Calcul : expliciter l'offre de calcul française et européenne et comment l'implémenter

Alain Franc[1] et Jean-Marc Frigerio[2]

Résumé. Le monde académique français et européen offre à l'ensemble des équipes de recherche et d'enseignement un accès à une diversité d'infrastructures pour le calcul hautes performance ou hautement distribué, différenciées selon les besoins croissants liés à l'émergence des nouveaux domaines d'application du calcul. Il est classique de distinguer trois types d'infrastructures : en HPC (high performance computing, avec forte parallélisation) : *(i)* les centres nationaux, administrés par GENCI, qui offre des machines avec plusieurs dizaines de milliers de cœurs, au niveau dit Tier 1 *(ii)* les centres régionaux, appelés mésocentres, qui pavent le territoire, et offrent des machines avec plusieurs milliers de cœurs, au niveau dit Tier 2, avec des chartes souvent propre à chaque mésocentre et *(iii)* en HTC (high Throughput computing) la grille de EGI (european grid infrastructure) qui élargit ses services au cloud et permet de distribuer un calcul sur un réseau de clusters. Après avoir rappelé quelques notions simples sur les parallélisations possibles d'un calcul, nous présentons les différents modes d'accès à ses infrastructures, ainsi que des indications pour le choix vers telle ou telle infrastructure en fonction de la nature et de la taille du problème posé. Enfin, ce paysage est en train d'évoluer rapidement par l'arrivée de besoins massifs en calcul et données liés au « deep learning » qui demande à la fois du calcul intensif et un accès rapide à de très grandes bases de données, avec des architectures dédiées.

Mots clés. Calcul scientifique, calcul intensif, infrastructures de calcul, parallélisation, données massives

Introduction

Hier, les prévisions météorologiques se faisaient avec une expertise forte pour l'état du ciel et des nuages ; aujourd'hui, les plus gros ordinateurs mondiaux sont au service de calculs et modèles numériques pour la prévision météorologique. Des capteurs de plus en plus diversifiés déversent en continu des masses énormes de données (images, séries chronologiques, séquences, signaux, etc.). Le calcul intensif, issu des besoins en résolution numérique des équations de la physique ou de la chimie, en recherche et technologie, se développe puissamment dans des domaines nouveaux d'analyse de données massives, que ce soit en analyse d'images, en méthodes d'apprentissage, en fouille de données. Les données devenant massives, le calcul devient intensif. Ainsi, selon le site de GENCI, organisme national responsable de la mise en œuvre et du développement du calcul intensif, les principaux domaines y ayant recours sont (voir http://www.genci.fr/fr/content/applications-majeures) : la modélisation du climat et de l'environnement, la médecine, la chimie et la biologie (conception de médicaments, dynamique moléculaire), des applications industrielles liées à l'automobile, à l'aéronautique et au spatial, à la finance, sans oublier l'astronomie, l'astrophysique et la physique des hautes énergies.

Lorsqu'on effectue un calcul sur son ordinateur personnel, on remarque assez facilement qu'il existe deux facteurs limitants qui peuvent en bloquer l'exécution :

[1] *BIOGECO, Inra, Université Bordeaux, 33610, Cestas, France*
[2] *Inria, Équipe Pleiade, 33405, Talence, France*
alain.franc@inra.fr

- le temps de calcul, qui peut devenir prohibitif (quiconque a réalisé une phylogénie par maximum de vraisemblance il y a une quinzaine d'année peut en témoigner…) ;
- la quantité de mémoire requise, phénomène d'autant plus courant de nos jours avec l'explosion des jeux de données massives.

Il existe trois types de solution pour franchir ces obstacles :

- changer d'architecture : distribuer les calculs et les données sur plusieurs éléments de stockage ou de calcul, en parallèle ;
- changer d'algorithme : l'optimisation des codes pour minimiser le temps de calcul ou la taille des données stockées même temporairement est une phase importante du développement (voir par exemple les algorithmes de programmation dynamique) ;
- programmer des heuristiques : lorsque le calcul de la solution exacte est trop coûteux, construire une solution approchée dont le temps de calcul est significativement plus court que pour la solution exacte (troquer de la précision contre du temps, voir par exemple les algorithmes gloutons).

Dans la suite de ce texte, nous nous intéresserons à la première solution : changer d'architecture machine ou logicielle, notamment *via* la parallélisation.

Les principes de la parallélisation

La parallélisation d'un programme consiste à allouer des tâches partielles d'un programme à des processeurs différents, de façon simultanée et ponctuellement indépendante. L'avantage essentiel est la mutualisation des ressources (dont le temps) entre plusieurs processeurs et/ou machine. On parle souvent de « passage à l'échelle » à savoir que, grâce à la parallélisation, un même problème peut être résolu à une échelle plus grande (maillage plus fin en cas de simulation numérique, taille du jeu de données plus grande en traitement des données, etc.). Il est clair que, pour la cohérence du programme global, les processeurs doivent à un moment ou un autre communiquer pour recevoir et envoyer des instructions et données. La gestion de cette communication, gourmande en temps, est un des points délicats de la parallélisation. Il existe ainsi plusieurs types de parallélisation : à mémoire distribuée (MPI) par gestion d'envoi de message ; à mémoire partagée (OpenMP) avec une même RAM pour plusieurs processeurs. On peut paralléliser les tâches (plusieurs tâches en parallèle sur un même jeu de données, MPI), ou les données (une même tâche sur plusieurs jeux de données, OpenMP). En fait, un cluster de calcul est classiquement construit comme un ensemble de machines (les nœuds), chacune étant constituée de plusieurs cœurs (ou CPU), habituellement 12 ou 24, qui ont accès à une même RAM (mettons 128 Go). Aussi, la parallélisation peut se faire par mémoire partagée sur un nœud (OpenMP) ou communication entre les nœuds (MPI). Le passage à l'échelle de plusieurs ordres de grandeurs se réalise donc avec une communication entre les nœuds par MPI. Il est plus facile d'assurer le passage à l'échelle en se cantonnant à des directives MPI de programmation entre nœuds.

La parallélisation d'un code fait partie de son développement, et est possible pour certains langages uniquement, classiques en calcul scientifique : Fortran, C, C++ et python. Il existe d'excellentes formations à la parallélisation des codes, que ce soit MPI, OpenMP, ou les deux (voir https://cours.idris.fr/php-plan/affiche_planning.php?total).

Centres de calcul et data centers

La parallélisation demande de concentrer en un même lieu (un centre de calcul) un très grand nombre de machines qui communiquent entre elles de façon très rapide, pour minimiser les coûts de communication. Une architecture courante pour cela est d'installer des bus InfiniBand qui permettent un réseau de type Ethernet à très hautes performances. Mais cela a un coût. Aussi, tant les coûts importants d'investissement (machines, communications…) que d'entretien (énergie, fluides…) des ordinateurs puissants ont rapidement conduit à une mutualisation de ces équipements entre plusieurs utilisateurs, notamment par le développement de centres de calculs, aujourd'hui qualifiés de « data centers ».

Le paysage national en calcul intensif est organisé selon une pyramide classique de ces centres, selon leurs capacités, dite des Tiers :

- cluster de laboratoire : Tier 3
- centre régional de calcul (mésocentres) : Tier 2
- centre national de calcul : Tier 1
- niveau européen : Tier 0 (PRACE)

Les principales caractéristiques de ces centres sont rappelées dans le Tableau ci-dessous.

Taille du jeu de données	Machine	Nombre de CPU	Disponibilité
≈ 10^3	Ordinateur personnel	1 - 8	Immédiate
≈ 10^3	Tier 3 (cluster de laboratoire)	10 - 128	Gérée au sein d'un laboratoire
≈ 10^4	Tier 2 (mésocentre)	> 1000	« Fair share » ou appel d'offre
≈ 10^5	Tier 1 (national)	> 20 000	Appel d'offre

Nb : la taille du jeu de données est estimée comme la taille d'une matrice pour un calcul matriciel.

Nous aborderons dans la suite de ce texte les façons d'accéder à la mise en œuvre de calculs sur des centres nationaux ou régionaux (Tier 1 et Tier 2). Les clusters de type Tier 3 sont en général gérés au niveau de laboratoires, ou de fédérations de laboratoires, ou rassemblés sous forme de « data centers » où le coût de l'hébergement, des fluides, et de la sécurité sont mutualisés (l'Inra met ainsi à disposition deux data centers, un à Toulouse et un en Ile-de-France).

Niveau national

L'ensemble des moyens de calcul intensif à vocation nationale est géré par un organisme, le *Grand équipement national de calcul intensif* ou GENCI (http://www.genci.fr). Le GENCI a en charge à la fois de mettre à la disposition des communautés scientifiques françaises l'accès au calcul intensif via les centres nationaux, mais également de faciliter et développer son utilisation.

Il existe quatre centres nationaux :

- l'IDRIS, au Cnrs, à Orsay ;
- le CINES, qui dépend du ministère de l'Enseignement supérieur, de la recherche et de l'innovation, à Montpellier ;
- le TGCC, qui dépend du CEA, à Bruyère le Chatel ;
- le CC-IN2P3, qui dépend du Cnrs (IN2P3), à Lyon.

Le CINES : Centre informatique national de l'enseignement supérieur (https://www.cines.fr/)

Les missions nationales du CINES sont :

- le calcul numérique intensif,
- l'archivage pérenne de données électroniques,
- l'hébergement de plates-formes informatiques d'envergure nationale.

L'IDRIS : Institut du développement et des ressources en informatique scientifique (http://www.idris.fr/)

L'IDRIS est le centre majeur du Cnrs pour le calcul numérique intensif de très haute performance. À la fois centre de ressources informatiques et pôle de compétences en calcul intensif de haute performance, il participe à la mise en place

de ressources informatiques nationales, au service de la communauté scientifique publique. Il comprend également un service d'aide aux utilisateurs étoffé, qui facilite grandement l'accès à ses ressources de façon optimale.

Le TGCC : Très grand centre de calcul du CEA (http://www-hpc.cea.fr/fr/complexe/tgcc.htm)

Le CEA accueille la première machine pétaflopique française : Curie, qui concourt à la satisfaction de besoins scientifiques propres aussi bien qu'externes. Il héberge aussi le CCRT (Centre de calcul dédié à la recherche et la technologie).

Le CC-IN2P3 : Centre de calcul de l'IN2P3 (à https://cc.in2p3.fr/**)**

Le centre de calcul de l'IN2P3, unité de service et de recherche du Cnrs, est orienté vers le traitement des données massives produites par les grands instruments scientifiques (physique des hautes énergies, télescopes, etc.). Il ouvre ses ressources également aux sciences de la vie et aux sciences humaines et sociales, qui font de plus en plus appel au traitement des données massives (big data). Il s'agit d'un centre de calcul dit HTC (high throughput computing).

L'accès aux centres nationaux sauf IN2P3 se fait de façon commune *via* GENCI et sa procédure DARI (http://www.genci.fr/sites/default/files/Livret-information-Genci.pdf).

Il s'agit de soumettre un dossier comprenant deux volets : un volet scientifique présentant la pertinence de recourir au calcul intensif, et un volet informatique montrant la compétence de l'équipe pour utiliser les machines des centres nationaux.

Seuls des projets nécessitant un très grand nombre d'heures de calcul (souvent quelques millions) sont pertinents pour un accès à un centre national. Aussi, cette pertinence se montre en général *via* un passage progressif à l'échelle : un programme à l'origine séquentiel est parallélisé, d'abord sur un cluster de laboratoire (Tier 3), en développement, puis passe en production sur un mésocentre, avec en général 1000 cœurs pendant quelques heures, afin de tester la qualité du passage à l'échelle. Le programme est alors mûr pour être mis en œuvre sur un centre national, sur 10 000 cœurs.

Les mésocentres

Un mésocentre est un centre de calcul ou un data center de taille moyenne dans la pyramide des Tiers (Tier 2), bâti en général autour d'un cluster de calcul offrant quelques milliers de CPUs. Il n'y a pas de règle générale de fonctionnement d'un mésocentre. Selon la région, ils sont gérés par une Université, une unité de service d'un EPST (le Cnrs en général), un GIP. Ils comprennent généralement un comité scientifique et un comité d'utilisateurs, où les laboratoires utilisateurs sont représentés, les financements sont souvent régionaux. On assiste actuellement à plusieurs initiatives de rapprochement et harmonisation de ces différentes politiques, *via* un equipex (equip@meso) et un rapprochement avec la grille de calcul (France-Grille, voir http://www.france-grilles.fr/accueil/). Une présentation générale des mésocentres est accessible à http://calcul.math.cnrs.fr/spip.php?rubrique7.

Les projets soumis à un mésocentre sont en général de taille inférieure aux projets soumis à un centre national, et doivent également reposer sur des codes parallélisés. De plus en plus de mésocentres offrent des accès à des machines hétérogènes, en RAM (parfois 1 To de RAM sur un nœud, ce qui est utile pour les assemblages de génome), des fermes GPU, etc.

Les mésocentres sont souvent les lieux de mutualisation des compétences et d'animation scientifique régionale du calcul intensif, par des journées scientifiques annuelles. Tous les mésocentres proposent en général des formations. Elles sont très utiles au vu des pratiques (scheduleur, soumission de jobs) et des paradigmes utilisés (parallélisme MPI, OpenMP, distribution). L'accès aux machines pour les différents projets peut prendre plusieurs formes selon les mésocentres : FairShare (règles de gestion de l'accès aux ressources en calcul et files d'attentes, selon des critères définis par la communauté utilisatrice), dépôt de projets pour attribution d'heures...

Notons que le groupe calcul (http://calcul.math.cnrs.fr/) ouvert à l'ensemble de la communauté académique impliqué dans le calcul tient un site affichant les principaux événements concernant les mésocentres (journées annuelles des mésocentres, d'equip@meso…), des formations (http://calcul.math.cnrs.fr/spip.php?rubrique39), et gère une liste d'échange et de partage d'expérience sur le calcul (http://calcul.math.cnrs.fr/spip.php?article11).

La grille : une infrastructure distribuée

Une grille de calcul est une mise en réseau d'un très grand nombre de machines distinctes, distribuées sur un territoire vaste, connectées par un réseau à très haut débit sur de longues distances. Autant la notion de centre de calcul consiste à concentrer en un lieu unique une très grande puissance de calcul, autant une grille consiste à distribuer massivement les calculs sur des machines distinctes. Cette infrastructure est particulièrement adaptée lorsqu'un grand nombre de calculs indépendants doit être réalisé : par exemple, répéter plusieurs milliers de fois un même calcul statistique ou une même modélisation dans le cadre de procédures de bootstrap. Une différence essentielle entre les centres de calcul et la grille se situe au niveau du débit de communications entre machines (débit de type internet pour une grille).

Cette technologie a été mise au point au début des années 2000 pour le traitement des données massives en physique des hautes énergies. Elle est particulièrement adaptée aux traitements de données massives produites en continu par des capteurs, lorsque la tâche de traitement peut être organisée en paquets, de façon indépendante entre les paquets, chacun sur une partie des données (parallélisation par les données). Il est possible de raffiner cette approche par le paradigme dit map/reduce où, après avoir distribué les paquets entre machine (phase « map »), une synthèse des résultats par paquet doit être construite (phase « reduce »). Pour une présentation des grilles et l'historique de leur conception, consulter le site http://www.in2p3.fr/presentation/thematiques/grille/grille.htm .

Il existe une infrastructure européenne de grille (EGI : european grid infrastructure), voir https://www.egi.eu/, qui offre un accès à des sites de calcul et de stockage. Elle est administrée par une fondation de droit hollandais basée à Amsterdam. Cette infrastructure rassemble environ 350 sites, répartis dans 56 pays, interconnectés pour former une infrastructure de grille intégrée, sécurisée et fiable qui fournit à plus de 40 000 utilisateurs dans le monde entier les ressources indispensables pour relever les défis du traitement de données à haut débit. La grande majorité des logiciels déployés sur la grille sont des logiciels libres. Chaque pays membre est une NGI (national grid initiative), qui assure la gestion et l'accessibilité des machines et serveurs (stockage et calcul) sur le territoire national. Pour la France, il s'agit d'un GIS, dit France-Grille, dont le pilotage est confié au Cnrs, IN2P3 (85 % des calculs sur la grille le sont pour des projets liés à la physique des hautes énergies).

Une présentation de France-Grille ainsi que la présentation des procédures en permettant l'accès se trouvent à http://www.france-grilles.fr/accueil/. La procédure pour devenir utilisateur est clairement expliquée à http://www.france-grilles.fr/devenir-utilisateur/. Elle consiste principalement à envoyer un mail à info@france-grille.fr. Si le laboratoire concerné est sous tutelle de l'un des organismes du GIS France-Grille (voir la liste à http://www.france-grilles.fr/presentation/gis/), la procédure est un peu lourde, mais simple. L'essentiel consiste à recevoir une accréditation *via* des certificats. Rapidement, il faut d'abord agréer le laboratoire, puis un contact dans le laboratoire, qui sert de référent pour l'agrément des chercheurs/ingénieurs qui souhaitent travailler sur la grille. Trois types de service sont offerts :

- un accès à la grille de calcul, essentiellement *via* l'API Dirac (en python), voir http://www.france-grilles.fr/catalogue-de-services/fg-dirac/ ;
- un accès au cloud académique, en plein développement (voir http://www.france-grilles.fr/catalogue-de-services/fg-cloud/) ;
- un accès à un service de stockage, *via* le middleware iRODS (voir http://www.france-grilles.fr/catalogue-de-services/fg-irods/).

L'accès à la grille est organisé en V.O. (virtual organisation), qui consistent en des communautés scientifiques plus ou moins grandes qui décident de mettre en commun ou d'ouvrir à la V.O. les machines sur lesquelles, *in fine*, les travaux seront exécutés (même les machines virtuelles et le cloud tournent sur un support réel). Après accréditation sur la grille, il faut être accrédité par une V.O. Il existe une V.O. « transversale », qui offre les services de la grille pour toutes disciplines. Chaque V.O. offre un catalogue de services, qui ont un squelette commun, accessible à http://www.france-grilles.fr/catalogue-de-services/.

Conclusion

En fait, le calcul intensif est en pleine (r)évolution, suite au basculement de la demande de la modélisation et simulation numériques vers le traitement des données massives, issues notamment de capteurs (grands instruments, images, séquençage, phénotypage…) et avec les algorithmes et outils de l'Intelligence Artificielle et de l'apprentissage. De nouveaux paradigmes émergent, comme l'accès au cloud, le paradigme spark-hadoop (calcul distribué) pour les analyses complexes à grande échelle, ainsi que l'accès à des clusters de calcul *via* des bloc-notes comme jupyter pour les langages de calcul (Julia, Python, R). L'idée est de diversifier au mieux l'utilisation des ressources, et de choisir celles adaptées au mieux pour optimiser les traitements : de même qu'il n'existe pas de « langage » universel qui soit optimal pour toutes les tâches, il n'existe pas d'infrastructure de calcul qui soit optimale pour tout projet. Il faut adapter ensemble les éléments du triplet projet – code – infrastructure.

Concernant le calcul intensif, la démarche la plus naturelle et pragmatique est de :

- se former aux interfaces de programmation pour paralléliser les codes quand utile (MPI, OpenMP) ; il existe plusieurs lieux pour cela, soit auprès des centres nationaux soit des mésocentres ;
- construire un passage à l'échelle progressif, en vérifiant à chaque étape que la « scalabilité » est bien assurée : on peut utiliser MPI sur un laptop avec huit cœurs, ce qui peut être un premier pas, puis sur un cluster de laboratoire, puis sur un mésocentre, puis sur un centre national quand la taille du jeu de donnée l'exige ;
- si le projet est plus orienté vers un traitement de données massives issues de capteurs, qui peuvent être traitées par paquets, penser à aller sur la grille, et pour cela se former à l'intergiciel Dirac.

Les centres de calcul, qui évoluent vers des data centers, vu le rôle de plus important que prennent les questions de traitement des données, non seulement fournissent des services pour produire des calculs, mais offrent également un environnement de formation et d'accompagnement, avec une dimension locale pour les mésocentres. Il est important, voire primordial, de s'insérer dans ces communautés scientifiques utilisatrices du calcul, au-delà de sa propre thématique, pour bénéficier de toutes ces évolutions.