44ème forum de l'ORAP :
"Comment concilier pérennité et performances : quels défis pour l'Exascale ?"
Synthèse & Recommandations pour préparer la transition vers l'exascale
Comité Scientifique de l'ORAP
15/07/2020

L'exascale n'est pas seulement une augmentation de la puissance des machines de calcul, c'est une rupture technologique majeure et également une modification importante des usages du HPC. Les architectures de calcul les plus récentes utilisent des accélérateurs de type GPU ou autres, disposent de hiérarchies mémoires complexes et nécessitent déjà des adaptations majeures des codes applicatifs pour être utilisées correctement, voire leur réécriture complète. Par ailleurs, les travaux scientifiques qui seront effectués sur les machines exascale feront de plus en plus appel aux données et à l'enchaînement/couplage de différents codes de calcul (cf présentations du forum ORAP n°43). C'est donc une révolution qui concerne à la fois la technologie et les usages qu'il faut affronter.

Le contexte international

Les pays fortement impliqués dans le HPC ont mis en place des politiques ambitieuses pour tirer le meilleur parti des révolutions en cours. C'est notamment le cas du Japon, des Etats-Unis ou de la Suisse et également de la Chine où HPC et Big Data sont pris en compte dans une dynamique conjointe. L'approche Japonaise est partie des applications pour développer un processeur et un réseau d'interconnexions spécifiques (approche *co-design*). Ainsi la nouvelle machine Fugaku et son processeur ont été conçus pour répondre prioritairement à neuf questions scientifiques s'inscrivant dans des grands défis sociétaux. Les codes de calcul correspondant à ces défis ont également été co-développés durant la phase de conception. La solution actuelle ne dispose pas d'accélérateur, mais il est d'ores et déjà prévu, pour des raisons énergétiques, une solution hybride pour la machine suivante. Ainsi, entre le début de la phase de pré-design du processeur japonais, à base de technologie ARM, et la mise en production de la machine Fugaku, il se sera écoulé sept ans pendant lesquels concepteurs du matériel et développeurs des applications auront pu travailler en équipe. L'approche américaine s'appuie d'abord sur du matériel existant (sans négliger d'explorer de nouveaux développements de composants et d'architectures), en particulier les GPU, et l'effort est majoritairement mis sur le développement de solutions logicielles pour exploiter au mieux les machines (modèles de programmation innovants, intergiciels permettant d'abstraire les architectures matérielles, bibliothèques spécifiques, nouvelles fonctionnalités des langages de programmation,...). Le projet '*Exascale Computing Project*' est un projet à 1.7 Milliard de $ sur sept ans, lancé en 2016, qui implique les six centres du DOE, 100 Laboratoires, et 1000 chercheurs. Ce projet est structuré par des CoD (Co-Design Centers), en lien fort avec les applications. L'objectif est de développer, sur le long terme, des applications phares et une pile logicielle (middleware) permettant d'assurer, avec des coûts de développement raisonnables, performances et portabilités des applications. La solution suisse, qui ne dispose pas des mêmes budgets, est plus orientée application. Elle est basée sur le matériel et la pile logicielle bas niveaux existant, tandis que l'effort est mis sur la refonte des applications reposant sur

des techniques avancées de génie logiciel avec un focus sur les applications météo/climat et matériaux. Les budgets dédiés à cette refonte sont comparables à ceux investis pour les équipements. Cette approche plus locale met en évidence le besoin de compétences informatiques spécifiques dans le développement des codes de simulation.

Ces approches pour l'exascale sont assez différentes, mais ont toutes en commun de **mettre en œuvre une vision globale et de long terme allant du matériel aux applications**. Une telle vision est nécessaire afin de disposer d'applications exploitant efficacement l'infrastructure matérielle. Elle suppose des investissements matériels combinés avec une stratégie explicite de long terme et un effort comparable pour les développements applicatifs. Les ruptures technologiques en cours rendent l'usage des supercalculateurs extrêmement technique et délicat. Des spécialistes du HPC et de l'informatique du calcul haute performance sont nécessaires à toutes les étapes du développement des codes de calcul, notamment pour concevoir des applications capables de s'adapter dans la durée à des architectures de calcul en évolution rapide.

La portabilité des grands codes communautaires

Si les codes *prototypes* permettent de tester des solutions techniques et de valider leur *scalabilité*, la production scientifique des supercalculateurs se fait essentiellement avec de grands codes communautaires développés sur le long terme. Ces codes, qui représentent souvent des dizaines d'années de développement et sont le fruit d'un savoir-faire et d'une expertise de pointe, sont vitaux pour une entreprise ou une communauté scientifique mais ils ont néanmoins des freins intrinsèques à leur évolution : nombre de lignes de code de l'ordre ou supérieur au million, connaissance partiellement perdue au cours du temps, capacité de l'équipe à développer du code adapté aux nouvelles architectures, coût induit par la validation du code, … En raison de la diversité des solutions techniques et de leur complexité, l'effort de portage ne peut se faire que dans une vision de long terme basée sur une forte expertise technique. La portabilité des performances, à la fois sur les architectures actuelles mais également sur celles qui viendront, va donc devenir un défi plus important que l'obtention des performances brutes (de courte durée). Dans ce contexte d'évolution rapide des solutions techniques, de nombreuses communautés restent attentistes et ne disposent pas encore d'applications adaptées aux nouvelles architectures, principalement en raison du manque de ressources humaines spécialisées.

Les progrès récents en informatique du HPC et en algorithmique ainsi que le génie logiciel qui permet de structurer les codes, de les rendre plus facilement adaptables et évolutifs, sont des éléments centraux pour assurer cette portabilité de performance et la pérennité des applications. Une des voies possibles consiste à séparer autant que possible les aspects spécifiques à l'architecture (qui peuvent nécessiter des changements fréquents et de bas niveaux, demandant des compétences spécialisées) de la partie applicative, dont les spécialistes du domaine doivent pouvoir garder la maîtrise. Le développement et/ou l'utilisation de bibliothèques, de langages spécifiques et de nouvelles fonctionnalités dans les langages existants sont également des approches prometteuses. Les options offertes sont nombreuses dans un paysage en évolution rapide, il est donc nécessaire de disposer de forces suffisantes pour explorer et valider les solutions possibles.

Renforcer et allier les compétences

Autant à cause de la complexité des phénomènes simulés que de celles des nouvelles architectures de calculs, il est devenu indispensable que les équipes qui développent des codes allient ces différentes compétences. Il faut pour cela des spécialistes, formés aux dernières évolutions du calcul

haute performance (génie logiciel, modèles de programmation, algorithmes et méthodes mathématiques) au sein de centres dédiés ainsi qu'un réseau d'experts auprès des équipes de développement servant de relais et d'interface. Cela permettra à la fois d'atteindre une masse critique et de mutualiser des compétences de haut niveau et de diffuser cette expertise de manière efficace au sein des communautés. Au-delà de ces experts en HPC, les équipes applicatives ont également besoin de chercheurs, spécialistes des outils de simulation, afin que ces derniers soient pleinement exploités. Cela suppose une reconnaissance de la modélisation numérique, au même titre que les activités expérimentales ou théoriques, au sein des communautés scientifiques.

Cette approche a déjà montré des succès : en France, on peut citer le Cerfacs, la Maison de la Simulation et les équipes de support avancés des centres de calcul nationaux et régionaux. Toutefois ces efforts sont très largement sous-dimensionnés, notamment en comparaison de ce qui se fait à l'étranger (USA, Japon, Allemagne, Espagne, ...). Les exemples des SimLab à Jülich, des centres de calcul intensif de Barcelone (BSC) ou de Suisse à Lugano (CSCS) montrent de plus que ces équipes doivent être co-localisées avec les équipes de support des supercalculateurs, ou a minima travailler en interaction étroite. De ce point de vue, la France est le seul pays où les centres de calcul intensif académiques ne sont pas des centres de recherche. Les moyens humains doivent prendre une importance égale, voire supérieure au matériel : utiliser de manière optimale un supercalculateur coûtant des centaines de millions d'euros impose des investissements conséquents dans le capital humain. L'Europe a lancé un programme ambitieux dans le domaine du HPC et la communauté française doit se mettre en capacité d'en profiter pleinement. Une nouvelle organisation et le redéploiement de moyens est nécessaire et n'est possible que par une action volontariste des directions des organismes.

Recommandations de l'ORAP :

- Mettre en place une politique ambitieuse et de long terme de recrutement de spécialistes du HPC, chercheurs et ingénieurs, sur des postes permanents afin de rattraper le retard accumulé ; comme indiqué par la Cour des Comptes dans son dernier rapport annuel, le déploiement de l'infrastructure doit s'accompagner d'une réflexion sur les ressources humaines associées.
- Créer et renforcer des centres de compétences pour mutualiser une partie de ces recrutements à travers les domaines scientifiques et pour faire émerger une communauté des sciences du calcul; renforcer les équipes scientifiques autours des codes applicatifs afin que ces centres soient les nœuds d'un réseau d'experts.
- Un effort particulier est nécessaire dans les années à venir afin de faire face à l'arrivée des premières machines exascale ; des efforts très importants sont envisagés sur le hardware, dans la continuité de ce qui s'est fait avec la création du GENCI, et une action d'envergure cohérente est également indispensable pour ce qui concerne le personnel, domaine dans lequel la France a pris un retard particulier.
- Mettre en place une stratégie de long terme pour l'achat des équipements en associant les utilisateurs et les concepteurs d'applications dans un esprit de *co-design*.
- Veiller à intégrer les possibilités offertes par les développements numériques à l'exascale dans la stratégie scientifique qui sera mise en place pour répondre aux challenges sociétaux, notamment pour prendre en compte l'accroissement des risques climatiques et les besoins dans les domaines biologie et santé. Ceci doit se faire de manière holistique permettant une large implication des domaines scientifiques en amont, ainsi que le développement de modèles numériques et des grands codes requis.
- Promouvoir la reconnaissance et l'essor des activités de simulation numérique au sein des communautés scientifiques, notamment pour les recrutements et les promotions.