CRIANN

Retour d'expérience d'entrainement de modèles d'IA sur architecture HPC Data Parallelism et Model Parallelism

Benoist GASTON
JCAD 2024 - Bordeaux - 4-6 Novembre 2024

CCFR CENTRE DE COMPETENCE HPC.HPDA.IA.QUANTIQUE

Agenda

Présentation CC-FR - Criann

- Supercalculateur Austral

Apprentissage Distribué (travaux hiver 2023-2024)

- Data Parallelism avec Pytorch - Optimisation avec ZeRO
- Model Parallelism avec Pytorch

Conclusions et perspectives

EuroCC Criann - Austral

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia

Partenaires du Centre de compétence français :

Avec la participation de

EuroCC

Centre de compétence HPC

- Réseau de 33 centres de compétence nationaux
  - Favoriser l'usage du HPC et des technologies associées (HPDA, IA & Quantum)
  - Fédérer l'écosystème
  - Développer de formations
  - Accompagner les besoins et les demandes
- Programme d'accompagnement à l'usage du HPC par les mesocentres
  - Porté par Criann & Romeo, partenaires MesoNET
  - A destination du secteur public et du secteur privé
    - Collectivités, administrations
    - Industrie, PME, startups

CRIANN

Présentation

- Statut d'Association, créée en 1992
  - ESR (ComUE Normandie Université et établissements affiliés)
  - Rectorat, établissements de santé
  - Collectivités territoriales
- Mutualisation d'équipements et de services à haut niveau de performance
  - Réseau régional Syvik
  - Centre de données régional
  - Calcul intensif (HPC)
- Equipe : ~15 ETP
- Certification ISO27001 et HDS

Austral - caractéristiques octobre 2024

A100-SMX—80GB	
Stream Multiprocessors	108 (128 cores / SM)
Tensor Core	432 (4 TC / SM)
Bandwith	2 TB/s
Interconnect	NVLink 600GB/s
Flops DP	9.7 TFlops / 19.5 Tflops TC
Flops SP	19.5 TFlops / 156 Tflops TC
Flops HP	312 Tflops with TC

AUSTRAL SUPERCALCULATEUR POUR LA RECHERCHE EN NORMANDIE

Nœuds de calcul

- 88 GPUs Nvidia A100 80 GB SXM4
  11 nœuds – Apollo 6500 Gen 10+ DLC nodes
- 23 808 cœurs@2.4 GHz - 95 TB RAM
  124 nœuds AMD Genoa – 2x96 cores - 768 GB RAM DDR5
  - Cray XD 2000 Gen11 DLC
- Nœud SMP 6 TB de RAM DDR4
  Superdome Flex - 224 cores@2.6 GHz Intel Cooper Lake
- 8 GPUs AMD Mi210 (veille technologique)
  2 nœuds 4 x GPU - Apollo 6500 Gen 10+

Interconnection
Slingshot 200 Gbit/s

Stockage
2 Po (1 Po NVMe)

Accès à distance 100 Gbit/s
5 frontales de connexion
5 serveurs de visualisation
Environnements interactifs pour l'IA

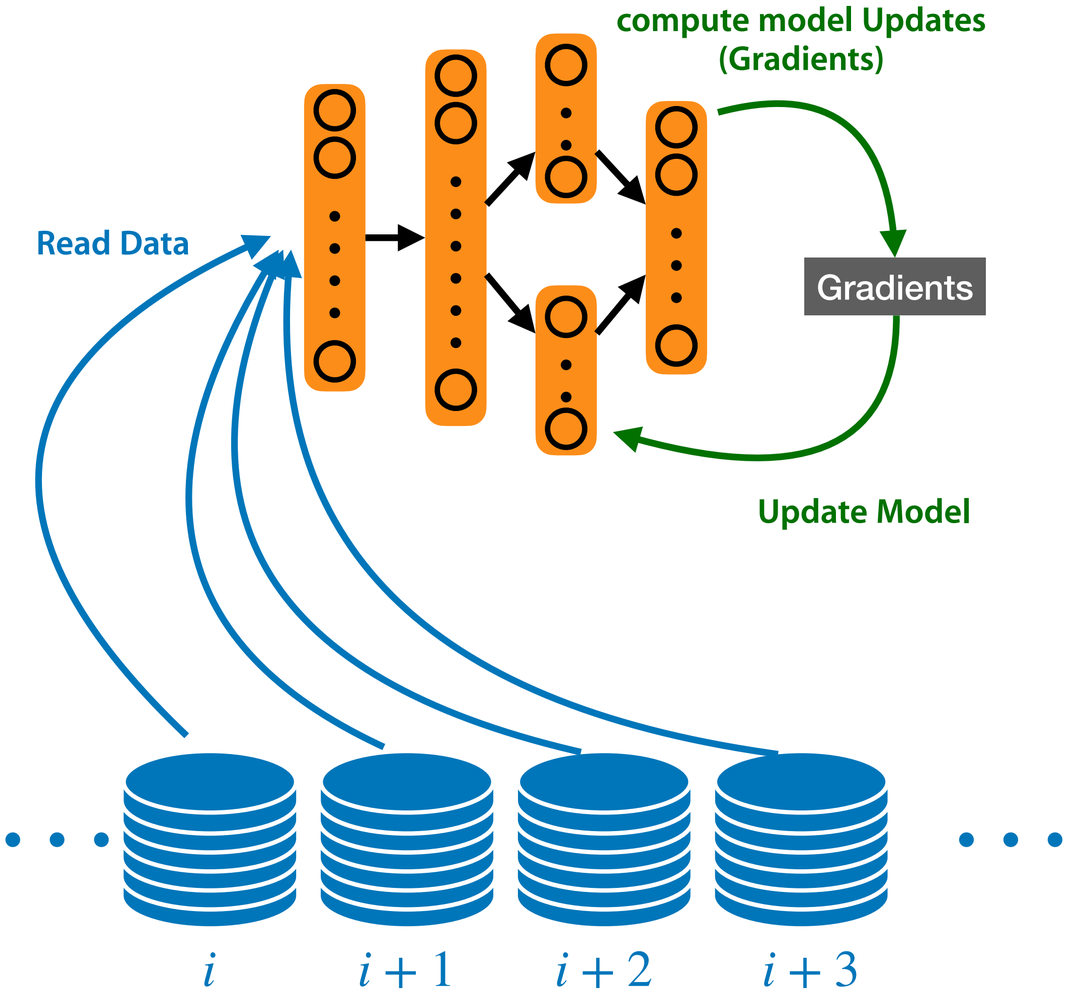
RedHat - Slurm - Lustre

CRIANN

Apprentissage Distribué

Apprentissage Distribué

Entrainement de modèle d'IA du point de vue HPC

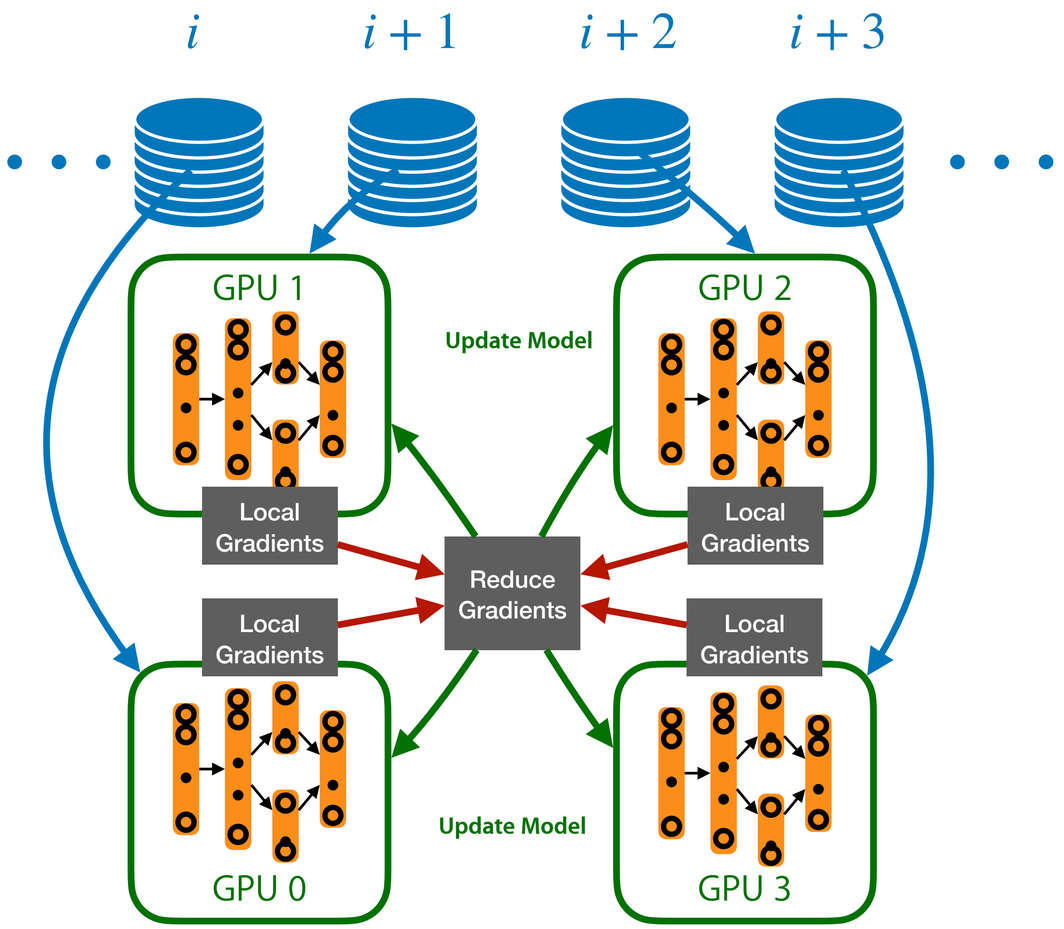


- Problèmes majeurs
  1. Modèle, Optimizer, Gradients : forte consommation mémoire GPUs => goulot d'étranglement
  2. Temps d'entraînement d'un modèle très long
     Généralement résolu avec le *Data Parallelism (DP)*
  3. Modèle trop volumineux pour tenir sur une unité de calcul (typiquement LLM)
     Généralement résolu avec la distribution de modèle *aka Model Parallelism (MP)*
- Au Criann
  - Data parallelism : régulièrement utilisé par quelques utilisateurs depuis plusieurs années
  - Model Parallelism : des demandes occasionnelles mais de plus en plus fréquentes

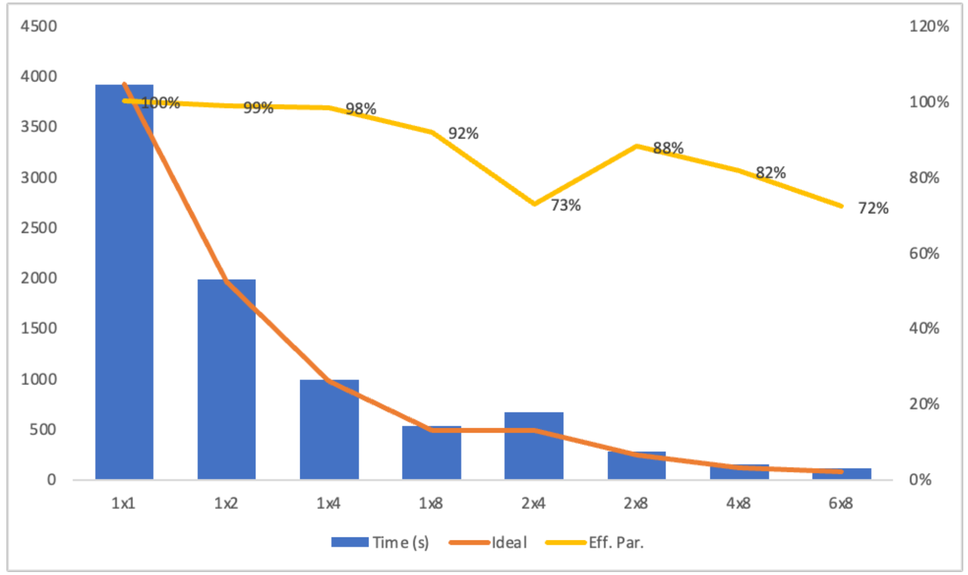
Data Parallelism

Data Parallelism (DP)

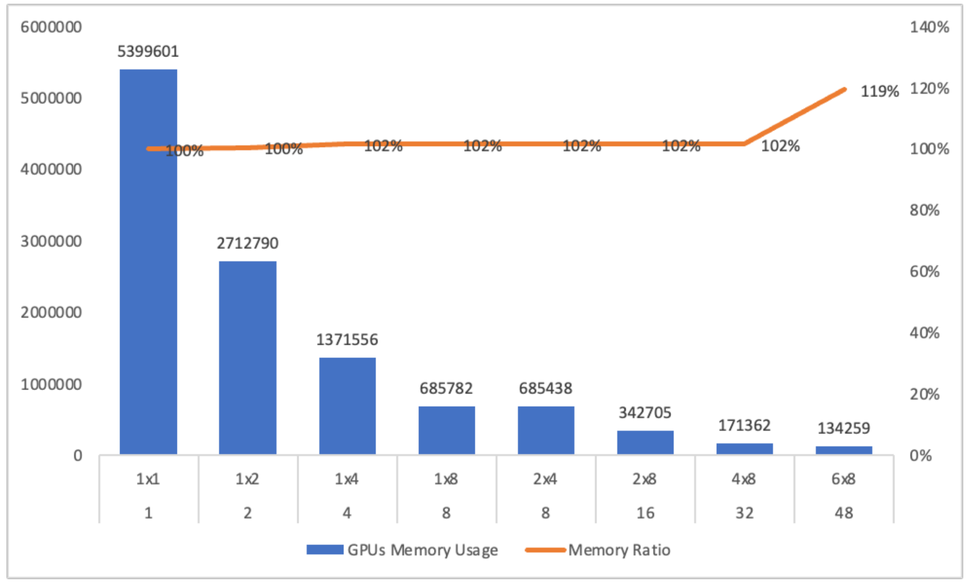
Principes



- Traitement des *batchs* données en parallèle
  - Distribution des données
  - Réplication du modèle sur chaque ressource
  - Calcul des gradients locaux
  - Agrégation des gradients à l'aide d'une réduction
- Avantage
  - Bonne efficacité de calcul => traitement de gros volumes de données
- Inconvénient
  - Mauvaise efficacité mémoire => taille des modèles limitée



Temps de restitution et efficacité parallèle en fonction du nombre de GPUs



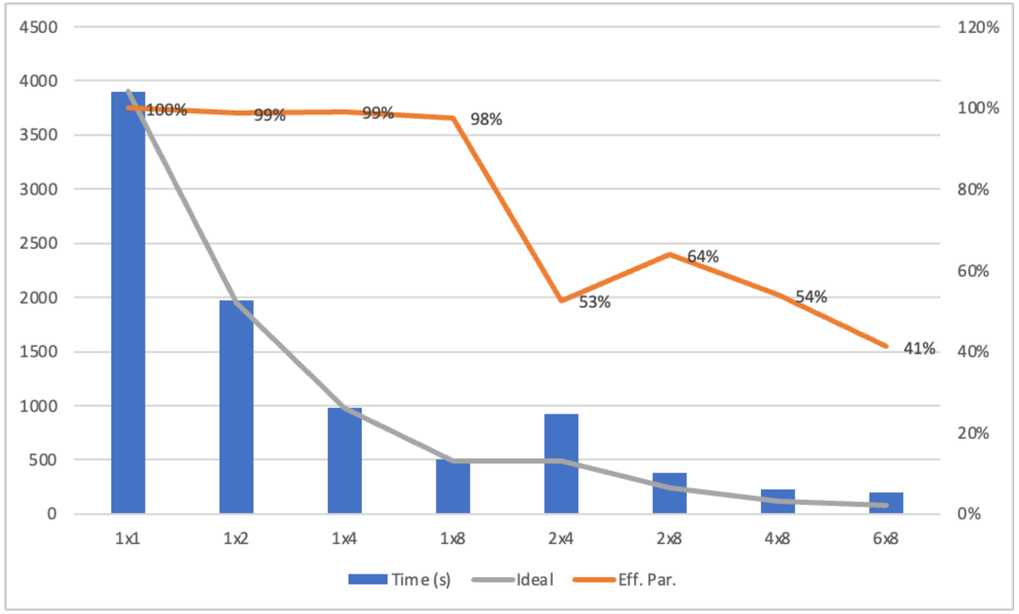
Consommation mémoire en fonction du nombre de GPUs

Data Parallelism (DP)

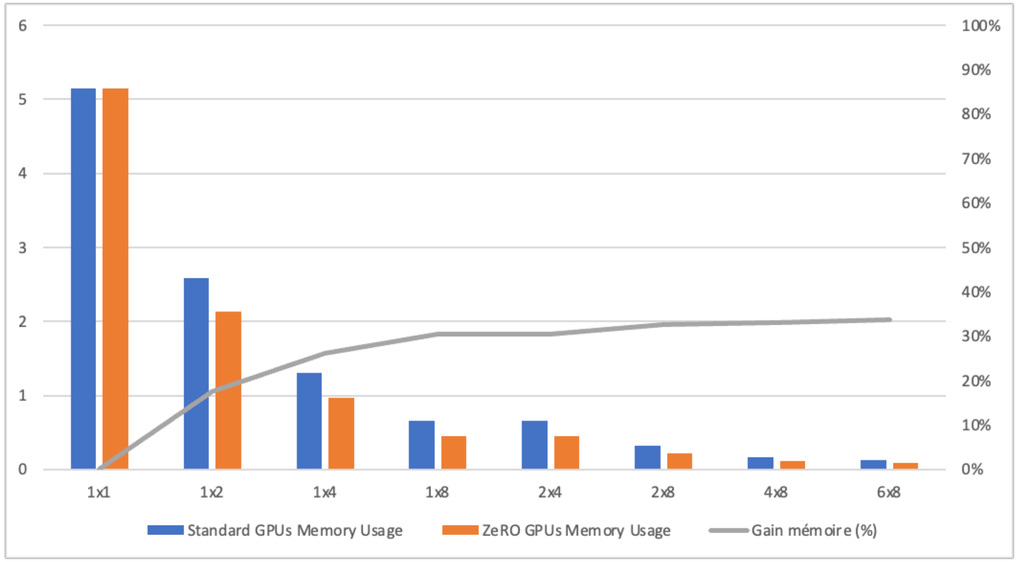
Tests et résultats

- Framework pytorch avec solution intégrée *DistributedDataParallel (DDP)*
  - Approche multi-gpus et multi-nœuds
    - 1 -> 48 GPUs A100
  - Intégration de SLURM (sous forme d'un package python)
  - Modèle gpu-2-XL Community
    - 1,5 Milliard de paramètres
  - Dataset
    - 10 000 items de Wikitext Graphcore (extrait des articles vérifiés de Wikipedia)

CRIANN



Temps de restitution et efficacité parallèle en fonction du nombre de GPUs



Consommation mémoire / méthode standard

Data Parallelism optimisation

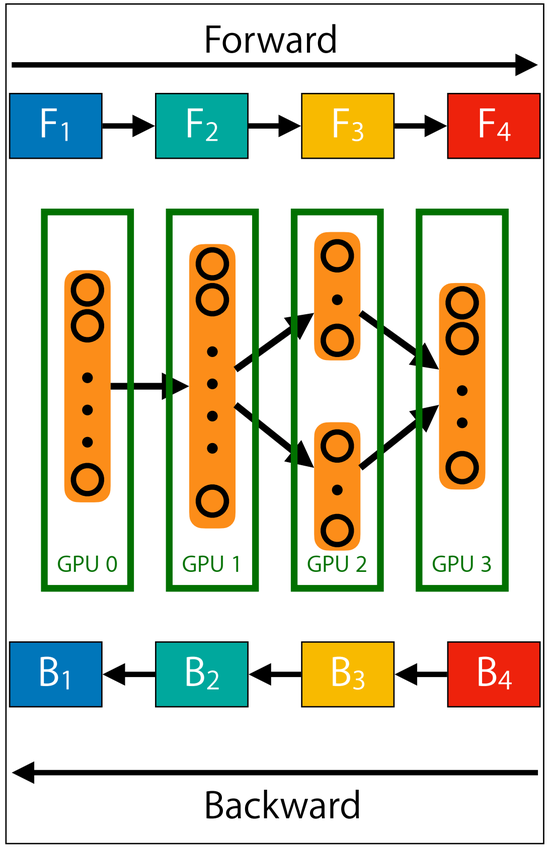
ZeRO (Zero Redundancy Optimizer)

- Optimiser l'entraînement de modèles massifs en machine learning
- Bibliothèque DeepSpeed (Microsoft)
  - Réduction de la consommation mémoire
  - Partitionnement / Distribution de l'*optimiser* et des gradients
- Tests et résultats
  - Comparaison avec le *Data Parallelism*

Model Parallelism

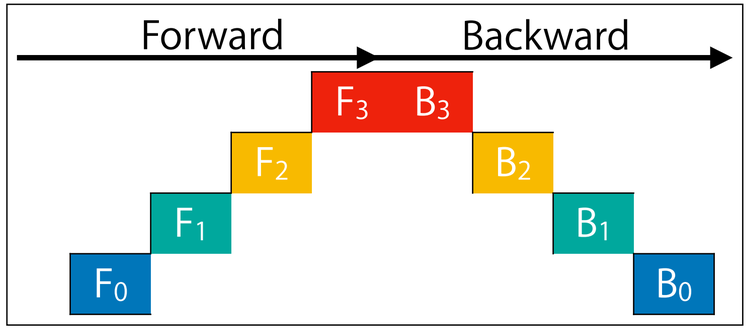
Model Parallelism (MP)

Principe

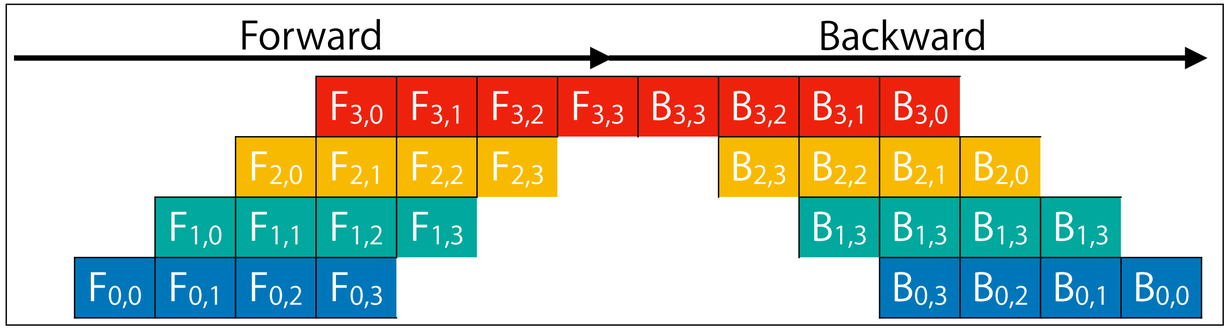


Distribution du modèle sur plusieurs GPUs

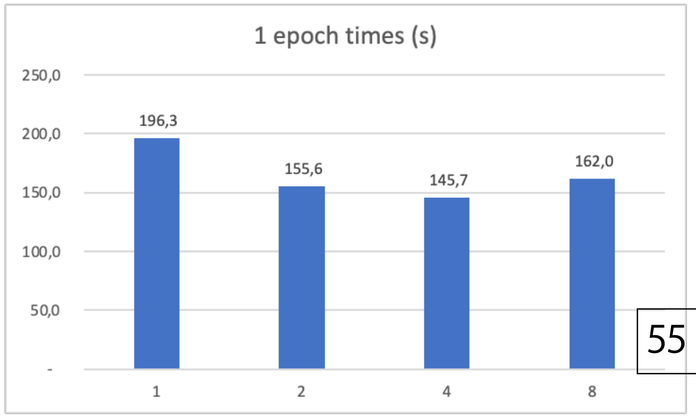
- Distribution d'un modèle sur plusieurs gpus
  - Traiter des modèles plus gros
  - Généralement découpage par couche
- Parallelisme par *pipelining* :
  - Division d'un *batch* de données en différent *chunks*
- Avantage
  - Bonne efficacité mémoire => traitement de très gros modèles
- Inconvénient
  - Faible parallélisation



Pipelining

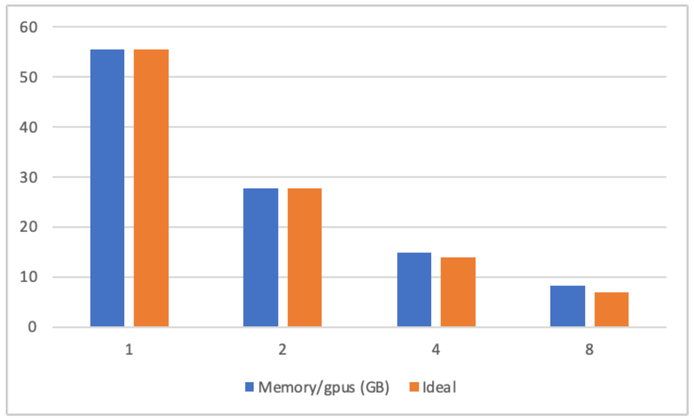


Pipeline Parallèle



55 GB/GPU

Temps par epoch en fonction du chunk : 8 gpus 52B paramètres



Empreinte mémoire par gpu - 6.5B paramètres

Modèle 55 GB

Temps par epoch en fonction du nombre de GPUs - 6.5B paramètres

Model Parallelism

Tests et résultats

- Framework pytorch (+Pipe pour le *pipelining*)
- Intranœud uniquement
  - 1, 2, 4, 8 GPUs A100
- Modèle-jouet avec architecture *transformers* de taille variable (nombre de couches)
  - 6.5, 13, 26, 52 milliards de paramètres
- Dataset
  - 2096 items de *Wikitext Graphcore* (extrait des articles vérifiés de Wikipedia)

Conclusions/Perspectives

Data Parallelism

- Conclusions
  - Régulièrement utilisé sur Austral
  - ZeRO efficace mais perte de scalabilité en multi-nœuds
  - À coupler avec les bonnes pratiques en chargement des données
- Perspectives
  - Mise en place d'un tutoriel
  - poursuivre l'étude de ZeRO
  - Comparaison avec d'autres méthodes/d'autres calculateurs

Model Parallelism

- Conclusions
  - Première approche satisfaisante
  - Parallélisation quasi-inexistante
- Perspectives
  - Poursuite des travaux sur cas d'usage réel

Questions ?

Le plateau de calcul intensif du Criann est cofinancé par la Région Normandie, l'État français et l'Union européenne (Fonds Feder).
MesoNET bénéficie d'un financement de l'Agence nationale de la recherche au titre des Investissements d'avenir.
Le Centre de Compétence EuroCC français est cofinancé par l'Union européenne et par l'État français.
Le réseau régional Syvik est cofinancé par la Région Normandie et par l'Union européenne (fonds Feder).
Le fonctionnement du Criann bénéficie du soutien de la Région Normandie.

RÉGION NORMANDIE

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

UNION EUROPÉENNE
Fonds européen de développement régional

CRIANN

Centre Régional Informatique et d'Applications Numériques de Normandie
www.criann.fr