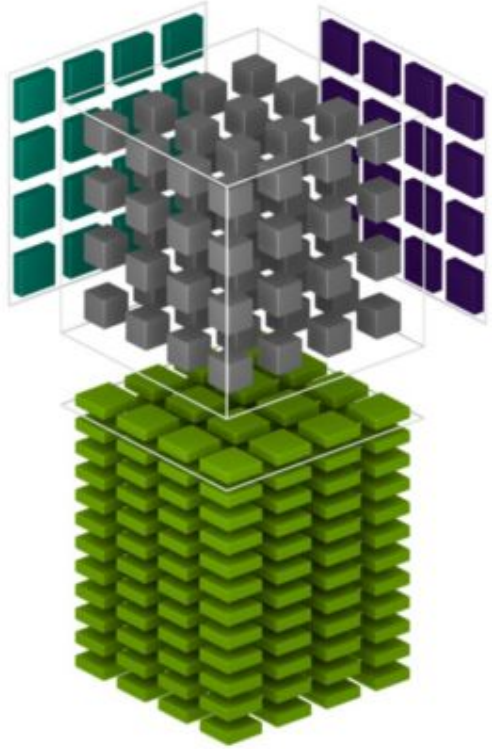
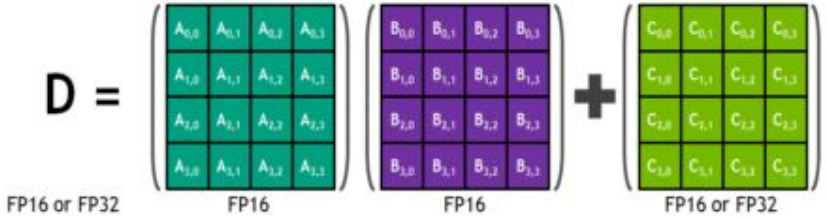
Tensor Cores



Les *CUDA Core* sont spécialisés pour le calcul vectoriel.

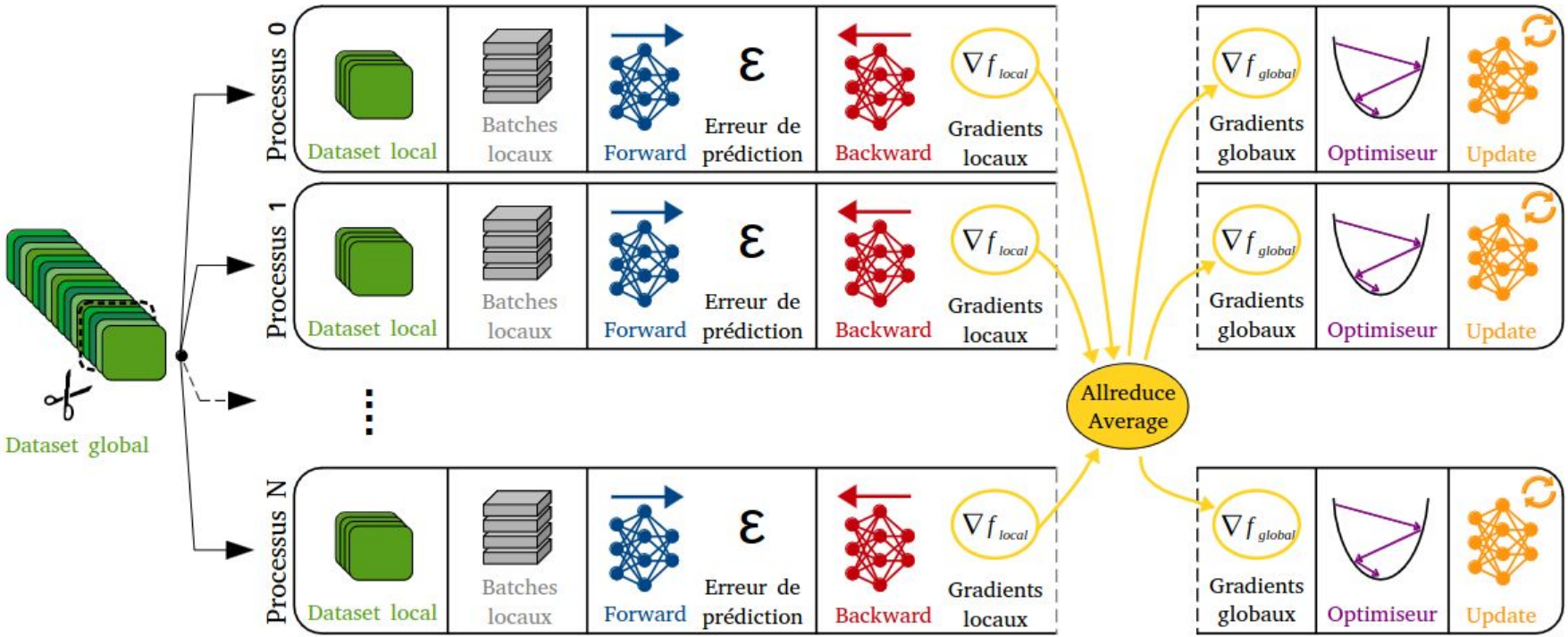
Les *Tensor Core* sont spécialisés pour le **calcul matriciel.**



Chaque *Tensor Core* est capable de traiter 64 opérations en 1 temps d'horloge.

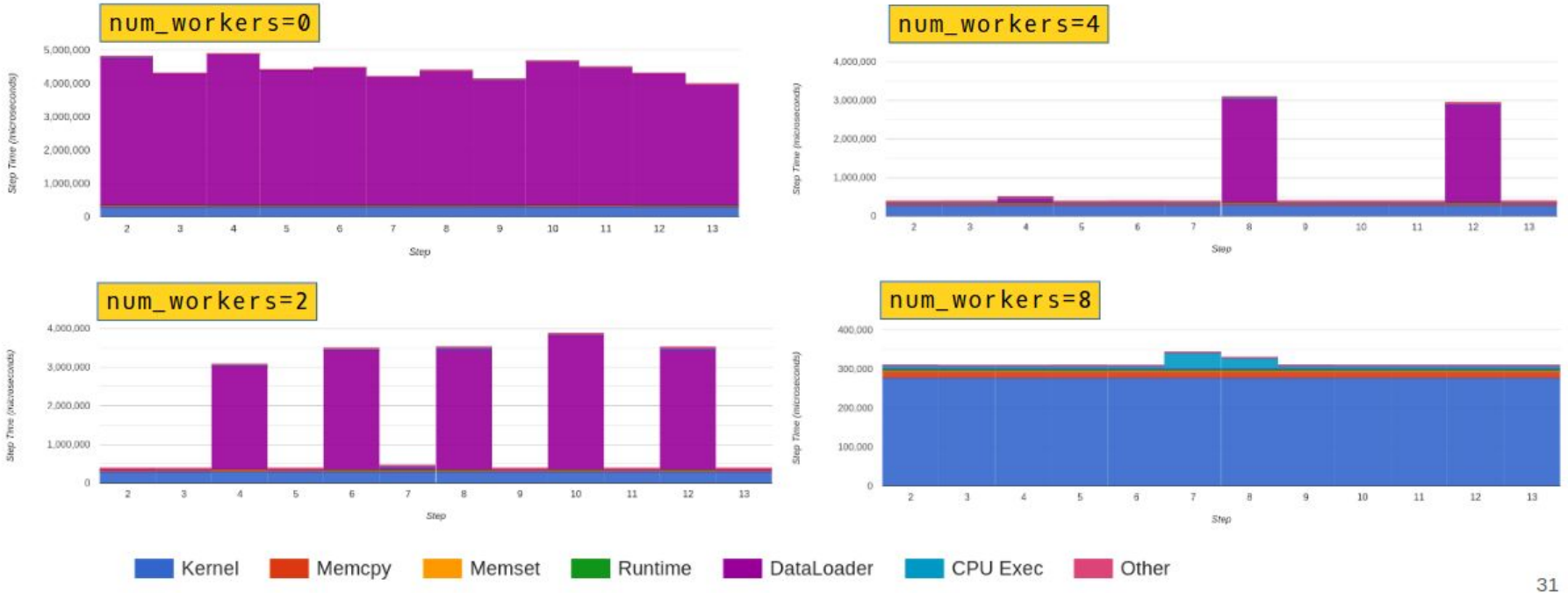
Source : NVidia

Distribution : Parallélisme de données



TP2_1 : Optimisation du DataLoader

- L'optimisation la plus efficace est l'augmentation du nombre de *workers*



Large Batches Rider

SGD AdamW

Weight Decay

LARS

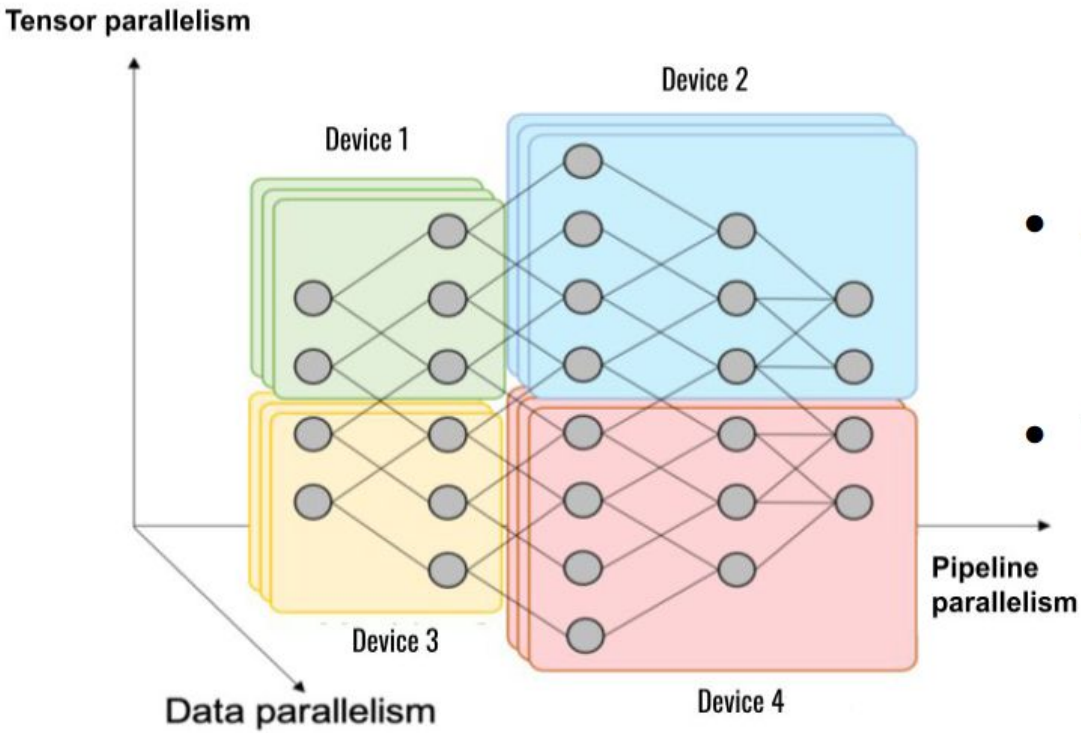
LAMB

Batch Scheduler

LR Scheduler

Warmup LR scaling LR Decay

3D Parallelism



- **Data Parallelism**
  - Simple à implémenter
  - Meilleure performance
  - Augmente la taille du batch (problème de convergence)
- **Pipeline Parallelism**
  - Effort d'implémentation.
  - Équilibre entre mémoire, performance et convergence.
- **Tensor Parallelism**
  - Effort important d'implémentation
  - Bonne accélération des calculs
  - **Bande passante très sollicitée** (implémentation Intra-nœud)

Conclusion

- Pay attention to **Dataloader**. It is often the bottleneck. Profile it, Optimize it, Boost I/O !
- **For usual size model**, use Data Parallelism, Tensor Core, Mixed Precision, Large Batch optimizations, on 1 or few nodes.
- **For huge model**, use dedicated library like Megatron-LM, Deepspeed, Colossal-AI on several nodes.