CATIE

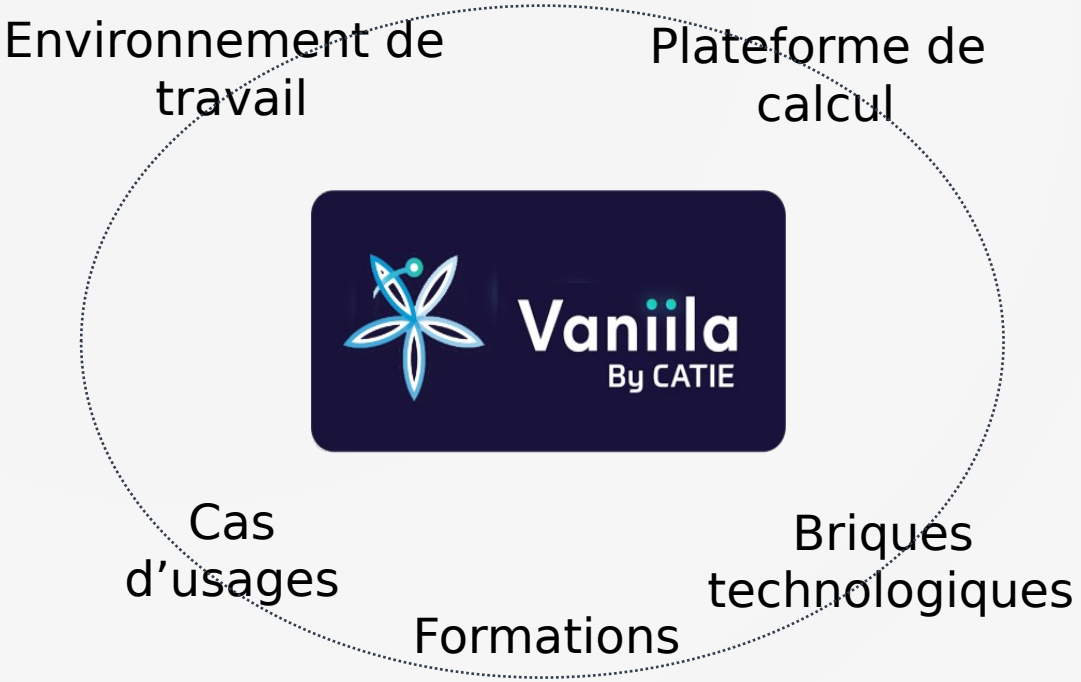
Solutions pour la société numérique

Plateforme Vaniila

Nicolas PHILIPPE

RÉGION Nouvelle-Aquitaine

Plateforme Vaniila



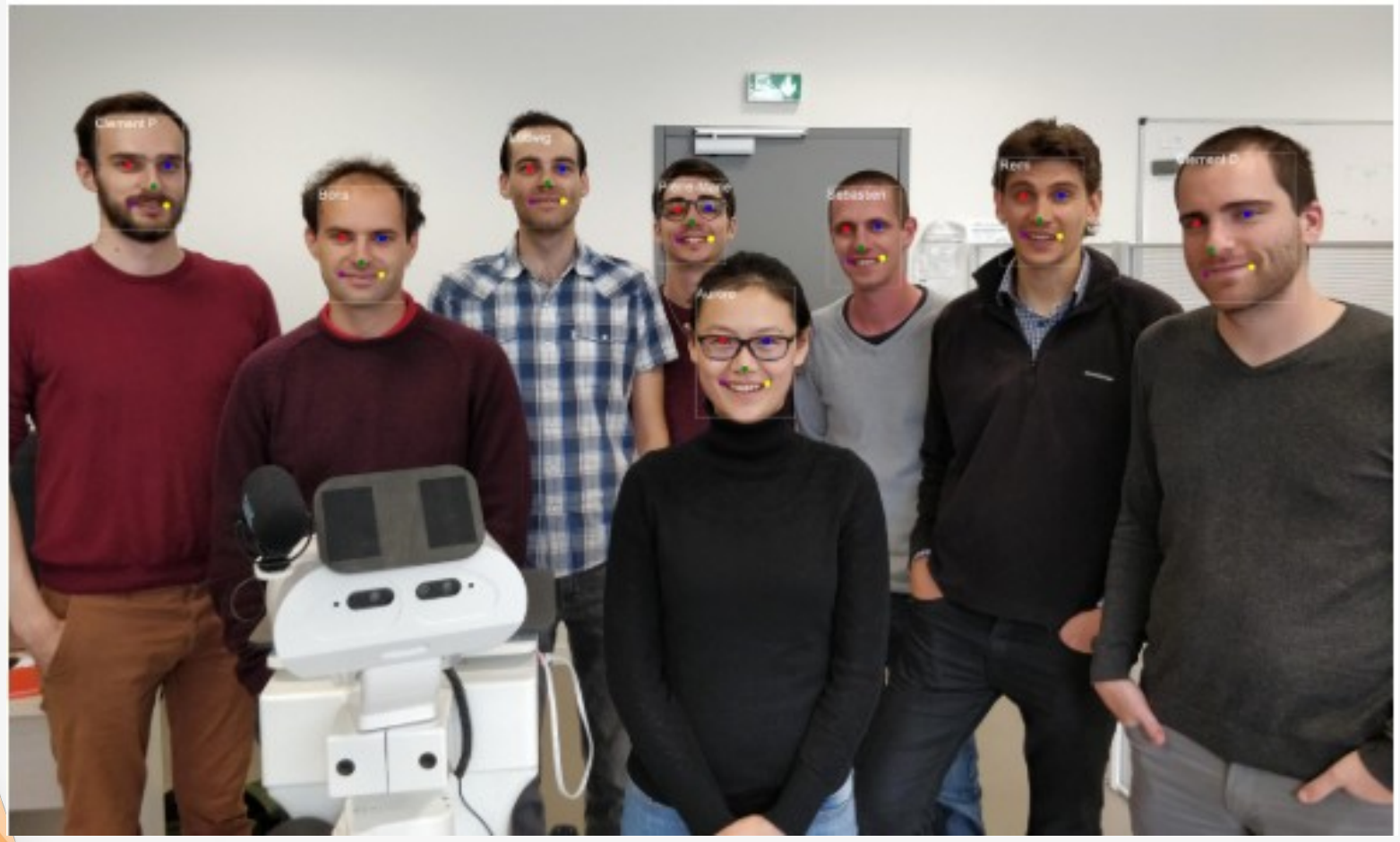
Plateforme Visant l'Accompagnement de Nouveaux Intervenants dans l'Intelligence Artificielle

- Démontrer le potentiel de l'IA
- Faciliter l'accès aux technologies associées
- Mettre à disposition des briques technologiques et « use cases » pour comprendre et s'approprier les technologies
- Faciliter le démarrage de projets IA
- Proposer l'accès à des ressources de calcul, accompagné par les experts du CATIE
- Proposer une réponse aux besoins de formation et d'accompagnement des entreprises

Le cluster de calcul

- 140 cœurs / 25 GPUs / 1,2 To de RAM /328 Go de mémoire vidéo
- Accès gratuit:
  - Via SSH
  - Via l'environnement de travail bientôt
- Hébergé dans les locaux du CATIE (ENSEIRB)

Cas d’application



CATIE

Solutions pour la société numérique

Deep learning distribué

Boris Albar

RÉGION Nouvelle-Aquitaine

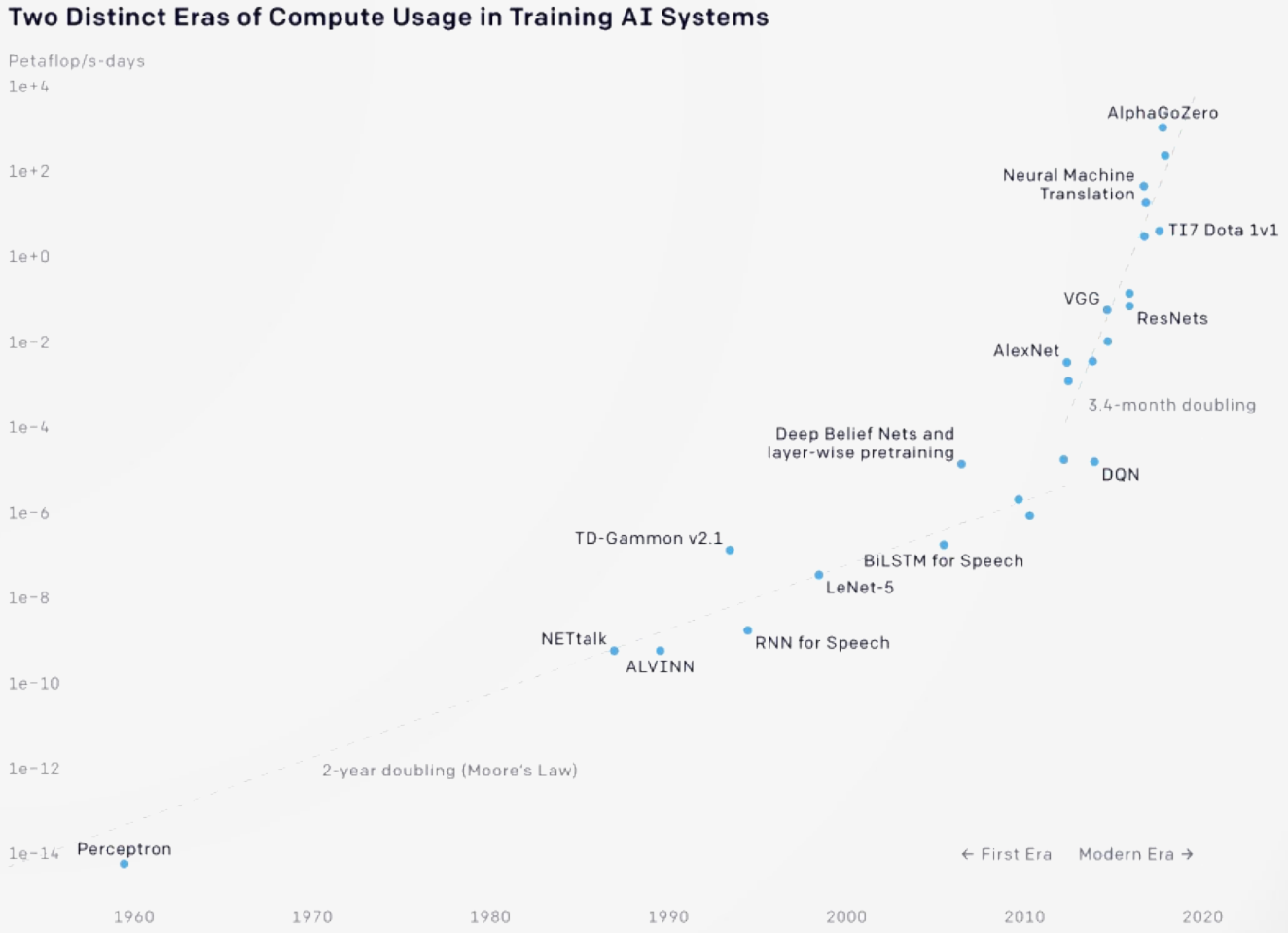
Cluster et machine learning

- Besoin en puissance de calcul important (notamment en nombre flottant)
- Explosion du deep learning (vision, traitement du langage, apprentissage par renforcement, ...)
- Augmentation de la taille des bases de données

Cluster et machine learning

- Processus itératif du développement des algorithmes de machine learning
- Besoin d'un retour rapide sur les expérimentations.
- Des modèles de deep learning récents peuvent prendre plusieurs mois d'entraînement sur un ordinateur classique.

Evolution des modèles



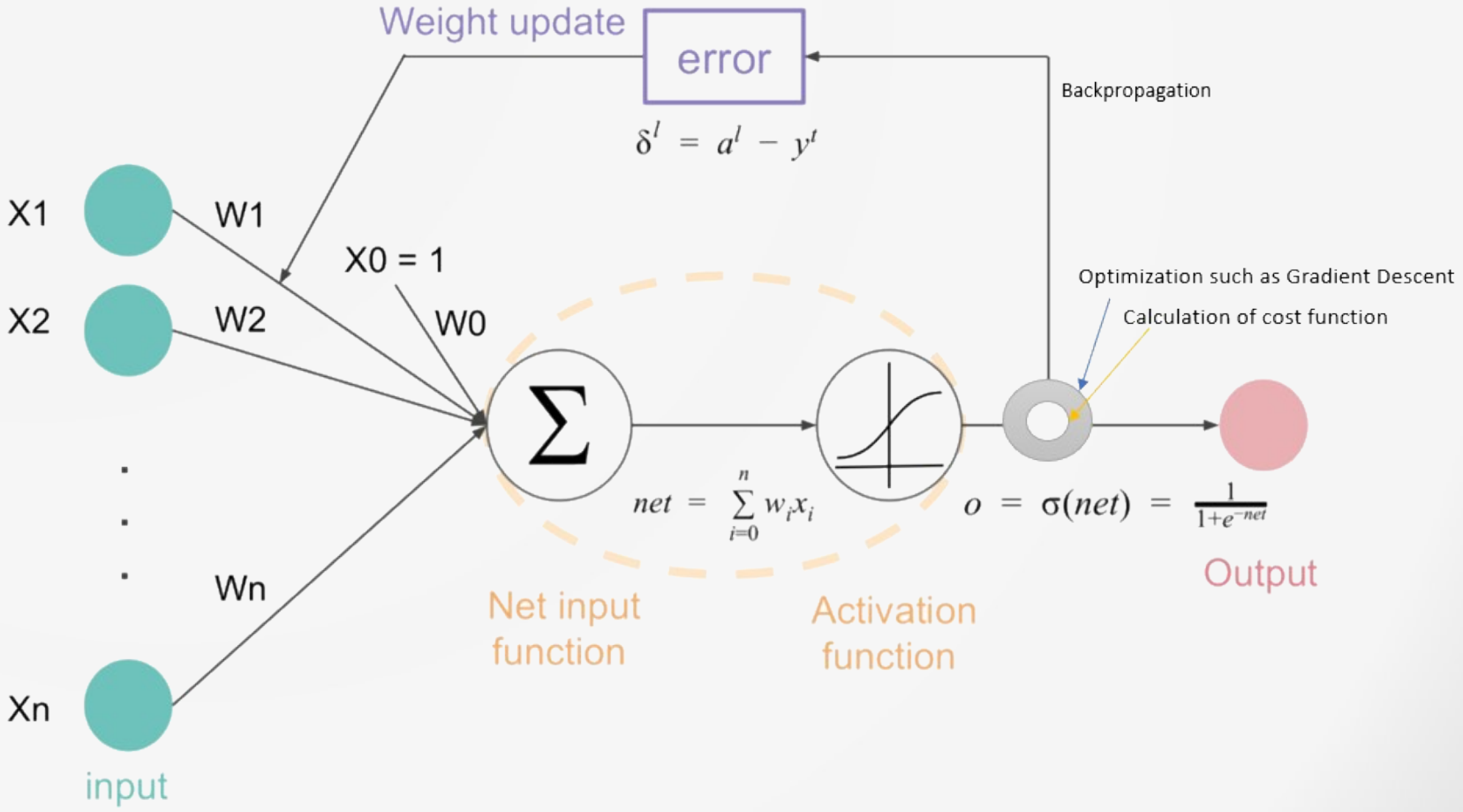
(Source OpenAI)

CATIE
Solutions pour la société numérique

L'utilisation des GPUs

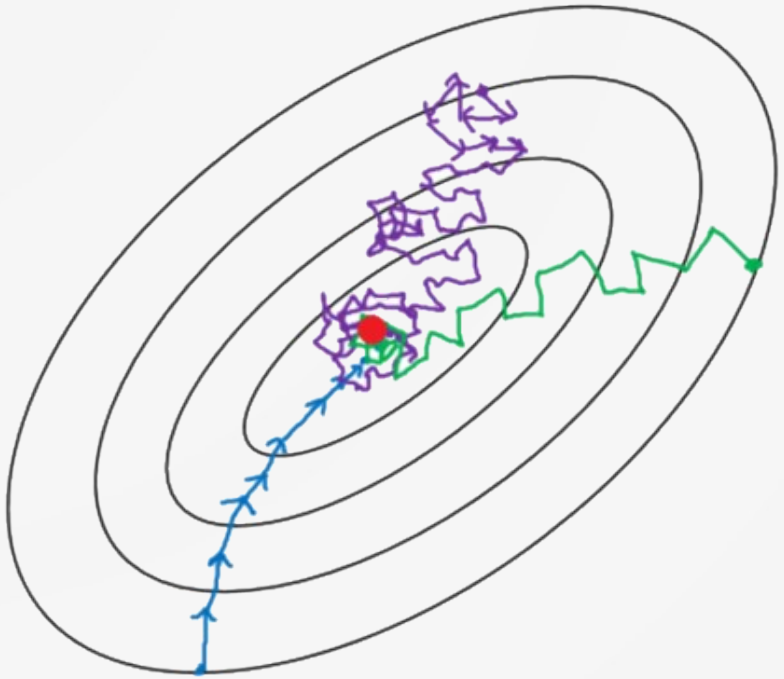
- Plus à l'aise avec le calcul en nombre flottant qu'un CPU
- Facilement parallélisable (sur des opérations simples)
- Possibilité d'avoir plusieurs GPUs par machine
- Universel

Deep learning (rappels)



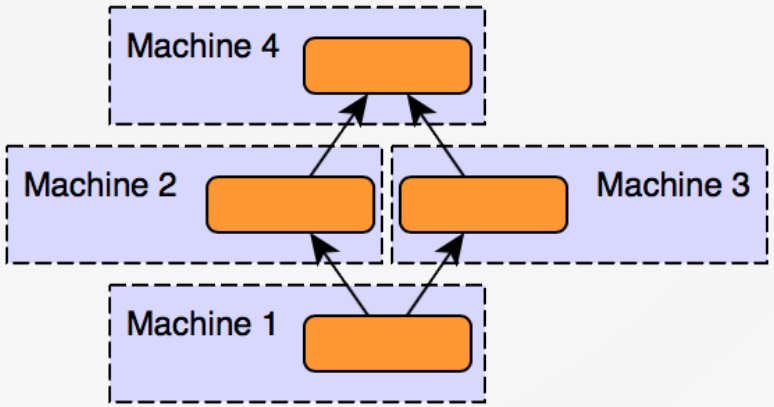
Deep learning (rappels)

A chaque itération, on calcule l'erreur pour un sous-ensemble des données, appelé minibatch.

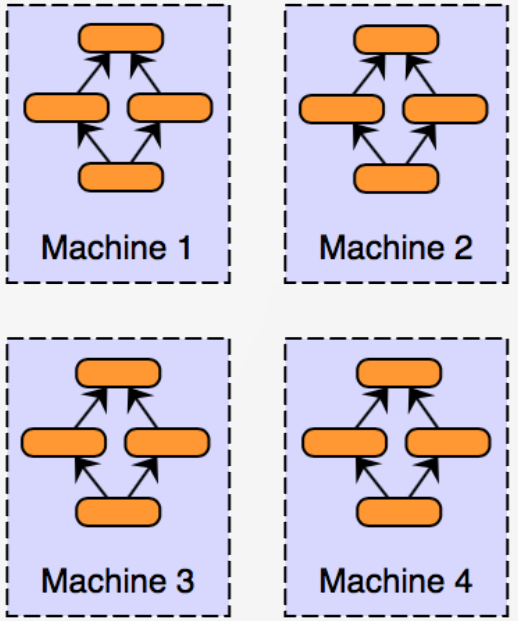


Deep learning distribué

Model Parallelism



Data Parallelism



Deep learning distribué

La parallélisation des données est la technique la plus couramment utilisée :

- le modèle est dupliqué entre toutes les machines (et/ou sur tous les GPUs d'une même machine).
- le coût de communication est assez faible.

Deep learning distribué

- La parallélisation du modèle est utile notamment, si le modèle ne rentre pas dans la mémoire d'un GPU.

- Le coût des communications est important (transfert des résultat en sortie des couches contenues dans le GPU).

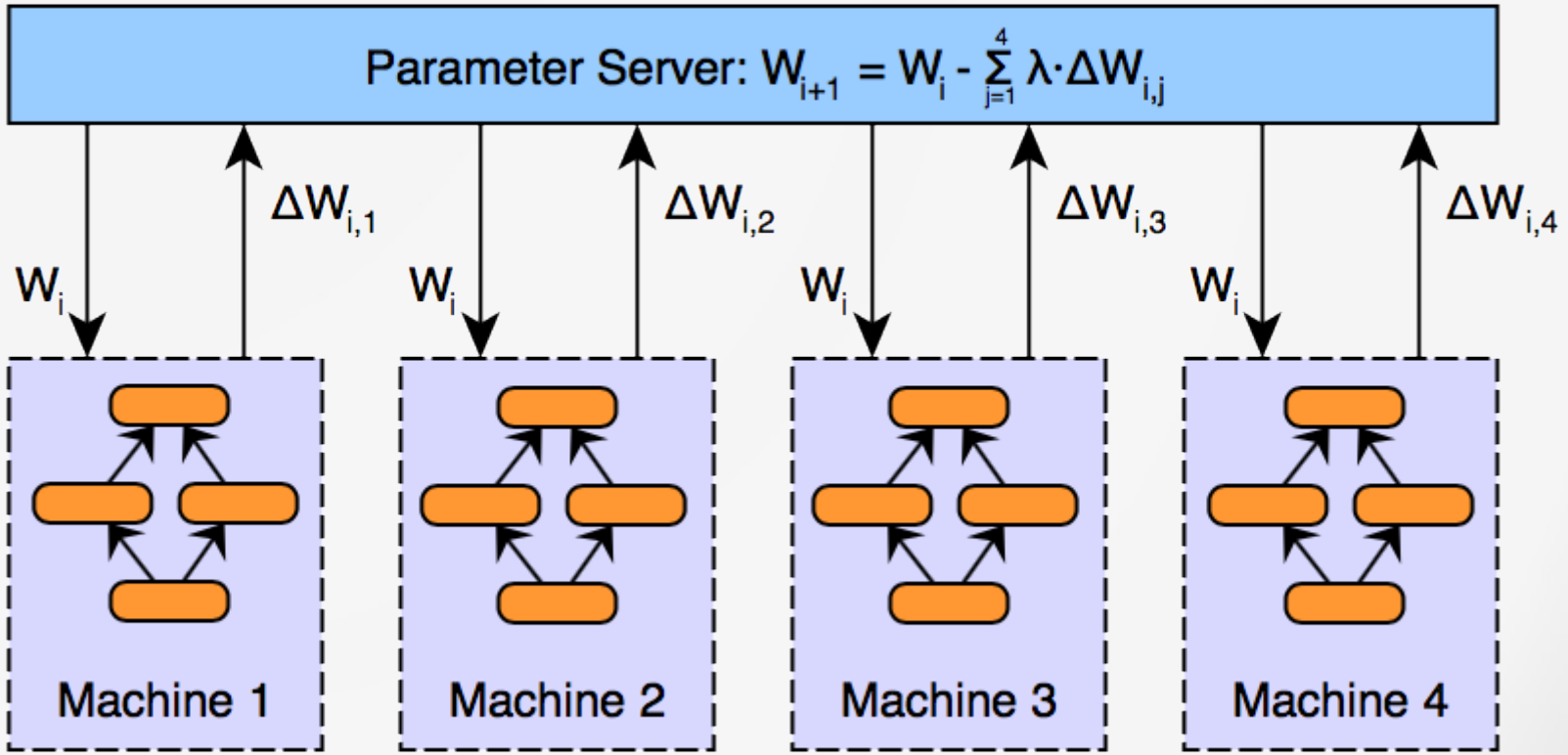
- De préférence, on utilise ce type de décomposition lorsque l'on a une communication rapide entre les machines. C'est le cas lorsque pour la communication inter-GPUs au sein d'une même machine.

CATIE
Solutions pour la société numérique

Descente de gradient distribuée

- Chaque machine traite une partie du minibatch.
- Les gradients sont mis en commun et moyennés.
- Les paramètres de tous les modèles sont alors mis à jour de manière synchrone.

Descente de gradient distribuée



CATIE
Solutions pour la société numérique

Descente de gradient synchrone distribuée

- Le coût le plus important corresponds à la mise à jour des modèles.
- Le temps de calcul d'un minibatch est limité par la machine la plus lente.
- On est aussi limité par la taille du minibatch. Si la taille du minibatch est trop petite, les GPUs seront sous-utilisés.

Descente de gradient synchrone distribuée

=> On cherche à augmenter la taille des minibatchs. Cela peut causer des problèmes dans l'apprentissage des modèles (manque de généralisation).

Ex: Fujitsu entraîne ImageNet en 74.7s sur 2048 Tesla V100 avec des "minibatchs" de taille 81920 en avril 2019.

CATIE
Solutions pour la société numérique

Descente de gradient asynchrone distribuée

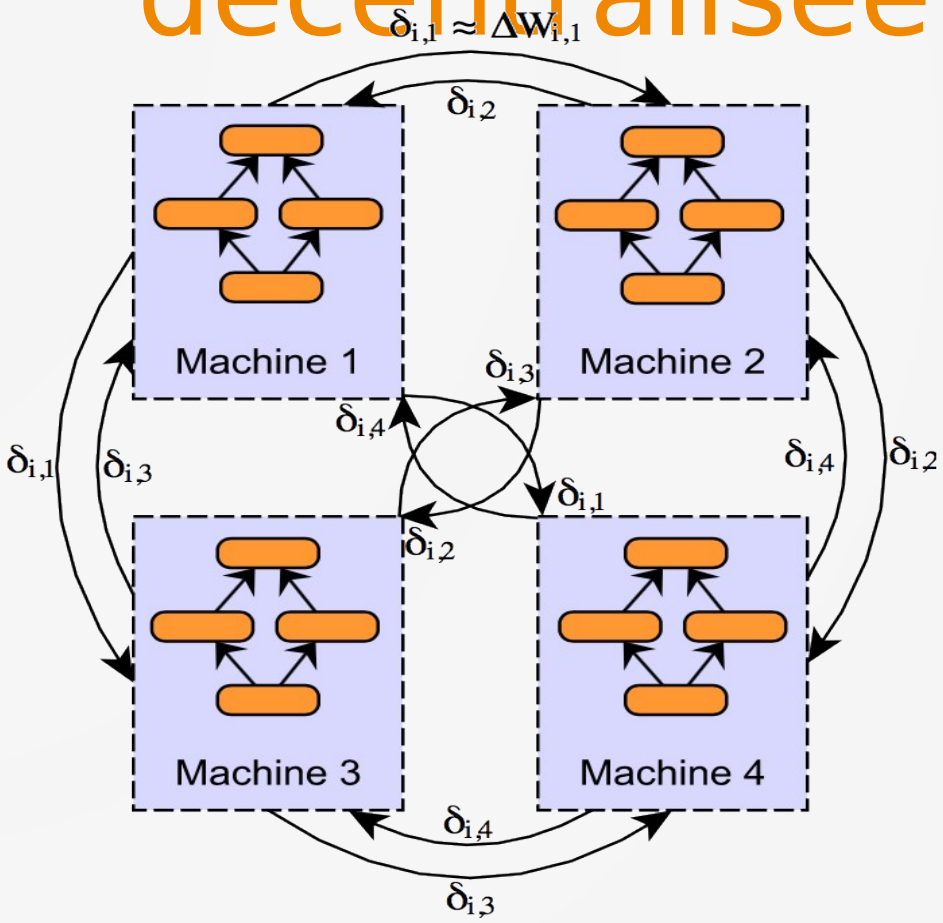
Pour limiter l'impact des différences de temps de calcul entre chaque machine, on peut utiliser une version asynchrone de SGD.

Les mise à jour des modèles sont réalisés de manière asynchrone à la fin du calcul par chaque machine.

Avantage: Utilisation maximale de chaque machine en fonction de ses performances.

Inconvénient: Accumulation de gradients calculés à partir d'ancienne version des paramètres.

Descente de gradient décentralisée



CATIE
Solutions pour la société numérique

MERCI POUR VOTRE ATTENTION

RÉGION Nouvelle-Aquitaine