

JCAD 2019 -Table ronde

"Evolution des usages des communautés SHS"

Céline Parzani
Tse-R / CNRS

Contexte TSE

- L'UMR TSE-R est composée de 150 chercheurs, 100 doctorants et 6 ingénieurs répartis en 11 groupes thématiques couvrant les différents aspects de l'économie (économétrie et économie empirique, économie expérimentale et comportementale, économie de l'alimentation, économie industrielle, économie publique, économie théorique, environnement et ressources naturelles, finance, macroéconomie, mathématiques de la décision et statistique, développement).
- En décembre 2017, création d'un **pole calcul scientifique**, composé de 2 ingénieures et d'un technicien
 - Assure l'**interface** entre les utilisateurs, le service informatique et les différents moyens de calcul internes ou externes
 - Gère les **moyens de calcul** : serveurs, cluster, abonnement cloud
 - Pilote les **projets autour du calcul scientifique** (formation, développement et optimisation de codes de calcul, expertise sur les méthodes numériques..)

Exemples de projets (1)

- Groupe "économie de l'alimentation" : analyse de la consommation alimentaire
 - **Données** : actes d'achat alimentaire d'un panel de ménages depuis 2001
 - **Méthodes** : modèle de demande de type choix discrets à utilité aléatoire, estimation des paramètres par maximisation de la vraisemblance, prédiction à partir de la simulation du modèle avec perturbations des paramètres
 - **Logiciels** : Matlab, Stata

- Groupe "économie publique" : calcul d'une fonction de probabilités d'élections inversées
 - Evaluation d'une fonction avec des combinatoires et des triples sommes infinies
 - **Logiciels/langages** : R, C++

Exemples de projets (2)

- Groupe « statistique et mathématique de la décision" : analyse des flux de transports aériens
 - **Données** : trajets entre aéroports (plusieurs milliers d'aéroports connectés entre eux)
 - **Méthodes** : modèle d'économétrie spatiale nécessitant l'inversion de matrice de très grande taille
 - **Logiciels** : Matlab, Stata
- Groupe "économétrie et économie empirique" : évaluation de politiques publiques du logement
 - **Méthodes** : Simulation d'un modèle du marché du logement, point fixe et équation de Bellman
 - **Logiciels/langages** : C++

Le calcul scientifique en économie

- **Diversité des outils mathématiques** : optimisation, statistiques, théorie des jeux, modèles agent, machine learning, EDO/EDP, Monte-Carlo, bootstrap, ...
- **Diversité des logiciels/langages** : Matlab, R, Stata, Mathematica, Python et les langages compilés, ...
- Problèmes de **grande dimension** (souvent plusieurs centaines de paramètres) stochastiques et non linéaires

Ce qui entraîne une **hétérogénéité dans les besoins HPC** comme par exemple

- mémoire vive de taille moyenne + nombre important de cœurs de calcul rapides adapté pour la réplication de calculs
- mémoire vive importante + nombre moyen de cœurs rapides pour la résolution de certains problèmes d'optimisation
- stockage important + mémoire vive de taille moyenne pour la manipulation de données de grande taille (plusieurs To)

Les difficultés et les besoins

- **Promouvoir le potentiel** du calcul scientifique en économie ce qui nécessite de travailler en amont sur des exemples probants pour convaincre. Le calcul est peu souvent mis en avant dans les publications du domaine,
- Nécessité de former les chercheurs aux **bonnes pratiques du calcul** : peu de codes parallélisés, expertise pour adapter les méthodes numériques et les logiciels au problème
- **Evolution du métier** (accès à de nouvelles données, évolution des ressources HPC, etc...) implique de mieux former les futurs doctorants
- **Grande variété dans la demande de formation** : Python pour le Big Data; R, Matlab des notions de base à l'optimisation des calculs avec la parallélisation, Machine Learning, OpenMP, MPI, portage sur GPU....
- Communauté un peu **réticente au dépôt de projets** pour accéder aux moyens mutualisés : difficulté à prévoir le planning des calculs, à évaluer les ressources nécessaires et besoin de support à l'utilisation

La question des données

- Bases de données de **grande taille** (plusieurs To)
 - Offrir des moyens de stockage adaptés
 - Grande diversité des données et fournisseurs
 - Développer, mutualiser des **outils génériques optimisés de pré-processing** des données
- **Confidentialité des données**
 - certaines (exemple CASD) ne sont accessibles qu'en **accès restreint et contrôlé** ce qui peut poser problèmes pour utiliser des ressources mutualisées
 - nouvelles règles liées au **RGPD**

Moyens de calcul

- Plusieurs **serveurs sous Windows**, certains mutualisés (doctorants, chercheurs, ..)

Nom	Mem (Go)	Proc.	Cœurs (log.)	Freq (Gh)	Stock. (Go)
Hal	192	X5687	24	3,6	276
Cassio	292	E7-4870	80	2,4	1000
Hobbes	192	E7-4870	16	2,4	276
Makalu	512	ES-2697	56	2,6	278

d'autres réservés à certains groupes thématiques car financés sur des projets

- Depuis juin 2019, **ANUBIS cluster de calcul** sous Linux, 5 noeuds de calcul (BI Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz) avec 28 coeurs et 192 Go de RAM par noeud.
- Un abonnement au **cloud Microsoft Azure** (utilisé notamment pour tester des GPUs)
- Deux projets ont été acceptés pour des heures de calcul à **CALMIP**
- Certains chercheurs utilisent des ressources extérieures (autres universités par exemple)

Exemples de solutions HPC autres labos

➤ University College of London

- 10 "big" servers (12 processors each, 96GB or memory each)
- 250 small servers (4 processors each, 16GB memory each)
- Stockage disque
- GPGPU cluster
- Une équipe d'une dizaine de personnes assure le bon fonctionnement

➤ LSE : London School of Economy

HPC 1	HPC 2
Intel v3 Haswell processors	Intel v4 Haswell processors
24 CPU cores	28 CPU cores
128GByte memory	1 TByte memory
1 TB local storage	1 TB local storage



Toulouse
School of
Economics

Economics
for the
Common Good

Manufacture des Tabacs
21, Allée de Brienne
31015 Toulouse Cedex 6
France

www.tse-fr.eu

T. +33(0)5 61 63 36 90
F. +33(0)5 61 63 35 86

