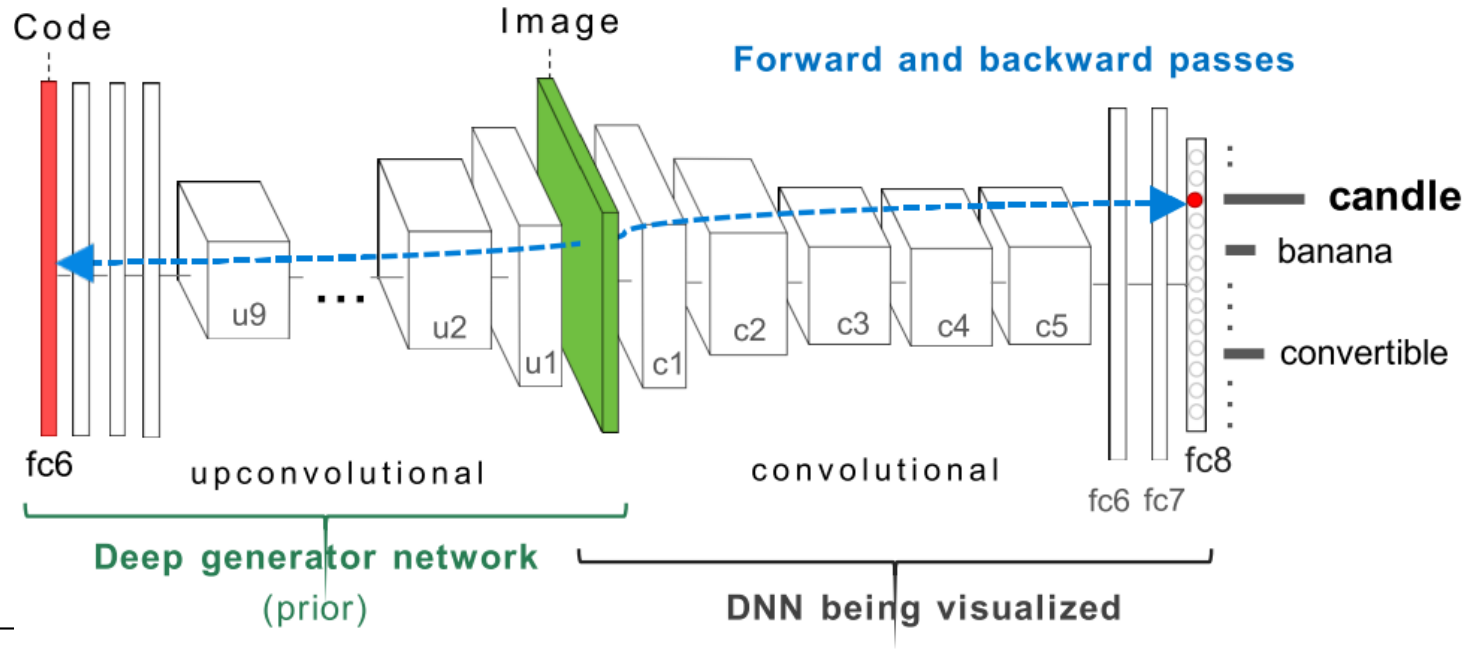
Le calcul base des DNN

Les réseaux de neurones profonds

- Structuration en couches fonctionnelles
  - Couches de convolution
  - SoftMax
  - Fully connected

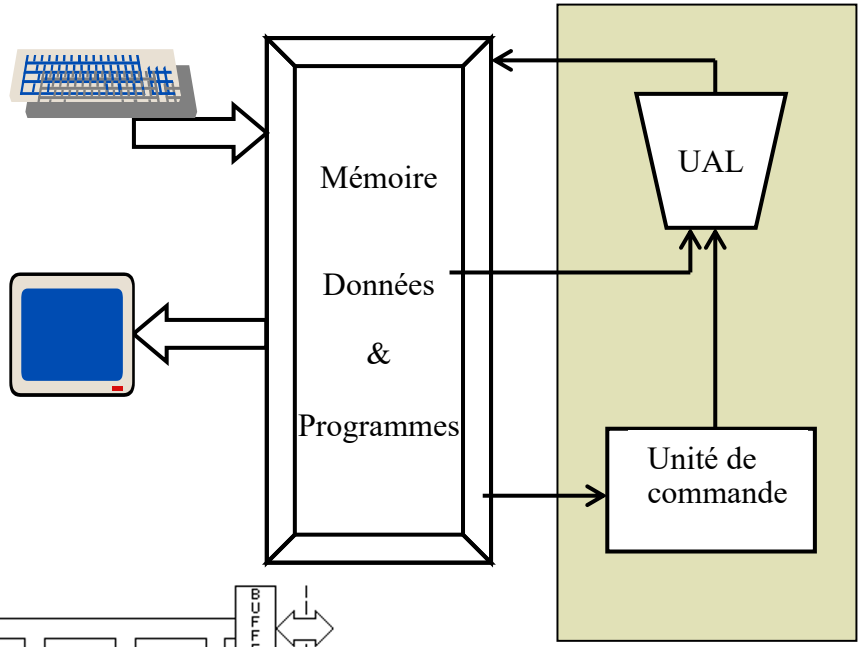
  …
- Parfois des millions (voire des milliards avec le LLM) de poids
  - ➔Opérations de calcul matriciel
  - ➔La taille du réseau est fixée
- Nécessite des compétences en design à part dans le cas des NAS



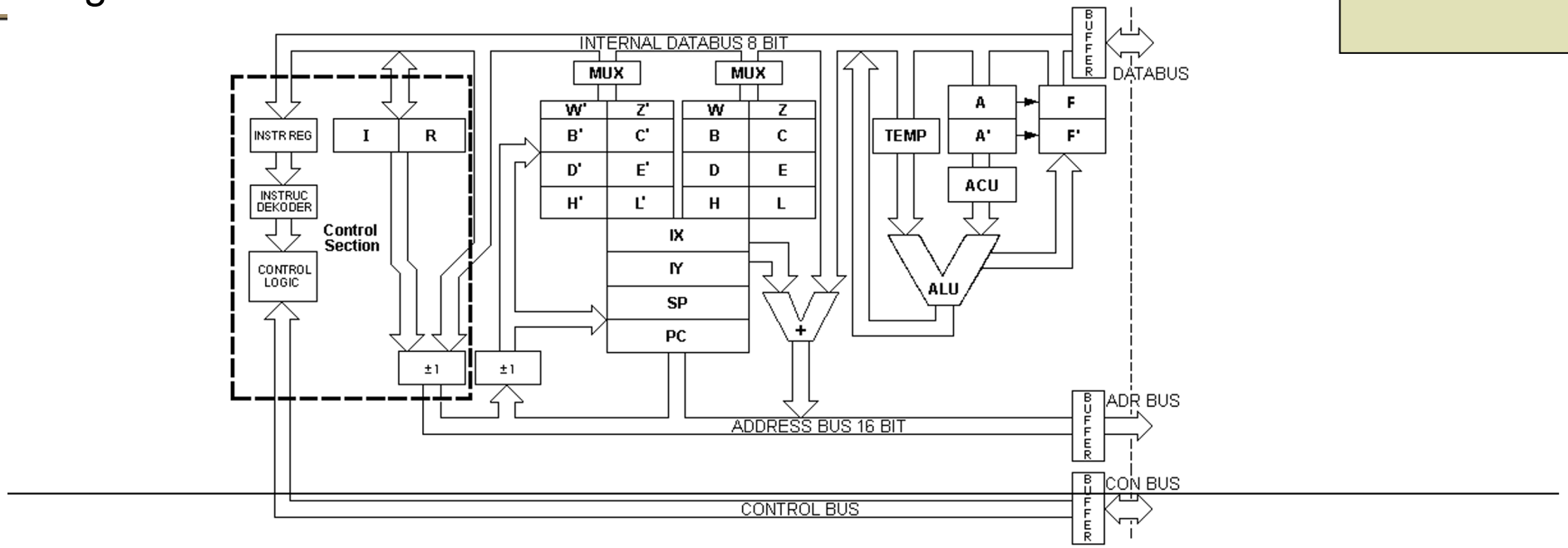


Fonctionnement d'un microprocesseur CPU

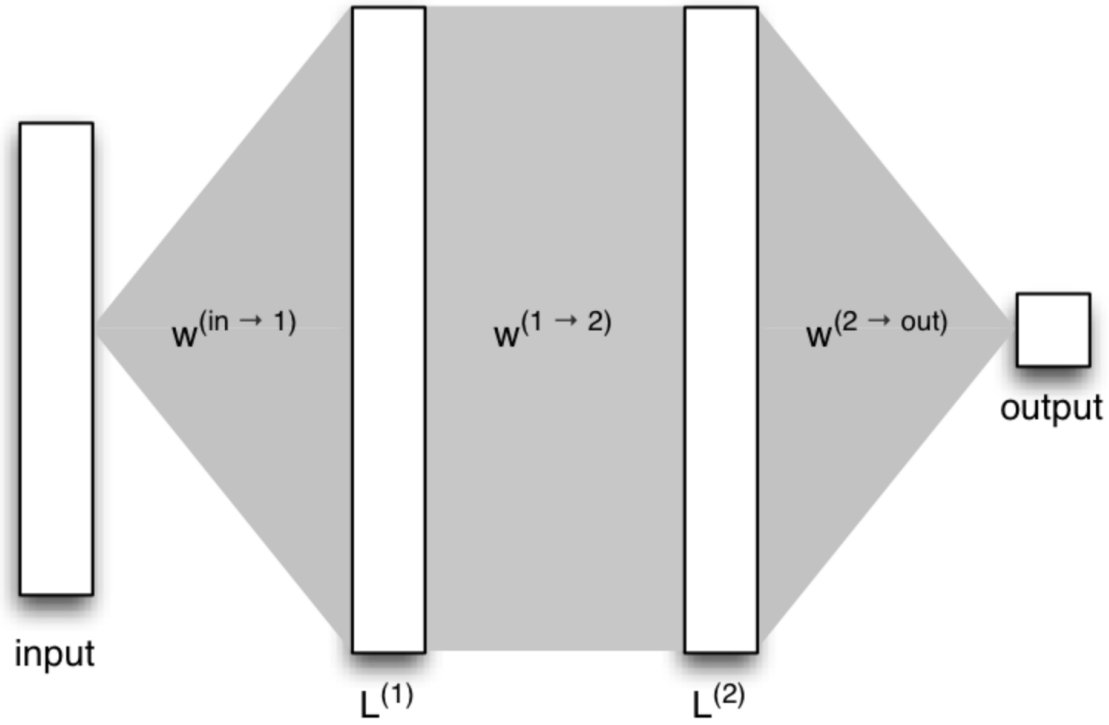
- Exécution des instructions en séquence
- Instructions de faible capacité
  - Addition, soustraction, comparaisons, branchements
- Peu de types de données
  - Booleens
  - Entiers
  - Nombres à virgule

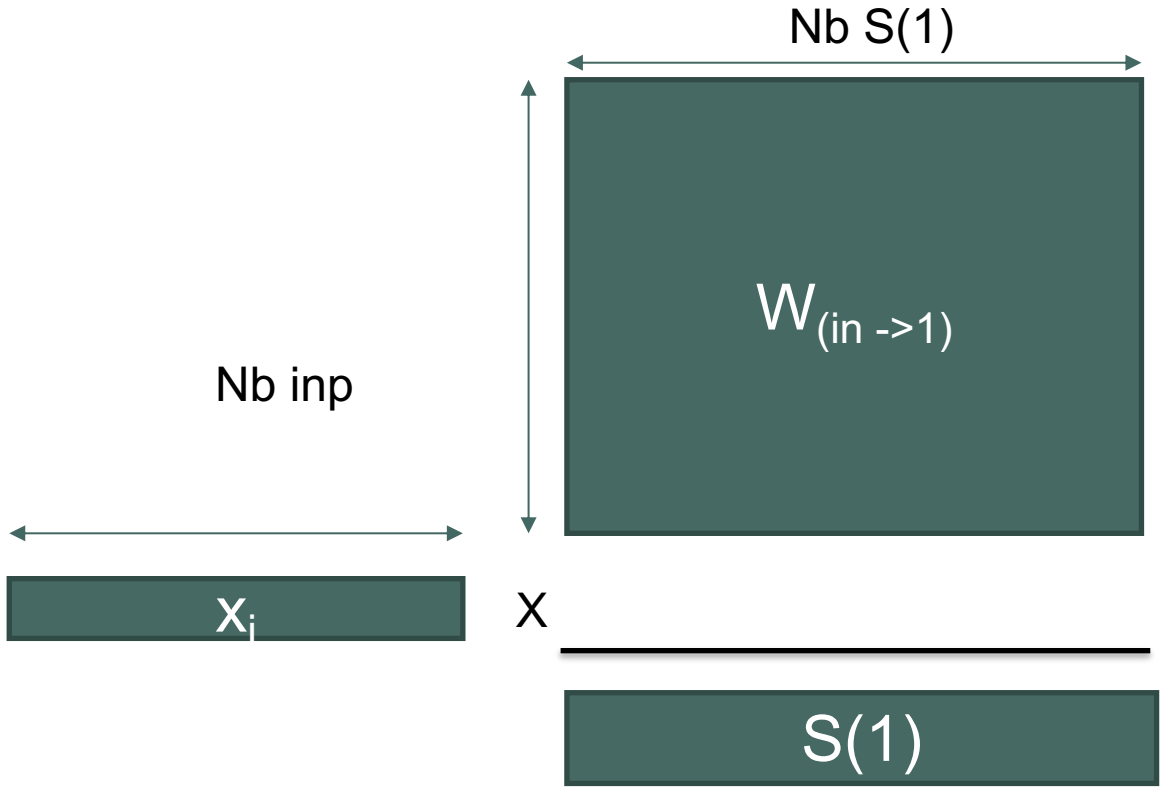



```
>0000H 3E40       LD A,40H
 0002H 3C         INC A
 0003H D380       OUT (80H),A
 0005H FE5A       CP 5AH
 0007H C20200     JP NZ,0002H
 000AH 76         HALT
 000BH FF         RST 38h
 000CH FF         RST 38h
 000DH FF         RST 38h
 000EH FF         RST 38h
 000FH FF         RST 38h
 0010H FF         RST 38h
 0011H FF         RST 38h
 0012H FF         RST 38h
 0013H FF         RST 38h
 0014H FF         RST 38h
 0015H FF         RST 38h
 0016H FF         RST 38h
 0017H FF         RST 38h
 0018H FF         RST 38h
```



Forme matricielle des ANN: exécution

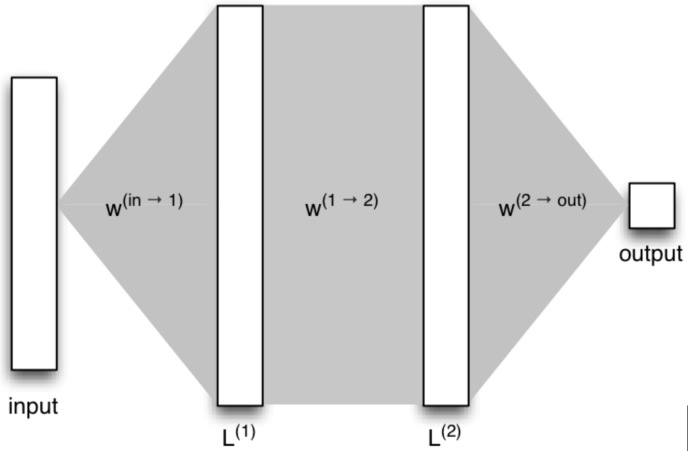


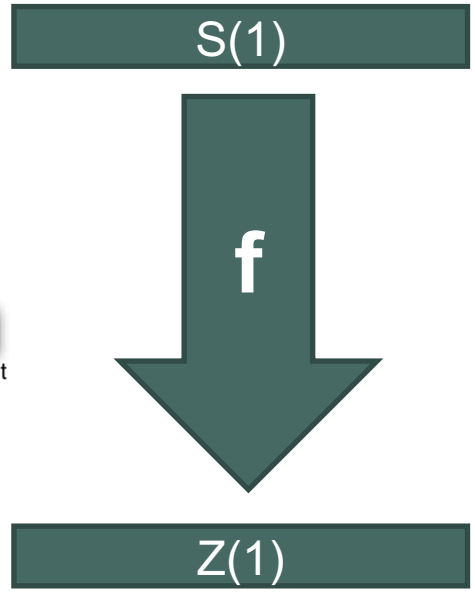


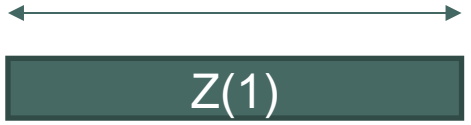
$$s_j^{(1)} = \sum_i x_i w_{i \to j}^{(in \to 1)}$$

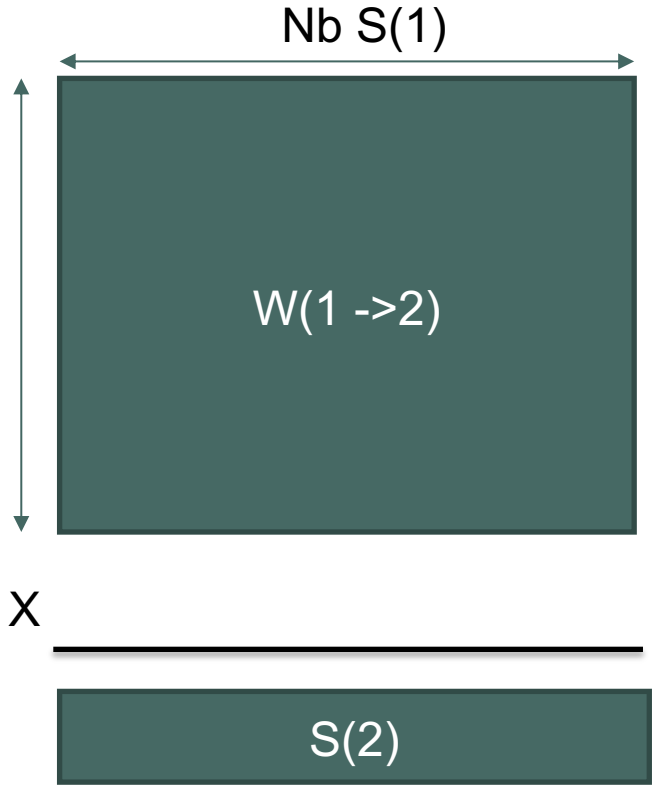
S(1)=[Input]x[W(in->1)]

Forme matricielle des ANN: exécution









$$s_j^{(2)} = \sum_i f^{(1)}(s_i^{(1)}) w_{i \to j}^{(1 \to 2)}$$

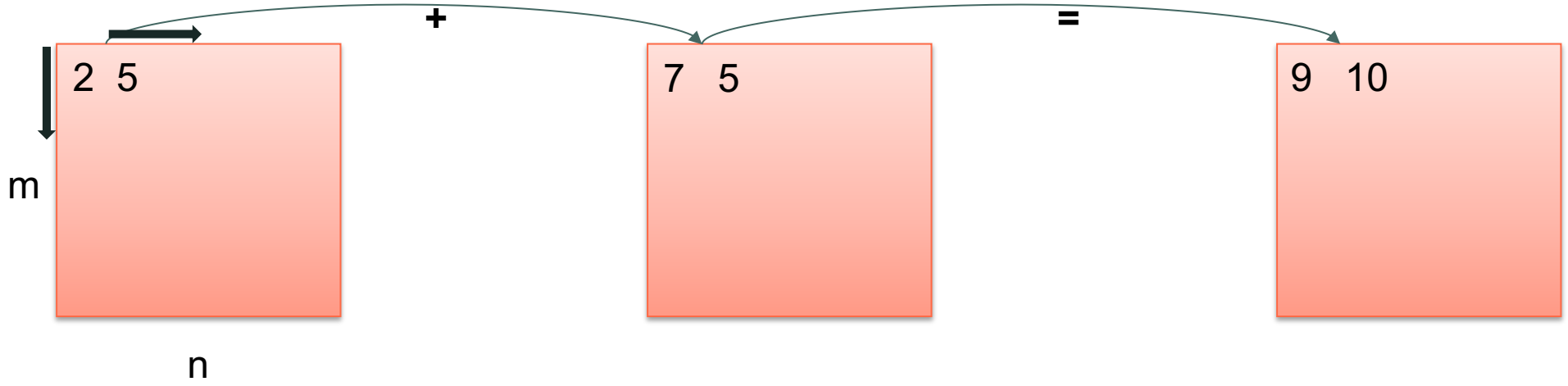
$$s_j^{(2)} = \sum_i z_i^{(1)} w_{i \to j}^{(1 \to 2)}$$

$z_i(1)=f(s_i(1))$

$S(2)=[Z(1)]x[W(1>2)]$

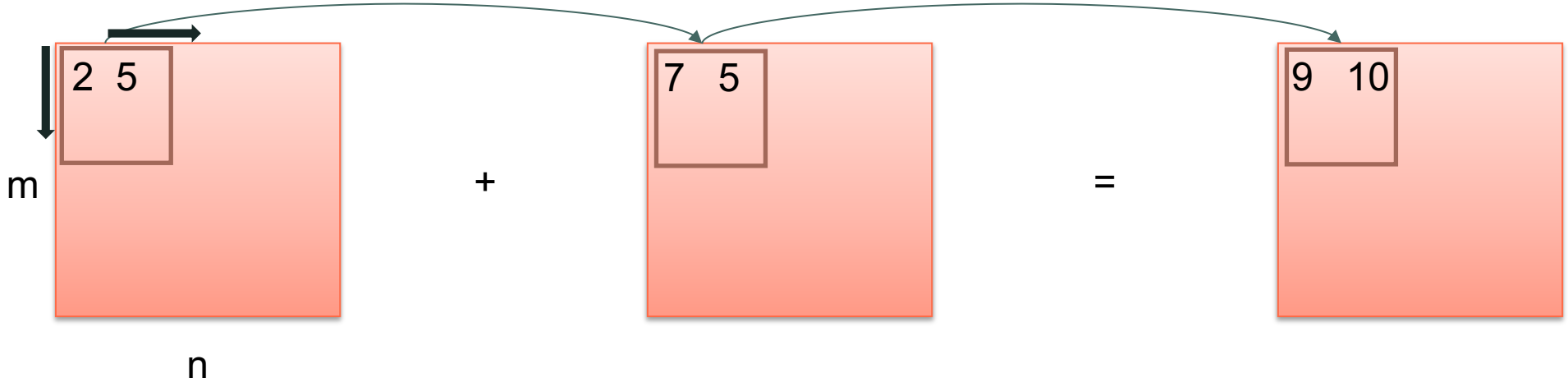
Travailler sur les matrices: Ajouter 2 matrices avec un CPU

- Double boucle => m * n opérations séquentielles



La factorisation d'opérations

- Idée: dans les calculs matriciels on a souvent la même opération appliquées à plusieurs données
  - Ajout d'instructions permettant de contrôler d'un seul coup plusieurs opérations
  - MMX, SSE pour le multimédia dans tous les processeurs
  - Avènement des GPU
- Une même opération par blocks voir par matrice entière

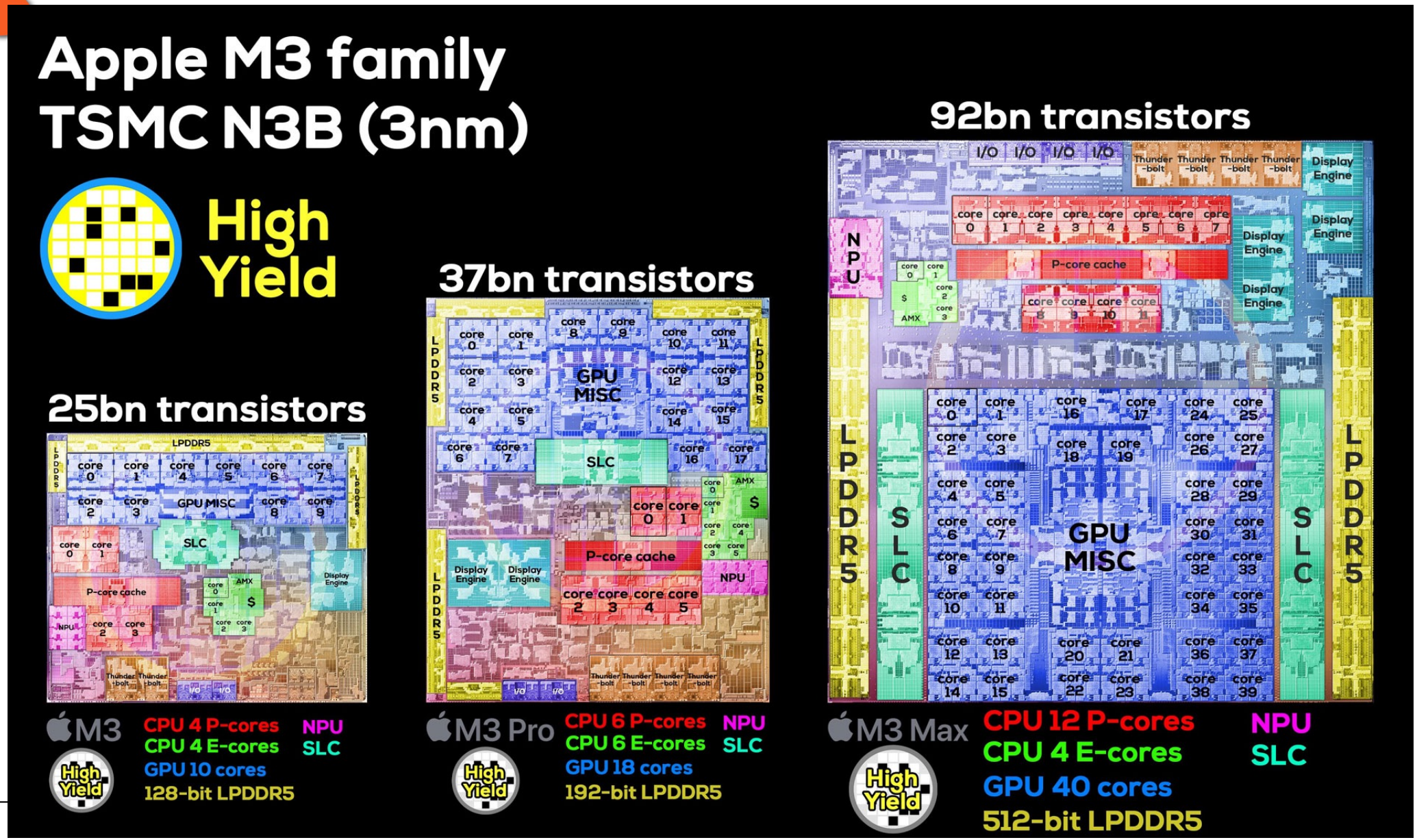


L'émergence des cartes graphiques 3D et CUDA

- Cartes 3D se sont démocratisées avec la 3dFx en 1996 pour faire des jeux vidéos
- En 1999, la société Nvidia a lancé des cartes de la série GeForce toujours pour la 3D et le multimédia
  - … Hors la 3D se base sur des opérations matricielles …
- En 2007 CUDA permet de faire du GPGPU: Utiliser les cartes graphiques pour calculer

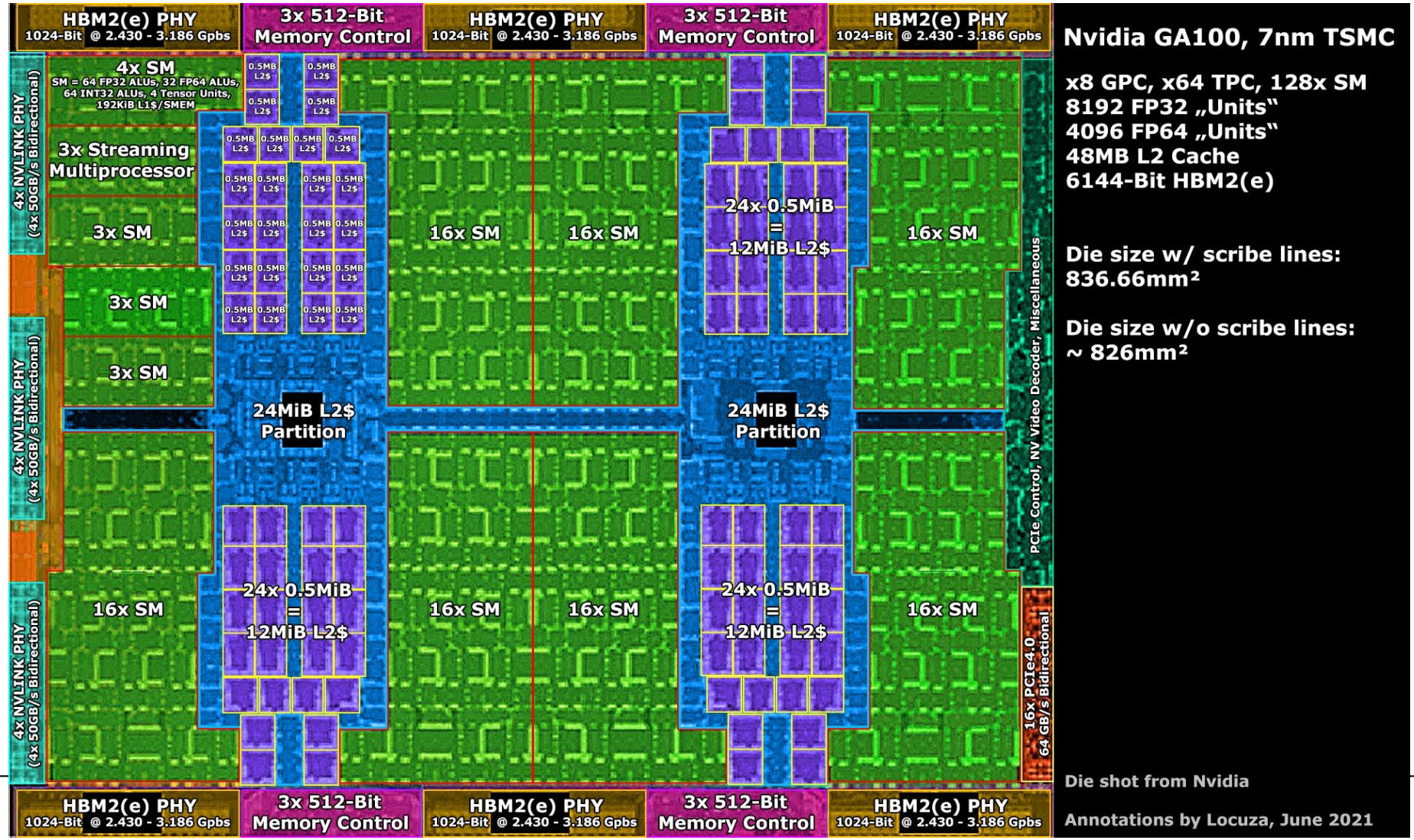
➔ Tous les processeurs se dotent de capacités GPU, sans égaler le matériel dédié

CPU, GPU, TPU, NPU : de multiples possibles



CPU, GPU, TPU, NPU : de multiples possibles

Nvidia A100, 300w par carte



Nvidia GA100, 7nm TSMC

x8 GPC, x64 TPC, 128x SM
8192 FP32 „Units"
4096 FP64 „Units"
48MB L2 Cache
6144-Bit HBM2(e)

Die size w/ scribe lines:
836.66mm²

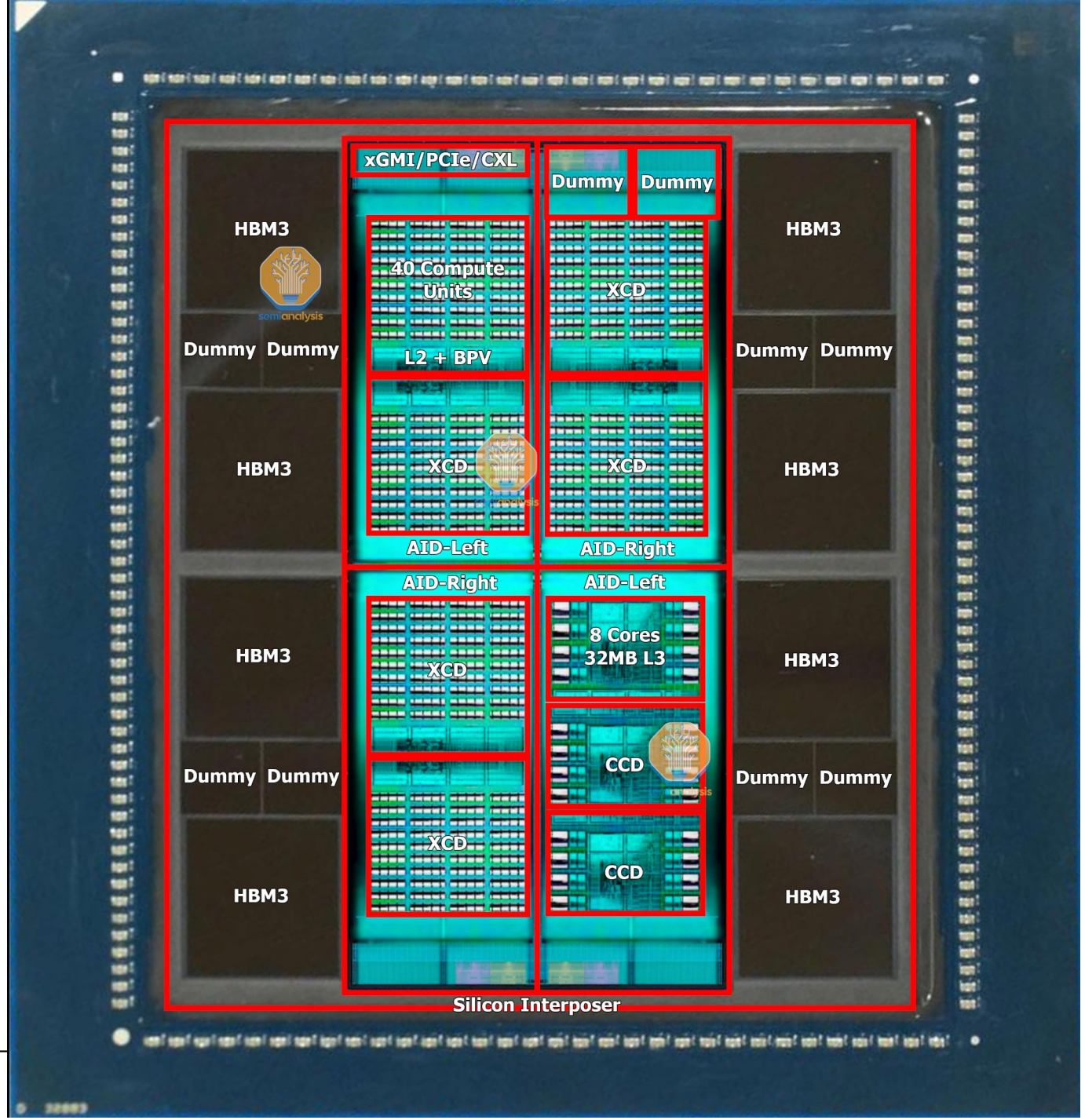
Die size w/o scribe lines:
~ 826mm²

Die shot from Nvidia

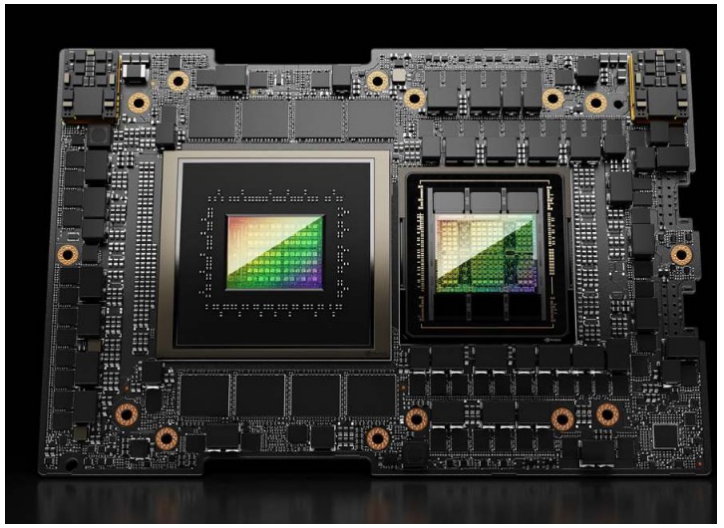
Annotations by Locuza, June 2021

CPU, GPU, TPU, NPU : de multiples possibles

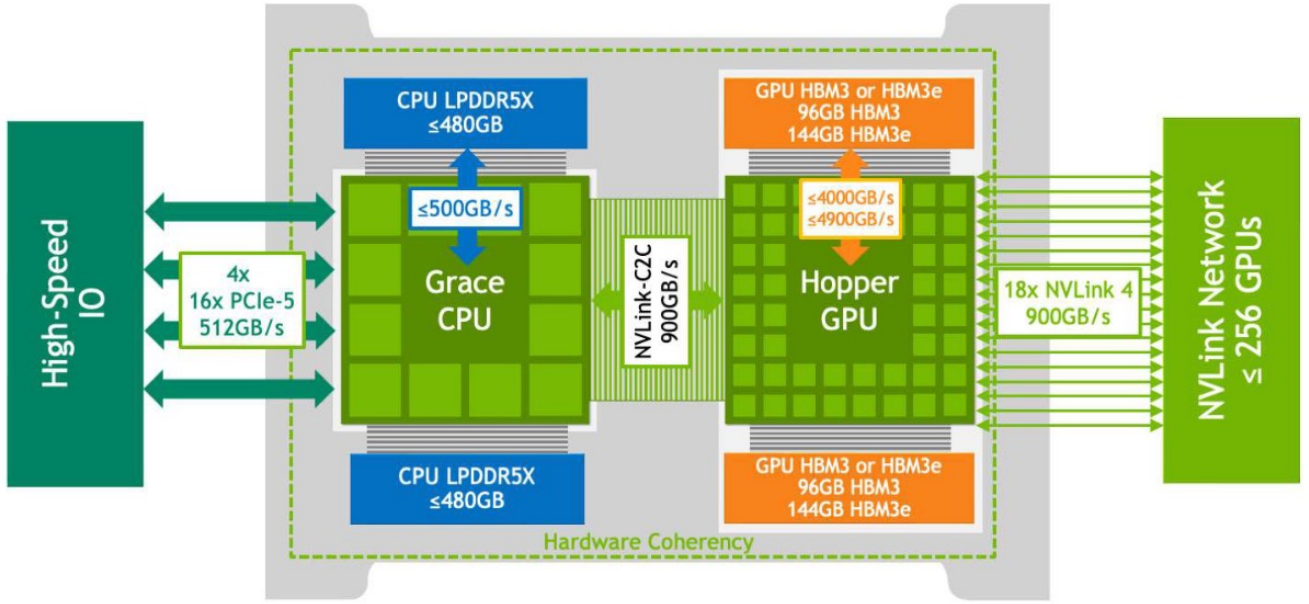
- AMD MI 300
- 6 nm
- 750 w/Carte



Nvidia GH200



- Atteint 1000w/puce
- En haute intégration on peut en placer 8 par serveur



Besoins en calcul

- Pour l'entrainement:
  - Des milliers de cartes en parallèle pour adapter les poids des neurones
  - Une infrastructure distribuée nécessitant un Datacentre (ou des) pour travailler en haute intégration sur le même modèle
  - Un refroidissement adapté, souvent liquide pour le circuit primaire

- Pour l'inférence
  - Besoin de puissance de manière ponctuelle mais de cartes possédant beaucoup de mémoire pour "faire rentrer" les paramètres des modèles
  - Décroissance de la précision des modèles qui permet de proposer des versions "dégradées" FP31->FP4

Coûts d'entrainement

- L'entrainement initial des systèmes générateurs est très consommateur
  - Training Meta's LLAMA took 3.7M (154166 ans) GPU hours on A100-80GB
- Une requête simple est moins consommatrice
  - Possible sur les puces intégrant un accélérateur (Ex: Macs Mx)
- L'adaptation à partir d'un modèle est possible sur machines individuelles

Microsoft et OpenAI vont construire un datacenter de 100 milliards $ pour le supercalculateur d'IA Stargate

2 avril 2024
Yves Grandmontagne

Supercalcul : Jean Zay se modernise pour l'IA en un temps record

Suite à la volonté française d'accélérer l'accès aux ressources pour les recherches en IA, le supercalculateur de l'IDRIS se dote de 14 nouveaux racks qu'Eviden est parvenu à fournir en à peine quelques semaines.

PC Search

PCMag editors select and review products independently. If you buy through affiliate links, we may earn commissions, which help support our testing.

Home > News > AI

Zuckerberg's Meta Is Spending Billions To Buy 350,000 Nvidia H100 GPUs

In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, the company's CEO Mark Zuckerberg says.

By Michael Kan January 18, 2024

Des annonces de plus en plus fracassantes

Le gouvernement annonce 35 sites « prêts à l'emploi » en France pour accueillir des

Les émissions de carbone de Google explosent à cause de l'IA

Etats-Unis. Trump annonce le projet Stargate, l'IA "made in USA" à 500 milliards de dollars

Intelligence artificielle : Microsoft va investir 80 milliards de dollars dans ses data centers en 2025

Mistral AI choisit Eclairion pour héberger son premier data center en France

Le champion français de l'IA générative a fait le choix d'Eclairio

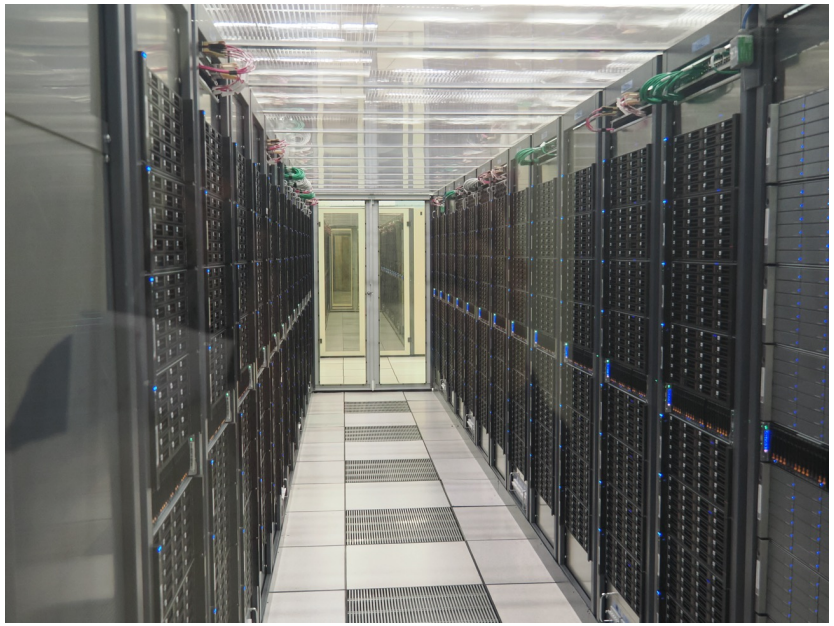
IA : les Emirats arabes unis annoncent la construction en France d'un data center géant

Ce « campus », dont la création a été entérinée jeudi à Paris en marge d'un sommet mondial sur l'IA, représente des investissements de 30 milliards à 50 milliards d'euros, selon l'Elysée.

Les Datacenters

- Hangars d'hébergement (ou containers) fournissant
  - Sécurité
  - Refroidissement
  - Electricité
  - Réseau

  Le tout redondé pour atteindre une disponibilité maximale



- Le PUE (Power Usage Efficiency) mesure l'efficacité énergétique des DC, c'est le ratio puissance consommée totale/puissance consommée par les serveurs
  - Un DC "moderne" arrive à des PUE <1.2

  ➔Mieux vaut un bon DC qu'un éparpillement des ressources

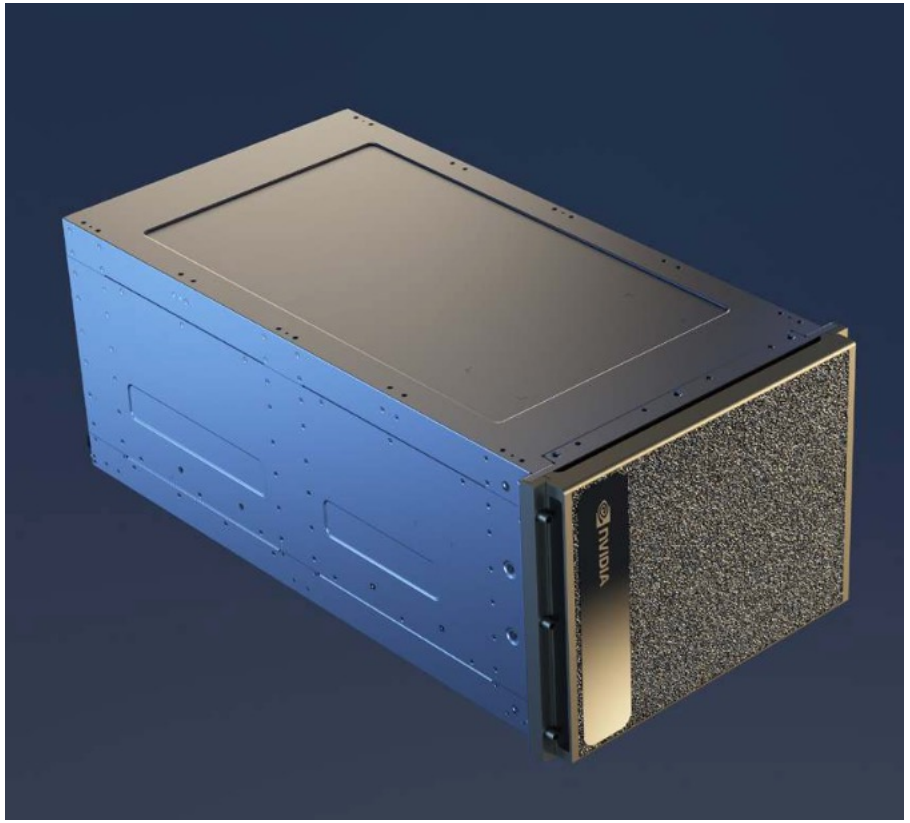
Le refroidissement et l’électricité

- Climatisation
  - Refroiddissement passif
  - Récupération de chaleur
- Electricité
  - Redondance

Urbanisation

- Des racks contiennent des machines dans un format standardisé

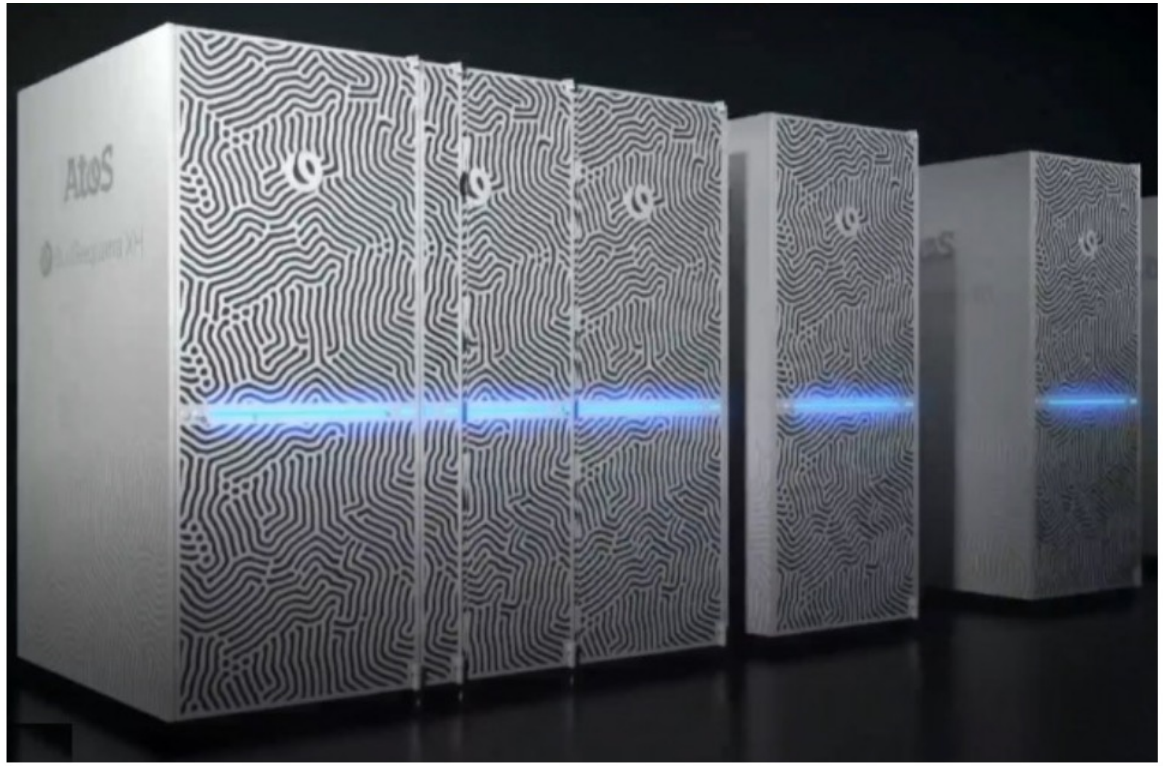
Serveur NVIDIA DGX H200



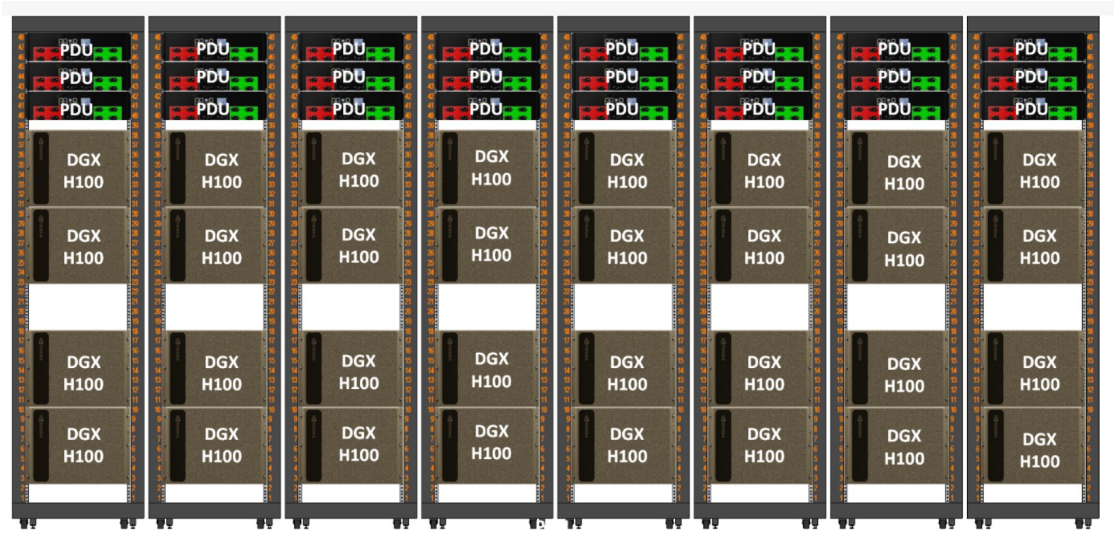
Specifications	
GPU	8x NVIDIA H200 Tensor Core GPUs, with 141GB of GPU memory each
GPU memory	1,128GB total
Performance	32 petaFLOPS FP8
NVIDIA NVSwitch™	4x
System power usage	10.2kW max*
CPU	Dual Intel® Xeon® Platinum 8480C Processors 112 Cores total, 2.00 GHz (Base), 3.80 GHz (Max Boost)
System memory	2TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet
Management network	10Gb/s onboard NIC with RJ45 100Gb/s Ethernet NIC Host baseboard management controller (BMC) with RJ45
Storage	OS: 2x 1.92TB NVMe M.2
Internal storage:	8x 3.84TB NVMe U.2
Software	NVIDIA AI Enterprise – Optimized AI software NVIDIA Base Command – Orchestration, scheduling, and cluster management DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky – Operating System
Support	Comes with 3-year business-standard hardware and software support
System weight	287.6lbs (130.45kgs)
Packaged weight	376lbs (170.45kgs)
System dimensions	Height: 14.0in (356mm) Width: 19.0in (482.2mm) Length: 35.3in (897.1mm)
Operating temperature range	5–30°C (41–86°F)

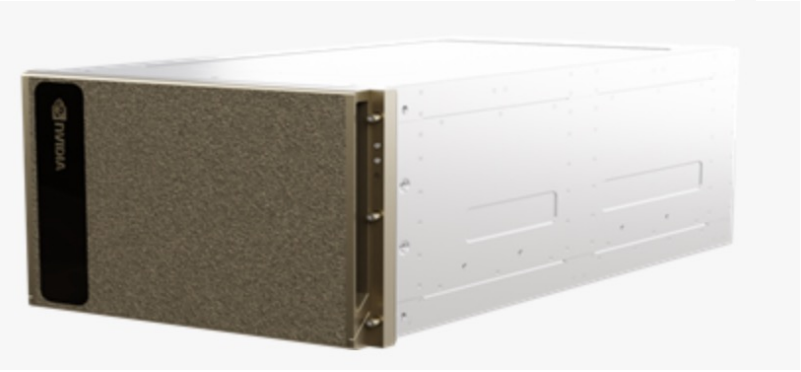
Les machines HPC à haute intégration

- Limite électrique EU : 20MW/Exaflop



Sur son système HPC XH3000, Atos pourra combiner des composants Intel, AMD, Nvidia (GPU ou Grace avec puce ARM en complément) ou SiPearl. (Crédit : S.L.)





Bilan

- L'IA la plus efficace aujourd'hui nécessite une grande quantité de données **de qualité**
- Les modèles de langages quand entrainés sur des corpus pertinents permettent de simplifier les relations entre des systèmes normalisés et les usagers
- Les coûts environnementaux de ces systèmes est important pour la partie initiale d'entrainement mais beaucoup moins à l'usage
- Beaucoup de solutions proposées sont du domaine de l'algorithmique classique et n'ont d'IA que la mode liée au mot.

Questions ?

herve.luga@univ-tlse2.fr