Relever les défis de la variabilité des Entrées/Sorties dans les simulations HPC post-petaflopiques

Matthieu Dorier [1]

Accessit du prix de thèse Gilles Kahn 2015

Matthieu Dorier a soutenu sa thèse [2] en décembre 2014 à l'École normale supérieure de Rennes, sous la direction de Gabriel Antoniu et Luc Bougé, au sein de l'équipe Inria/IRISA KerData. Il effectue actuellement un stage post-doctoral à Argonne National Laboratory (Illinois, États-Unis).

En 2008, la communauté du calcul hautes performances (HPC) atteignait le *Petascale* avec Roadrunner d'IBM, un supercalculateur de 122400 cœurs atteignant une performance de 1.375 Petaflops (1.375×10^{15} *floating point operations per second*). La barre du million de cœurs a été atteinte en 2012 avec le supercalculateur Sequoia à LLNL, et l'ère de l'*Exascale* (10^{18} flops) est attendue pour 2020. Une telle puissance de calcul est mise à profit dans de nombreux domaines de recherche tels que les sciences de la Terre, la biologie, le climat ou l'astrophysique, domaines dans lesquels les simulations à large échelle sont employées pour mieux comprendre les phénomènes physiques qui nous entourent. Ces simulations ont vocation à remplacer

1. http://people.irisa.fr/Matthieu.Dorier
2. Consultable à l'adresse https://tel.archives-ouvertes.fr/tel-01099105v1

des expériences réelles qui peuvent être trop coûteuses, trop dangereuses ou simplement irréalisables, comme les études portant sur la jeunesse de l'univers.

Les entreprises utilisent également les supercalculateurs pour diminuer leurs coûts de conception. Les simulations hautes performances ont en effet l'avantage d'être plus rapides et moins chères que la conception et les tests de prototypes réels. De plus, ces simulations peuvent être reproduites et les modèles virtuels peuvent être évalués dans des conditions variées avec une très grande précision.

Mais comme le dit Ken Batcher, « *a supercomputer is a device for turning compute-bound problems into I/O-bound problems*[3] ». En effet, de plus grosses machines mènent à une production accrue de données. Ces données doivent être stockées et traitées efficacement en vue d'en tirer un résultat scientifique. L'approche traditionnelle de gestion de données consiste à stocker les données produites par la simulation dans des fichiers pendant que celle-ci s'exécute, et à analyser ces fichiers plus tard, lorsque la simulation est terminée. On observe cependant un fossé de plus en plus large entre les performances des systèmes de stockage et les performances des systèmes de calcul dans les supercalculateurs récents. Par exemple, alors que le supercalculateur Jaguar d'ORNL (premier du Top 500 en novembre 2009 et Juin 2010) fournit un débit de 240 Go/s vers son système de stockage, pour une performance de calculs de 1,75 Petaflops, son successeur Titan (premier au Top 500 en novembre 2012) fournit un débit de stockage seulement six fois supérieur (1,4 To/s) pour une puissance de calculs dix fois supérieure (17,59 Petaflops). Ce fossé rend obsolètes les approches traditionnelles pour les entrées-sorties (E/S), qui prennent alors une part grandissante du temps d'exécution des applications et sont sujettes à une *variabilité* croissante de leurs performances.

D'une part, il devient donc nécessaire d'optimiser la pile logicielle des E/S à tous les niveaux, de la simulation jusqu'au système de fichiers, dans le but d'en améliorer les performances ainsi que la predictibilité de ces performances. Cela implique également d'améliorer la manière de gérer une concurrence croissante au niveau du système de fichiers, non seulement entre des centaines de milliers de processus constituant une seule application, mais également entre un nombre croissant d'applications qui s'exécutent sur la même machine et en partagent le système de stockage.

D'autre part, il devient inévitable de rapprocher les tâches d'analyse et de visualisation de la simulation elle-même afin d'éviter de stocker de larges quantités de données. Cette tendance soulève de nouveaux défis liés aux moyens dont disposent les simulations pour communiquer efficacement leurs données et partager ces dernières avec les outils d'analyse sans dégrader leurs performances.

3. Traduction : « un supercalculateur est un appareil transformant un problème limité par les performances de calcul en un problème limité par les performances des entrées/sorties. »

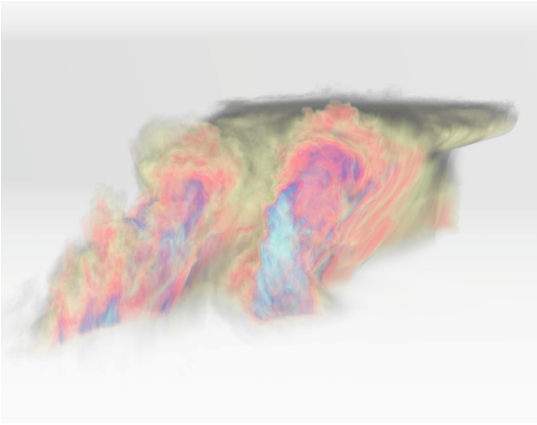

FIGURE 1. Visualisation in situ de la simulation atmosphérique CM1.

Enfin, la consommation énergétique des futurs supercalculateurs est un problème de plus en plus important dans la communauté HPC. Alors que les machines actuelles consomment une puissance d'environs 10 MW, la *Defense Advanced Research Projects Agency* (DARPA) a imposé une limite de 20 MW pour les futures machines Exascale. Cela représente une multiplication par deux de la consommation d'énergie pour des plateformes qui devront être mille fois plus performantes en termes de calculs. Cette contrainte ne pourra être satisfaite seulement par des améliorations matérielles, mais nécessitera par la conception d'approches logicielles plus économes en énergie. Les mouvements et le stockage de grandes masses de données constituent notamment des tâches coûteuses en énergie et doivent être optimisées en conséquence.

Dans cette thèse, nous avons tout d'abord proposé et implementé une nouvelle méthode de gestion des E/S, appelée Damaris [4] [1] qui se sert de cœurs dédiés aux E/S sur chaque nœud multicœur, ainsi que de mémoire partagée. Damaris permet de réaliser les tâches de traitement de données et d'E/S de manière asynchrone, et de cacher la variabilité de ces dernières en conséquence. Nous avons évalué Damaris sur trois plateformes différentes, notamment le supercalculateur Kraken (11^{e} du Top 500 au moment des expériences) avec la simulation atmosphérique CM1 et la simulation de dynamique des fluides Nek5000. En permettant le recouvrement des E/S et des calculs, et en regroupant les données dans des fichiers plus volumineux tout en évitant les synchronisations entre cœurs, notre solution offre d'importants gains de performance. Ainsi à une échelle de 9000 cœurs sur Kraken, Damaris a permis de diviser par 3.5 le temps d'exécution de CM1, et de multiplier par 15 le débit des

4. `http://damaris.gforge.inria.fr`

E/S, tout en permettant une compression asynchrone des données. Damaris a également été utilisé pour coupler les simulations avec des logiciels de visualisation parallèle [2], permettant ainsi une visualisation *in situ* des données(voir Figure 1), basée sur un envoi direct des données de la simulation vers le logiciel de visualisation sans passer par des fichiers. Enfin, nous avons montré que l'utilisation de Damaris permet aussi un gain au niveau de la consommation énergétique des supercalculateurs [5].

Dans un deuxième temps, nous avons également proposé de nouvelles approches pour résoudre le problème de variabilité des E/S induit par l'exécution de plusieurs applications indépendantes [3], et pour modéliser et prédire les futurs accès au système de fichiers effectués par les applications afin d'en optimiser les E/S [4].

Ces travaux ont été réalisés dans le contexte de collaborations avec l'université d'Illinois à Urbana-Champaign et *Argonne National Laboratory*, notamment dans le cadre du *Joint Laboratory for Extreme-Scale Computing* [5].

Références

[1] Matthieu Dorier, Gabriel Antoniu, Franck Cappello, Marc Snir, Leigh Orf. *Damaris : How to Efficiently Leverage Multicore Parallelism to Achieve Scalable, Jitter-free I/O*, Proceedings of the 2012 IEEE International Conference on Cluster Computing (CLUSTER '12).

[2] Matthieu Dorier, Roberto Sisneros, Tom Peterka, Gabriel Antoniu, Dave Semeraro. *Damaris/Viz, a Nonintrusive, Adaptable and User-Friendly In Situ Visualization Framework*, Proceedings of the 2013 IEEE Symposium on Large Data Analysis and Visualization (LDAV '13).

[3] Matthieu Dorier, Gabriel Antoniu, Rob Ross, Dries Kimpe, Shadi Ibrahim. *CALCioM : Mitigating I/O Interference in HPC Systems through Cross-Application Coordination*, Proceedings of the 2014 IEEE International Parallel & Distributed Processing Symposium (IPDPS '14).

[4] Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu, Rob Ross. *Omnisc'IO : A Grammar-Based Approach to Spatial and Temporal I/O Patterns Prediction*, Proceedings of ACM/IEEE 2014 Supercomputing Conference (SC '14).

[5] Orçun Yildiz, Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu. *A Performance and Energy Analysis of I/O Management Approaches for Exascale Systems*, in Proceedings of the 2014 Data-Intensive Distributed Computing (DIDC '14) workshop.

5. http://publish.illinois.edu/jointlab-esc/