Synthèse JEXP 2021 #2
Les GPUs

Table des matières

Lien vers l'indico : https://indico.in2p3.fr/event/23941/
Date : 30 avril 2021 à 9h30
Participants : 56

But de la session

Cette deuxième journée des expériences 2021 en visioconférence a pour but une rencontre entre les utilisateurs et les agents du CC-IN2P3 afin d'échanger sur les services disponibles ainsi que les besoins futurs des utilisateurs sur le sujet des GPUs.

Durant cette réunion, cinq courtes présentations suivies de discussion ont eu lieu. Le statut des GPUs au CC-IN2P3 a été présenté par Betrand Rigaud (CC-IN2P3), puis l'utilisation des GPUs avec les notebooks Jupyter a été présenté par Benard Chambon (CC-IN2P3). Ensuite, trois cas d'utilisation des GPUs au CC-IN2P3 ont été présentés, tout d'abord les besoins en GPUs du groupe LSST à l'APC par Bastien Arcelin (APC) puis le Deep Learning pour l'analyse des données de LISA par Natalia Korsakowa (SYRTE/OdP), enfin l'utilisation des GPUs par le groupe L2IT par Charline Rougier (L2IT).

Le statut des GPUs au CC-IN2P3 (Bertrand Rigaud, CC-IN2P3)

- # Résumé

1. Architecture des workers :
   a. K80 : 10 workers, 40 GPU au total, Infiniband interconnection, sortent de garantie en juin.
   b. V100 : 12 workers, 48 GPU au total, NO Infiniband interconnection.
2. K80 vs V100
   a. AI Benchmark utilisé au CC (https://pypi.org/project/ai-benchmark/#modal-close).
3. Accès :
   a. Accès au queue réglementé → besoin de faire une demande d'accès.
   b. Soumission classique sur Grid Engine (https://doc.cc.in2p3.fr/fr/Computing/job-types/job-gpu.html).
4. Logiciels fournies :
   a. Librairies : CUDA, OpenCL, OpenMPI
   b. Logiciel : Singularity
5. Statut de la ferme GPU :
   a. commande sur machine interactive pour voir l'état de l'utilisation des GPUs : $ gpu_info
6. Utilisation de la ferme :
   a. Ralentissement de l'utilisation des GPUs fin 2020.
   b. Bonne reprise de l'utilisation début 2021 avec l'arrivée des V100 2ème génération.
7. Hardware spécifique et R&D :
   a. Fournit une carte RTX8000 48GB pour l'équipe ATLAS au L2IT.
   b. Des besoins spécifiques comme celui-ci peuvent être étudiés par le CC-IN2P3.

Disponibilité des GPUs par les Notebooks Jupyter (Bernard Chambon, CC-IN2P3)

- # Résumé

1. Qu'est-ce que la plateforme Jupyter Notebook (JNP) ?
   a. Objectifs :
      i. JNP fournit des notebooks avec l'interface JupyterLab
      ii. Permet de développer et d'exécuter des codes de manière interactive pour plusieurs langages et fournit un terminal UNIX.
   b. Accès au service :
      i. accessible pour tous les utilisateurs avec les identifiants « calcul ».
      ii. Accès via un navigateur : https://notebook.cc.in2p3.fr
   c. Environnement d'exécution :
      i. Exécution de JNP avec les groupes primaire et secondaires.
      ii. Accès aux espaces HOME, THRONG, SPS et cvmfs
      iii. Différents langages disponibles avec un kernel spécifique :

        - Anaconda 3
        - Noyaux disponibles : ROOT via Python, ROOT via C++ Cling interpreter, R, Go, Julia : https://doc.cc.in2p3.fr/en/Getting-started/access/jn-platform.html
    d. Contrôle des ressources :
        i. Mémoire limitée à 2GB par utilisateur.
        ii. Arrêt de la session au bout de 72h
2. L'accès aux GPUs depuis la plateforme Jupyter :
    a. Objectifs :
        i. Fournir des notebooks s'exécutant sur une machine dotée d'un GPU.
        ii. Chaque utilisateur possède sont propre GPU (non partagé).
    b. Accès au service :
        i. Seulement pour les utilisateurs autorisés
        ii. Choix du modèle et du nombre de GPUs ainsi que la mémoire du serveur.
    c. Environnement :
        i. Notebook GPU s'exécuteront sur modèle K80.
        ii. Les logiciels suivant seront disponible : CUDA, PyCUDA, cuPy, Pytorch, TensorFlow + cuDNN
    d. Statut actuel :
        i. Un seul host avec GPU K80, 135GB de mémoire et 10Gbs d'ethernet.
        ii. Une image Docker avec CUDA 10.2, PyCUDA 2020.1, Pytorch 1.7.1, TensorFlow 2.3 + cuDNN 8.1.
        iii. Pour le moment l'accès est seulement réservé au « beta testeurs »
    e. Prévu :
        i. Nouvelle K80 en cours d'installation.
        ii. 10 hosts K80 prévu pour la fin de l'année.
    f. Comment tester ?
        i. Demande d'accès via le User Support : https://cc-usersupport.in2p3.fr
        ii. Se connecter via https://notebook.cc.in2p3.fr (et déconnexion obligatoire pour libérer le slot GPU).

• Discussion

- Patrice Lebrun : Quelle est la limite sur le nombre de beta testeurs ?
    - Bernard Chambon (CC) : 2 pour le moment. 4 possibles si 1 GPU par testeur. Mais c'est le tout début, à stabiliser avant d'ouvrir plus largement.

- Alexandre Boucaud (APC) : Retour en tant que testeur. Très agréable, libre d'installer n'importe quelle librairie python dans son espace utilisateur. Accès maintenu sur une certaine durée, permet de travailler sur plusieurs jours, connexion maintenue (et kernel aussi en cas de déconnexion).
  Par contre est-il possible de réserver quelques notebook sur une journée pour une formation ? Avec accès révocables pour les participants. Utilise google collab pour le moment.
    - Bernard Chambon (CC) : Ce n'est pas possible pour le moment, mais c'est envisageable dans le futur.

- Cécile Barbier : Prévoyez-vous l'achat de nouvelles cartes GPU ?
- Sylvain Caillou : Y a-t-il de nouveaux modèles de carte qui retiennent votre attention pour le futur ?
  - Bertrand Rigaud (CC) : On regarde pour remplacer les K80 pour grossir la ferme, on pense à des V100 ou bien la nouvelle génération A100 ? A voir suivant le coût. Mais ces achats dépendent de la demande.
  - Benoit Delaunay (CC) : On garde une homogénéité V100 sur 2021. A100 à voir plutôt pour 2022.

- Christophe Deroulers : Comment le rapport 4 CPU pour 1 GPU dans la ferme de GPU a-t-il été déterminé ?
  - Bertrand Rigaud (CC) : GE ne sait pas gérer des GPUs, il faut donc demander le bon nombre de CPU par GPU. Historiquement sur les K80 on a des GPU à 4 coeurs. Le rapport 4 CPU pour 1 GPU est indispensable. Limitation arbitraire, mais spécifique pour le bon fonctionnement de GE et au partage de ressource.

Besoins GPUs du groupe LSST à l'APC (Bastien Arcelin, LSST)

- # Résumé

1. LSST : Legacy Survey of Space and Time
   - Relevé de galaxies
   - Début des observations en 2022 pour 10 ans et 60 Pb de données.
2. Deux projets mené au CC :
   - Séparation de galaxies avec des AutoEncodeurs variationnels (https://arxiv.org/abs/2005.12039). TensorFlow v1 (CUDA - Singularity image)
   - Comparaison des IPUs Graphcore et des GPUs Nvidia pour les applications en cosmologie. TensorFlow v2 (CUDA v.10.1.105 - Singularity image)
   - Beta testeur notebook GPU : Estimation des paramètres de forme de galaxies (regression). TensorFlow v2 (CUDA v.10.1.105 - Singularity image).
3. Estimation utilisation GPU au CC-IN2P3 par l'équipe :
   - 2020 : ~ 1000h (K80 + V100).
   - 2021 : ~ 2000h (essentiellement V100).
4. Questionnement sur l'utilisation de nouvelle technologies au CC :
   - Nvidia A100 ?
   - Nouveau hardware : IPU, OPU, TPU ?

- # Discussion

  - Rachid Lemrani (CC-IN2P3) : Pourquoi parles-tu de ces nouvelles technologies ?
    - Bastien Arcelin (APC) : C'est simplement pour explorer ces nouvelles technologies. De plus, l'IPU donne un temps d'entraînement plus court. Et pour poser la question si ces technos étaient envisagé au CC.

- Ghita Rahal (CC-IN2P3) : Quel est l'effort à fournir pour passer d'une technologie à une autre?
  - Bastien Arcelin (APC) : TensorFlow permet maintenant d'adapter son code de façon très simple en quelques lignes.
  - Remarque : Alexandre : Voir la très bonne présentation au dernier workshop GPU, utilisation de ONNX pour abstraire un modèle de ML de n'importe quel framework (tensorflow -> pytorch par exemple). Gros efforts faits sur l'interopérabilité sur les framework. Même chose pour le hardware. Présentation sur ONNX lors du dernier Workshop GPU IRFU/IN2P3 [1] .

- Deroulers Christophe : Est-ce que vous avez une idée des performances relatives GPU et TPU ? (Merci pour la comparaison GPU – IPU).
  - Bastien Arcelin (APC) : je n'ai pas fait ces tests donc pas de réponse.

DL for LISA data analysis (Natalia Korsakowa, SYRTE/OdP)

- # Résumé

1. LISA = Laser Interferometer Space Antenna
2. Cet interféromètre permettra d'accéder à des fréquences en onde gravitationnelle plus haute que pour des interféromètres terrestres et donc de nouveaux phénomènes physiques.
3. Sources et bruits :
   a. Trou noirs binaires massifs.
   b. Binaire galactique compacte.
   c. Extreme Mass Ratio Inspirals.
   d. Trou noirs binaires d'origine stellaire.
   e. Fond diffus cosmologique. …
4. Méthode d'inférence
5. Estimation des paramètres :
   a. la réalisation d'une inférence d'un problème générale est impossible.
   b. Usage d'approximation :
      i. Markov Chain Monte Carlo/Nested Sampling
      ii. Variational inference
   c. Quelques exceptions pour des modèles simplifiés :
      i. Gaussian misxture models
      ii. Invertible models
6. Usage de Réseaux Neuronaux (NN) pour paramétrer les profils entre données observées et simulées.
7. Estimation du Jacobien avec hypothèse d'un Jacobien triangulaire pour accélérer l'estimation.
8. LISA Data Challenge : apprentissage de transformation conditionné par des données réels.
9. Conclusion :

a. Nouvelle manière de faire de l'inférence Bayesienne pour l'analyse de données d'ondes gravitationnelles.
b. La grande partie du temps de calcul consommé est due au temps d'entraînement des algorithmes.

- **Discussion**

- Rachid : Does the GPU usage will rise up in time ?
  - Natalia : The limit will be the memory of GPU in further works.

L'usage des GPUs par le groupe L2IT (Charline Rougier, L2IT)

- **Résumé**

1. L2IT : nouveau « laboratoire des deux infinis » à Toulouse
   a. Créé en « FRE » en 2019
   b. UMR en 2020
   c. Composé aujourd'hui de 9 scientifiques sur différents sujets : Physique des particules, Ondes gravitationnelles, Physique Nucléaire et Calcul, Algorithme et Données.
2. Graph Neural Networks (GNN) : Reconnaissance des motifs de traces au HL-LHC
   a. Limitation de la physique par l'efficacité des ressources logiciels et de calculs utilisés.
   b. Nouveaux efforts pour implémentés des algorithmes GNN réalistes pour le HL-LHC.
   c. Principe : arXiv:2003.1160
3. GNN :
   a. Nodes = hits ; edges = 2 hits successifs sur une trace
   b. Création du graphe de représentation et du graphe cible sur CPU.
   c. Étape d'entraînement et d'inférence sur GPU.
4. Besoin spécifique : avec une architecture de modèle simple :
   a. Étape d'entraînement :
      i. ~ 100 GB de mémoire
      ii. Temps d'exécution ~ 1 semaine et ~ 1 mois pour un modèle plus complexe.
      iii. Benchmark sur différents GPUs : pas de différence sur le temps d'exécution.
   b. Etape d'inférence :
      i. ~ 6 GB de mémoire
      ii. Temps d'exécution : 0,3 s/évènement.
5. Ce que l'on a appris :
   a. Limitation en mémoire : découper les tenseurs mais augmente le temps de calcul.
   b. Utiliser plusieurs GPUs pour accélérer le temps de calcul.

- # Discussion

- Ghita Rahal (CC) : Pourquoi y a-t-il une différence dans le benchmark avec la RTX ?
  - Charline Rougier (L2IT) : La RTX Quadro n'est pas du tout faite pour tourner avec de la précision 64 et notre modèle tourne en 32. Le temps d'exécution est donc comme attendu plus long.

- Alexandre : Augmentation de la batch size ?
  - Charline Rougier (L2IT) : Oui c'est vraiment quelque chose qu'on veut faire à l'avenir.

- Bertrand Rigaud (CC) : Avez-vous identifié ce dont vous aurez besoin pour l'inférence ?
  - Charline Rougier (L2IT) : Pour le moment on n'a pas observé de différence, mais on va faire de nouveaux benchmarks.

- Eric Fede (CC) : Etes-vous capable de quantifier ce que vous perdez quand vous passez sur des GPUs avec moins de mémoire ?
  - Charline Rougier) : Si on a un GPU avec beaucoup de mémoire, on peut accélérer notre training mais avoir un host CPU boosté est également très utile. L'optimal est d'avoir les deux.
  - Catherine Biscarat (L2IT) : On est actuellement en développement R&D, pour la suite on ira sur des modèles plus complexes qui prennent plus de mémoire, donc pour le moment on en est à l'étude et on n'a pas encore quantifié cela.

- Renaud Vernet (CC) : Temps d'attente GPU ? Volumétrie de l'infra GPU doit elle augmenter ?
  - Bastien Arcelin (APC) : Estimation du double en temps de calcul pour 2021 et 2022 par rapport à 2020. En ce début 2021 le temps d'attente sur la queue GPU est très correct.

  - L2IT : Peut être, des GPUs de bonne qualité pour la R&D et des plus bas de gammes pour l'inférence et la prod.

- Gillaume Baulieu : Peuton imaginer de connecter la ferme CPU et GPU ?
  - E. Fede (CC): Le modèle de batch actuel n'est pas compatible avec cette utilisation pour l'instant.
  - Bertand Rigaud (CC): possibilité de partager un GPU avec plusieurs utilisateurs → chez Nvidia il y a possibilité de le faire mais pas encore en place au CC et la mutualisation est envisagée au CC.
  - Alexandre : Le phase d'inférence des algorithmes peut être traitée sur CPU, il serait souhaitable de pouvoir stocker les poids de la phase de training effectuée sur GPU, afin de les migrer sur CPU. ML Flow ? [3]

Liens partagés :

- [1] Présentation ONNX https://indico.in2p3.fr/event/22938/contributions/93166/attachments/63109/86704/ML_KM3NeT_IN2P3_IRFU.pdf
- [2] Benchmarking TPU, GPU, and CPU Platforms for Deep Learning : https://arxiv.org/abs/1907.10701
- [3] MLFlow : https://mlflow.org/