SEPTEMBRE 2013

HighPerformanceComputing

Le magazine référence du calcul intensif et du Big Data

Volume I | Numéro 6

France

0x06

/DECOUVRIR

RIP, ROADRUNNER

Eloge et autopsie du premier calculateur pétaflopique

/L'ETAT_DE_L'ART

MONTE CARLO QUANTIQUE

Parallélisme massif et méthodes probabilistes en chimie

/DEVELOPPER

OPTIMISEZ VOS APPLICATIONS MPI

Après les processus et les benchmarks, les connexions InfiniBand !

/L'ATELIER_CUDA

COPPERHEAD, TOUT CUDA EN PYTHON

La productivité de l'un, la puissance et la versatilité de l'autre

SIMULATION NUMERIQUE : QUATRE CAS D'USAGE EXEMPLAIRES

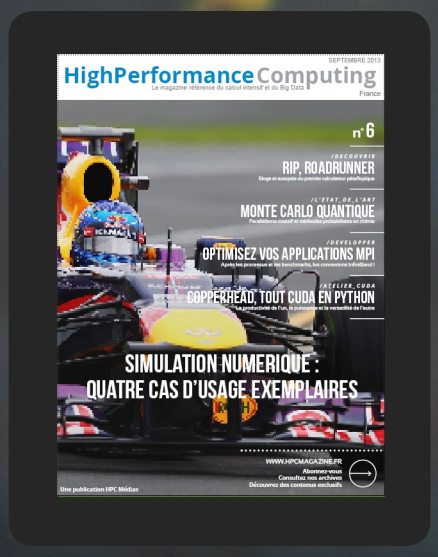
WWW.HPCMAGAZINE.FR

Abonnez-vous
Consultez nos archives
Découvrez des contenus exclusifs

Une publication HPC Médias

EDITO

SEPTEMBRE 2013



Format Desktop / Tablettes

"Technical computing"

Je ne sais pas pour vous, mais ici, à la rédaction, les vacances ont été plutôt courtes. Entre la préparation de nos éditions Europe et Americas (toutes deux en anglais), de gros travaux techniques sur nos sites Web, la formation de nouveaux collaborateurs et la veille technologique que nous vous devons, l'été à été chaud... mais à l'intérieur ! Loin de nous en plaindre, nous sommes plus qu'impatients de vous faire partager, au delà des livraisons mensuelles de votre magazine préféré, le fruit de ces efforts volontiers consentis.

En attendant - pas trop longtemps - voici un numéro qui fait la part belle à la simulation ou, comme disent désormais les fournisseurs de solutions HPC d'entreprise, au "technical computing". On croirait presque que le terme "HPC", historiquement connoté académique voire recherche, fait peur au client...

Allez, laissons au marketing ses pudeurs et regardons ensemble les choses évoluer. L'ouverture de plus en plus large du calcul intensif au secteur privé sera probablement la contribution la plus décisive aux prochains passages à l'échelle. Intel, IBM, Google... les voilà qui prennent le relais de nos institutions pour faire avancer la machine à un rythme qui leur convient mieux. Faut-il le regretter ? Le résultat, jusqu'à présent, c'est un formidable enrichissement du trésor de nos connaissances disponibles en Open source...

Bonne lecture !

Frédéric MILLIOT
Directeur de la rédaction
frederic@hpcmagazine.com

CARRI Systems

Architecte technologique depuis 1992

intel

Découvrez notre gamme **XLR dédiée aux accélérateurs sur ports PCI Express**

Les nouveaux serveurs de calcul XLR de CARRI Systems sont désormais compatibles avec les coprocesseurs Intel® Xeon® Phi™ 5110P.

HighServer XLR4i

Appuyez vos innovations de pointe avec le traitement hautement parallèle du coprocesseur Intel® Xeon Phi™. Chaque puce peut traiter plus d'un téraFLOP en performances de pointe à double précision, les meilleures performances de traitement parallèle par watt de tous les processeurs Intel® Xeon®

intel inside™ XEON PHI™

Coprocesseur Intel® Xeon® Phi™

HighServer XLR4i
133548

- Intel® Xeon® E5-2620 à 2 Ghz
- 32 Go DDR3 1600 MHz ECC REG
- 2 x 1 To RAID Edition (7 200 tr/min, 64 Mo de cache)
- 1 x Intel® Xeon® Phi™ 5110P avec 8 Go GDDR5 @ 5.0 GT/s
- Chipset vidéo intégré
- 2 x Gigabit Ethernet
- Rack 2U
- 2 x alimentations redondantes 1620 W
- Linux Centos®
- Garantie 3 ans sur site pièces et main d'oeuvre (hors accessoires)
- France Métropolitaine uniquement (hors îles)

À partir de 5 990 euros HT

Découvrez nos solutions XLR sur

www.carri.com

intel Technology Provider Platinum 2013

intel inside™ XEON®

intel inside™ XEON PHI™

intel inside™ SOLID-STATE DRIVE

Ultrabook, Celeron, Celeron Inside, Core Inside, Intel, Logo Intel, Intel Atom, Intel Atom Inside, Intel Core, Intel Inside, Logo Intel Inside, Intel vPro, Itanium, Itanium Inside, Pentium, Pentium Inside, vPro Inside, Xeon, Xeon Phi, et Xeon Inside sont des marques de commerce d'Intel Corporation aux Etats-Unis et dans d'autres pays.
Intel, le Logo Intel, le Logo Intel Inside, Xeon et Xeon Phi sont des marques de commerce d'Intel Corporation aux États-Unis et dans d'autres pays.

AU SOMMAIRE

CE MOIS-CI

/actus
06 L'essentiel
IBM fédère autour de POWER, NVIDIA avale PGI, des Xeon avec fonctions réseau intégrées, le projet DataScale, les Méso-challenges et + encore...

/en_couverture
22 Simulation : quatre *success-stories*
Des témoignages qui racontent en détails ce que la simulation numérique apporte de bénéfices concrets aux équipes qui l'adoptent...

/verbatim
38 Marc Simon - SGI
Retour sur la conception technique de Pangea - Très haute densité et efficacité énergétique - La vraie valeur ajoutée d'Hadoop...

/l'état_de_l'art
46 Monte Carlo quantique en chimie
Une implémentation efficace de l'approche QMC pour l'exploitation massive du parallélisme intrinsèque des méthodes probabilistes en chimie.

/découvrir
54 Fin de course pour Roadrunner
Après cinq ans de service actif, le premier calculateur pétaflopique tire sa révérence - non sans un dernier témoignage de l'au-delà.

/développer
58 Optimisez vos applications MPI
La performance des codes MPI dépend de plus en plus des inter-connexions. Revue de détail et contribution au débogage de MVAPICH2.

/atelier_CUDA
64 Copperhead : Python parallélisé
Allier la productivité de Python et l'efficacité de la programmation parallèle sur GPU, telle est sa vocation. Une prise en mains s'impose !

© HPC Médias 2013.

Ont collaboré à ce numéro :

Michel Caffarel
Brian Catanzaro
Amaury de Cizancourt
Frédéric Milliot
Alex Roussel
Anthony Scemama
Jérôme Vienne
et...
Franz Schubert

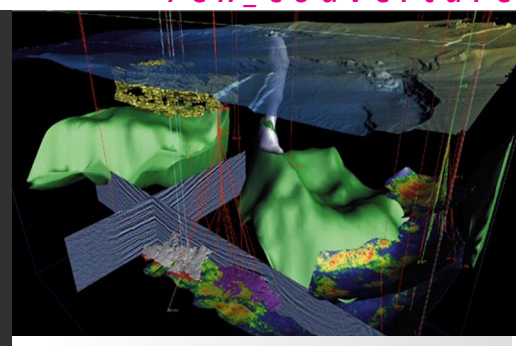
+33 9 50 96 64 00

REDACTION
Suggestions, commentaires, propositions d'articles...
laredaction@hpcmagazine.fr

ABONNEMENTS
Un problème avec votre abonnement ?
abonnements@hpcmagazine.fr

PUBLICITE
Annoncez dans le seul média intégré dédié à l'univers HPC.
pub@hpcmagazine.fr

/en_couverture

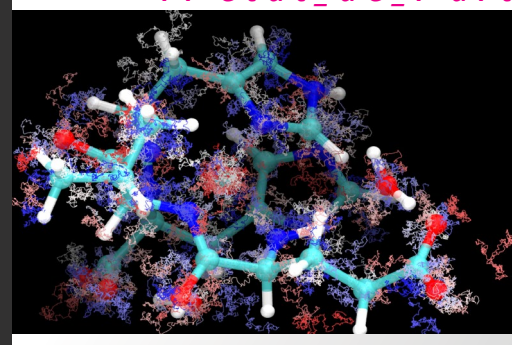


Sismique
La simulation pour voir sous plusieurs couches d'océan, de sels et de roches
Page 29

/verbatim

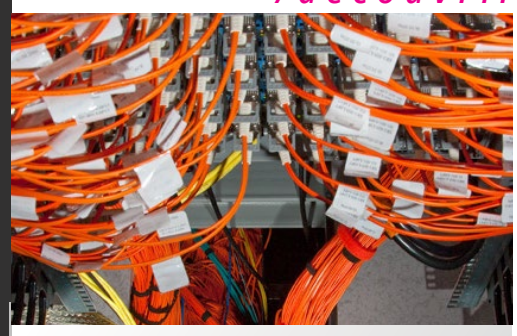
Privé
Les coulisses de l'appel d'offre Pangea, le calculateur privé le plus puissant au monde
Page 40

/l'état_de_l'art



Probabiliste
Une approche radicalement nouvelle de la simulation numérique en chimie
Page 46

/découvrir



Décommissionné
Roadrunner est aujourd'hui débranché. Son testament est riche d'enseignements
Page 54

HighPerformanceComputing
HPC MEDIAS
11, rue du Mont Valérien
92210 St-Cloud - France

PNY®

SERVEURS HPC de calcul par GPU

Accélérez les procédures de calcul scientifique et technique avec les GPU NVIDIA Tesla.
Les serveurs TYAN et ASUS de PNY, associés aux GPUs Tesla de dernière génération, constituent l'architecture la plus rapide et la plus efficace au monde.

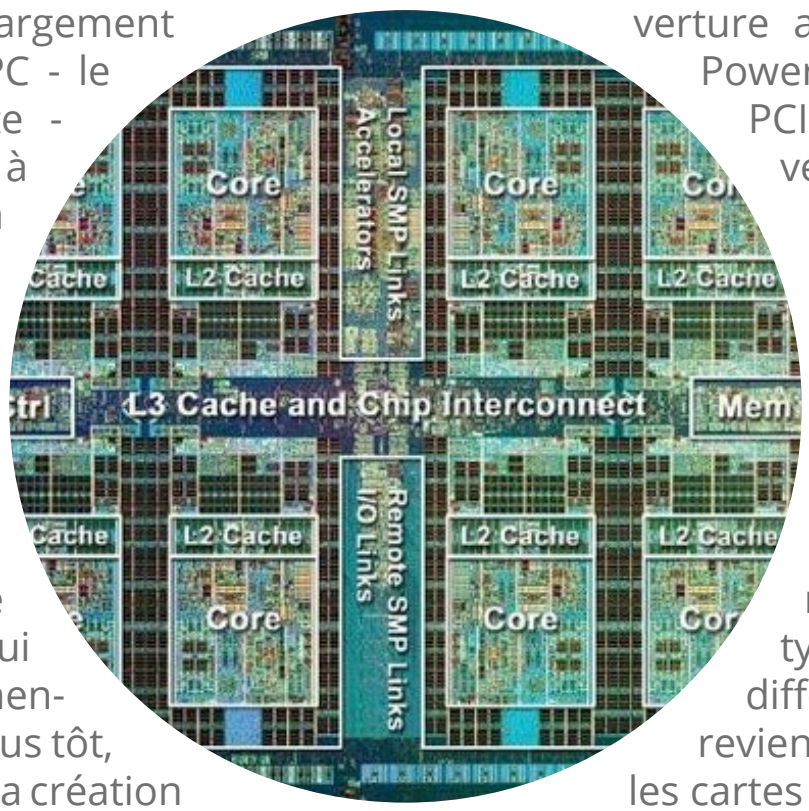
Plus d'informations sur produits, visitez www.pny.eu/tesla et le configurateur PNY
http://pny.quadrok-selector.com/

@PNYProFR

ASUS® TYAN

/ACTUS

IBM fait consortium autour de POWER



L'architecture x86 s'est largement imposée sur le marché HPC - le dernier TOP500 en atteste - mais IBM est bien décidé à défendre ses positions. Si la présentation officielle du Power8 dodécacores (qualificatif ajouté à notre dictionnaire métier à l'occasion de cet article) a bien eu lieu lors de la conférence Hot Chip qui s'est tenue à Stanford fin août, ce n'est pourtant pas le sujet qui y a suscité le plus de commentaire. Quelques semaines plus tôt, en plein été, IBM annonçait la création du consortium OpenPOWER, qui regroupe notamment Google, Mellanox, NVIDIA et Tyan. A la clé, l'ouverture des brevets aux nouveaux partenaires, afin qu'ils développent des serveurs, des équipements réseaux et de stockage, et des accélérateurs GPU optimisés pour les systèmes Big Blue.

Avec une activité en baisse de 25 % au deuxième trimestre, IBM a fini par admettre qu'il ne pouvait plus défendre seul son architecture, qu'il lui fallait constituer autour d'elle un écosystème qui garantisse sa pérennité. Sans surprise, IBM et NVIDIA vont travailler de concert au portage de CUDA vers des architectures GPU sur POWER. De même, Mellanox, déjà partenaire OEM d'IBM, va pouvoir optimiser ses cartes InfiniBand pour l'offre POWER. Une démarche simplifiée par l'ouverture aux standards du marché. Power8 se rallie en effet au bus PCIe 3.0, tandis qu'une nouvelle interface CAPI (Coherence Attach Processor Interface) permettra de faire communiquer Power8, GPU, FPGAs et autres ASICs au sein même des calculateurs. Dans l'annonce, IBM précise que le consortium devrait donner naissance à de nouveaux types de serveurs destinés à différentes charges de calcul. Il reviendra à Tyan d'en concevoir les cartes mères.

Reste une interrogation : quel rôle va jouer Google dans l'affaire ? Le géant du Web cherche-t-il à s'affranchir d'Intel, au moins en partie, pour ses serveurs conçus principalement en interne ? Un porte-parole a déclaré au Wall Street Journal que le consortium avait le potentiel de faire de l'architecture POWER une option viable pour les applications exécutées dans ses datacenters... Déclaration pour le moins ambiguë, qui n'engage à rien de ferme mais qui pèsera son poids au moment de (re)négocier avec Intel ou AMD. Le porte-parole Google a souligné l'effort d'ouverture de la part d'IBM et le vent nouveau d'innovation qui souffle désormais sur POWER. Ultime sursaut ou renaissance d'une design performant mais jusqu'ici verrouillé ? Les mois à venir diront si le marché est réellement demandeur... **AR**

A nos lecteurs : nos sites Web vont connaître une mise à jour technique et graphique majeure pendant la seconde moitié de septembre. C'est pour la bonne cause mais ce genre d'exercice n'allant jamais sans quelques petits désagréments d'usage, nous vous prions à l'avance de nous en excuser.

COMPILATEURS HPC : NVIDIA ENTRE DANS LA DANSE

C'est officiel depuis fin juillet : The Portland Group Inc. (PGI), un des deux éditeurs indépendants de compilateurs dédiés HPC (avec CAPS), appartient désormais à NVIDIA. En pleine restructuration, STMicroelectronics se sépare de cet actif stratégique devenu pour lui accessoire. En 2000, "ST Micro" avait pris le contrôle de PGI pour 17 M$. Le montant de la transaction avec NVIDIA n'a pour sa part pas été communiqué.

Créé en 1989, PGI commence à travailler sur le processeur Intel i860, pour lequel il avait développé des compilateurs Fortran et C. Cette incursion d'Intel sur le marché RISC fut un échec. Qu'à cela ne tienne, les équipes de PGI participent alors à la mise au point de HPF (High Performance Fortran) et vont connaître leur heure de gloire en développant toute une série de compilateurs C/C++ et Fortran pour le premier calculateur téraflopique, le vénérable ASCI Red du laboratoire américain Sandia. L'arrivée de STMicroelectronics ne bouleverse pas l'autonomie de l'éditeur, qui continue de décliner ses compilateurs sur diverses plateformes dont Linux, Windows et MacOS.

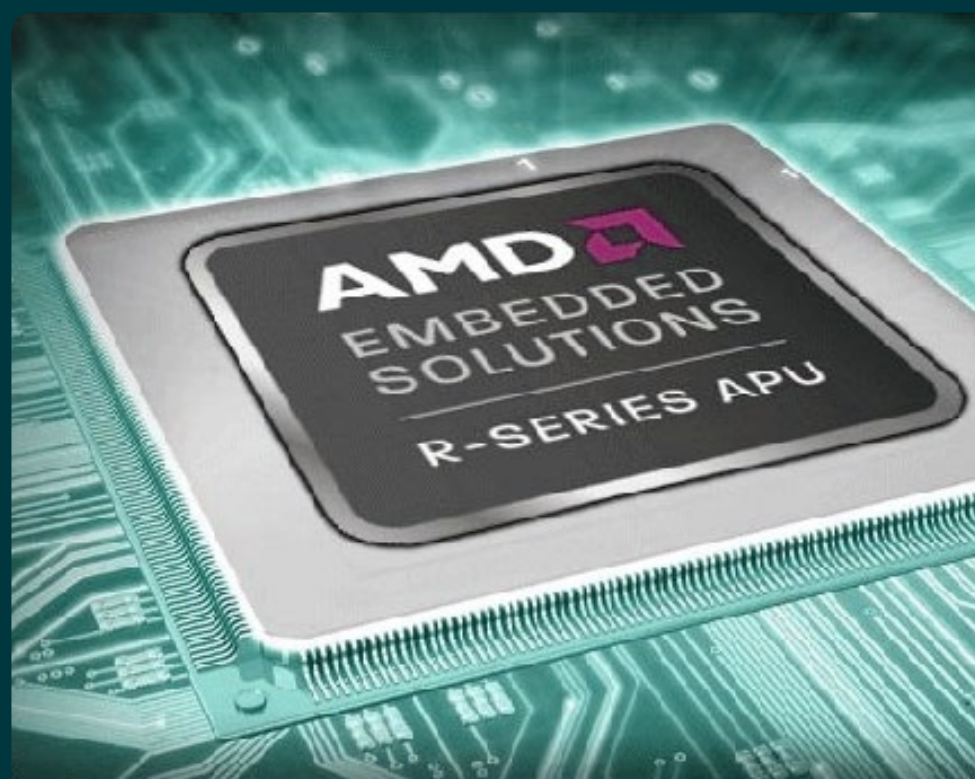
Plus récemment, PGI s'est beaucoup impliqué dans le développement de compilateurs pour les architectures GPU et ARM, et a joué un rôle déterminant dans l'élaboration d'OpenACC. C'est chemin faisant que les liens se sont tissés avec NVIDIA, avec en particulier le développement de compilateurs dédiés à CUDA.

L'acquisition de PGI va permettre à NVIDIA de muscler ses équipes de développement et de rivaliser avec Intel dans le domaine stratégique des outils de programmation. Comme le déclarait Ian Buck, responsable de l'engineering chez NVIDIA, "*derrière toute grand architecture processeur, il y a une grande équipe compilateur. La nôtre devient plus grande. Et meilleure.*"

A priori, la marque PGI devrait perdurer, et le travail continuer sur les développements CPU x86 et GPU. Mais on peut s'interroger sur le devenir du support AMD Radeon et Xeon Phi, qui figurait encore dans les roadmaps avant l'été. Premier effet tangible d'un changement de cap chez PGI, plusieurs pages relative à PGCL, le compilateur OpenCL pour ARM, ont disparu de son site Web. Ce compilateur permettait notamment de générer des applications Android. Une mise à jour de PGCL, attendue le 21 août, n'a pas été publiée. Signe que l'équipe ARM de PGI s'est déjà mise au travail sur la puce ARM 64 bits "Denver" ?

AMD étend son offre embedded

L'architecture x86 a elle aussi sa carte à jouer dans les applications embarquées. En témoigne la nouvelle gamme Embedded R-Series d'AMD, qui compte des CPU bicores (RE264X, RE272X) et quadricores (RE464X) cadencés entre 2,2 et 3,7 GHz pour un TDP compris en 17 et 35 Watts. Outre un certain nombre d'applications domestiques, le fondeur vise l'intégration "haute performance" tel que l'acquisition de signal ou la robotique orientée calcul. A noter, ces annonces s'accompagnent d'une remise importante sur l'achat conjoint de processeurs graphiques Radeon E6460 et E6760.

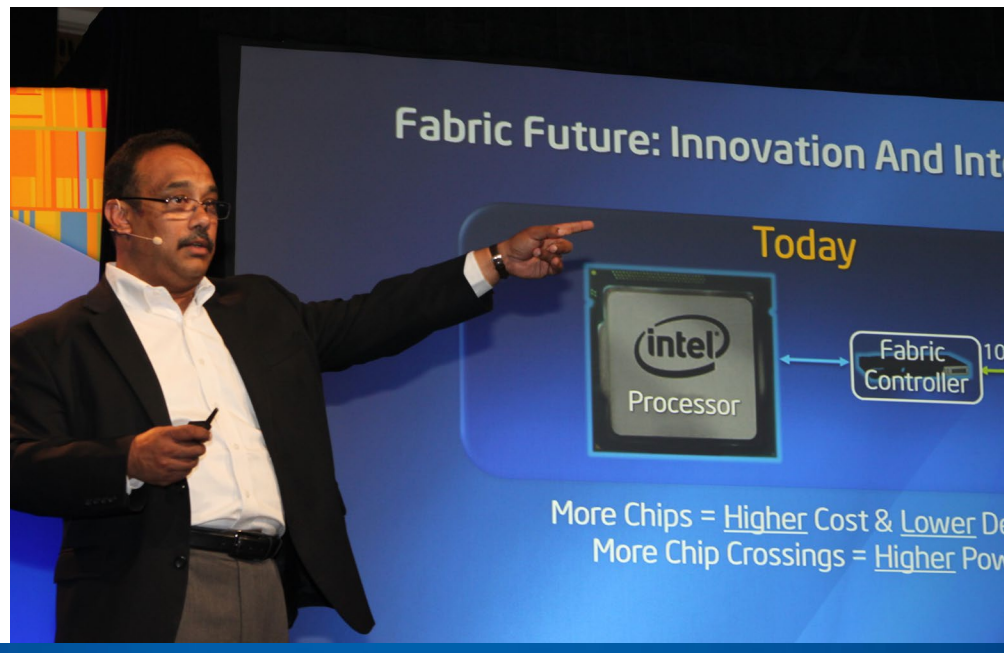


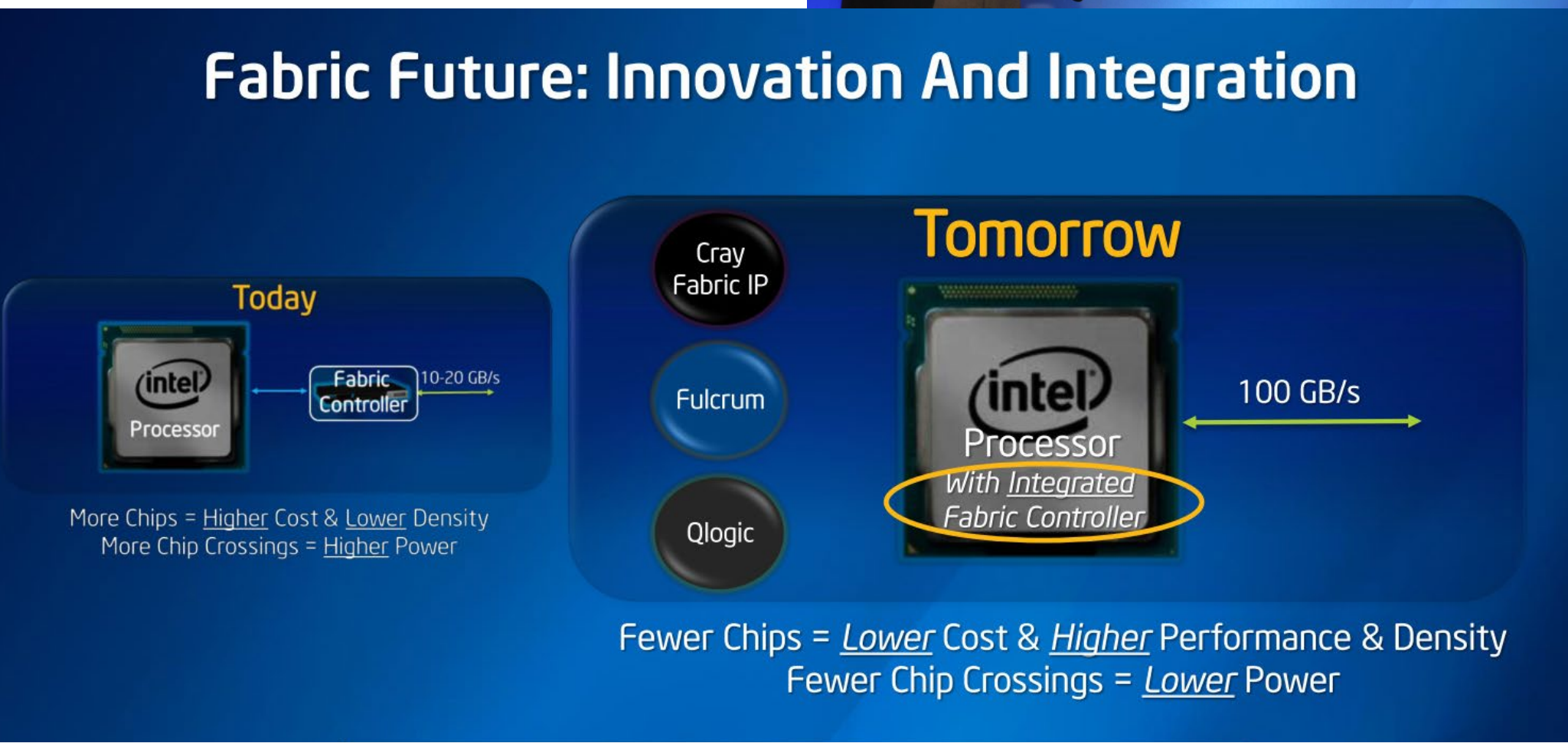
/actus

Intel Datacenter Days : l'intégration au CPU, condition nécessaire pour l'accès à l'exaflops

"Réimaginer le datacenter" : tel était le mot-d'ordre officiel de l'Intel Datacenter Day, qui s'est tenu le 22 juillet dernier à San Francisco. Transverse comme il doit, le rendez-vous réunissait notamment les experts Intel en matière de HPC et de Big Data. Et compte tenu de l'influence grandissante de l'américain dans ces deux domaines - qui selon nous n'en font qu'un - l'audience n'a pas manqué...

Pour entendre notamment Rajeeb Hazra, VP en charge du Technical Computing, dresser un point d'étape sur la stratégie HPC du fondeur.





"*Plus que jamais*", assure-t-il, "*nous sommes engagés sur la voie de l'exascale*". Et de poursuivre en affirmant que "*les premières machines de cette classe de puissance devraient entrer en production d'ici la fin de la décennie.*" Jusque-là, on n'est pas surpris, même si la succession de confirmations officielles concernant cette échéance a de quoi rassurer sur l'état des progrès technologiques qui impactent notre communauté. Mais le plus intéressant était sans doute les quelques éclaircissements concernant l'approche pour y parvenir...

> Intégrer de nouvelles capacités à Xeon

"*Plus vous avez de puces, plus les coûts sont élevés et plus basse est la densité*" a répété Rajeeb Hazra pour expliquer sa stratégie de développement produit. "*Regardez les grands calculateurs du Top10 : le réseau consomme entre 20 à 30 % de la puissance électrique du système complet. Cet obstacle majeur, nous avons choisi de le franchir en intégrant les fonctions réseaux directement dans les Xeon de prochaines générations.*" Voilà qui explique rétrospectivement les

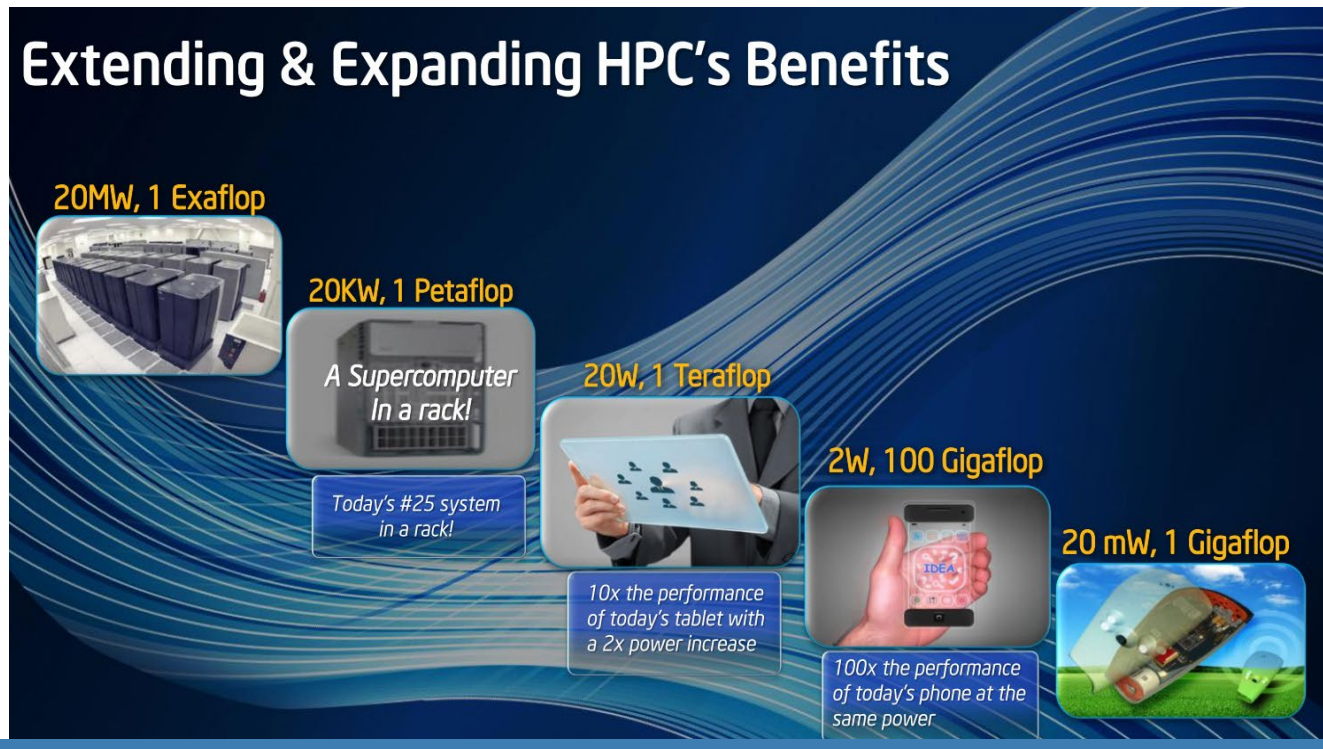
acquisitions de Fulcrum et surtout des activités InfiniBand de Qlogic et des brevets Cray relatifs à l'interconnexion en 2012. C'est toujours en vertu de la loi de Moore qu'Intel va pouvoir intégrer les contrôleurs réseau, le routage et en partie la gestion du stockage directement au CPU. Diminuer le nombre de composants améliore évidemment l'efficacité énergétique, mais pas seulement. On peut tabler sur des performances nettement revues à la hausse, avec des vitesses d'échanges passant nativement à 100 Gbps contre 10 à 20 « seulement » aujourd'hui.

Cette approche SoC transposée aux environnements serveurs semble d'ailleurs irréversible. Comme d'autres grands noms du silicium avant lui, Rajeeb Hazra a évoqué l'intégration de la mémoire sur le processeur (Stack Memory). En revanche, aucune précision "publique" sur les spécifications d'un tel Xeon 3D ni sur sa date de disponibilité.

> Toujours plus de softs Intel

Autre axe de développement suivi par Intel, le logiciel. La marque a en effet largement étoffé son catalogue ces derniers trimestres, avec des initiatives qui ont parfois surpris. C'est pourtant "*la condition* sine qua non *à une démocratisation du HPC en entreprise*" selon Rajeeb Hazra. "*Nous avons étendu nos outils bien au-delà des simples librairies et compilateurs, pour aller vers les outils de niveau cluster. C'est d'une grande importance pour le TCO des calculateurs, mais aussi pour leur efficacité.*"

Ce portfolio d'actifs softs, Intel continue de l'enrichir, avec notamment une distribution Hadoop annoncée en février et, plus récemment, une version maison de Lustre. Ces solutions commencent d'ailleurs à donner naissance à un écosystème rentable, au sein duquel on trouve la plupart des grands noms du HPC traditionnel, SGI et Cray en tête. C'est là sans doute l'enseignement le plus important de cette belle journée d'été. **AR**



DU HPC DANS TOUS LES TERMINAUX

"*Atteindre l'exaflops dans une puissance électrique limitée à 20 MW, toute la communauté est aujourd'hui en train de travailler pour y parvenir à la fin de cette décennie*" confirme Rajeeb Hazra. Il estime également que, d'ici 2021-2022, de deux à cinq systèmes exa seront déployés dans le monde. Par contrecoup, les technologies développées pour cela devraient ressortir bien au-delà de la communauté HPC, et impacter l'industrie IT dans sa globalité (voir l'illustration ci-dessus). "*Nous allons pouvoir offrir un Pétaflops pour une puissance électrique de 20 kW, et ce dans un simple rack. C'est un élément de démocratisation absolument extraordinaire pour le calcul intensif...*" ajoute Mr Hazra. La déclinaison des puissances disponibles sur différents types de terminaux laisse songeur : un Téraflops sur une tablette, des smartphones d'une puissance multipliée par 100... même les objets intelligents pourraient à terme bénéficier de ce surcroît de puissance à faible coût énergétique. L'unité de base, le Gigaflops, Intel pense pouvoir d'ici là le fournir pour seulement 20 mW consommés. "Cool", serait-on tenté de dire à plus d'un titre....

/actus

PNY sur tous les fronts

Grosse rentrée en perspective pour l'OEM n° 1 de NVIDIA, dont l'offre *compute* et visualisation s'enrichit sensiblement. Avec d'abord la nouvelle carte Quadro K6000, disponible physiquement à l'heure où vous lirez ces lignes et en exclusivité PNY pour au moins deux mois. Au menu, tout simplement le GPU le plus efficace jamais fondu à ce jour (448 cœurs, 1030 / 515 Gflops sp/dp), et pas moins de 12 Go de GDDR5 à 144 Go/s. Avec une résolution mono-écran maximale de 2560 x 1600 pixels, K6000 affiche 1,3 milliards de triangles par seconde, notamment en OpenGL 4.3. Un produit très technique, donc, qui s'adresse en premier lieu aux contextes académiques, industriels, créatifs et médicaux dans lesquels la visualisation multiécrans accélérée est indispensable à l'utilisation des applications.

Une nouvelle gamme de serveurs HPC et technical computing d'origine ASUS vient également compléter la famille Tyan distribuée par PNY, pour offrir aux PME et ETI des solutions HPC clés en mains évolutives. D'ores et déjà disponibles eux aussi pour la France et l'Italie, ces serveurs se distinguent par leur configuration multiGPU qui permet, selon le parc applicatif auquel ils se destinent, de déployer des accélérateurs K10, K20 ou K20X en cluster ou en grille. Nous aurons prochainement l'occasion de revenir en détails sur ces machines - tests à l'appui - dans la mesure où elles sont appelées à jouer un rôle important dans la démocratisation du calcul intensif.

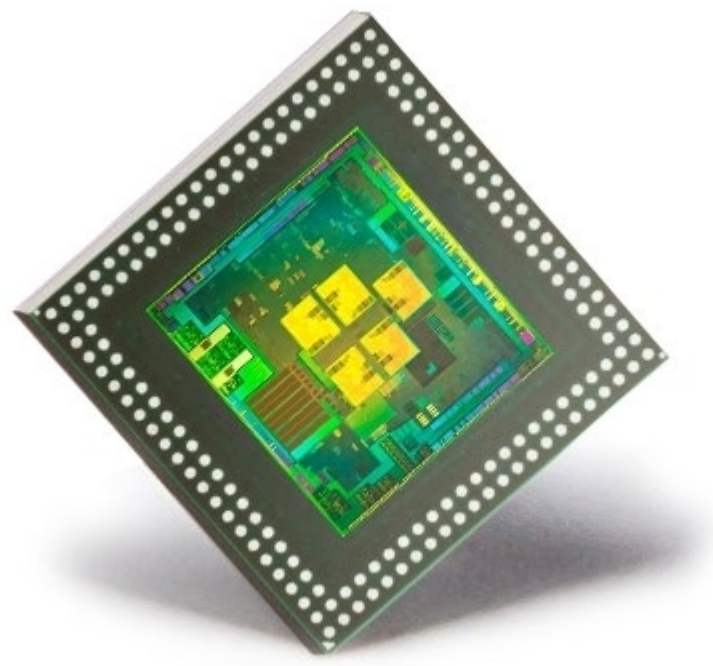
Enfin, mais c'est un secret, PNY devrait annoncer d'ici la fin de l'année un nouvel accélérateur GPU Tesla - nom de code "Atlas" - conçu autour du processeur Kepler GK110b en cours de finalisation chez NVIDIA. Malgré notre insistance, rien ne filtre concernant les spécifications de ce produit, qui devrait probablement s'appeler K40. On peut toutefois penser qu'il bénéficiera des mêmes améliorations que Quadro K6000, à savoir notamment 12 Go de GDDR5, un augmentation significative du nombre de cœurs, un TDP de 225 W et une pleine compatibilité PCIe 3.0. De quoi relancer les hostilités avec le côté bleu de la force, dont les prochains Xeon Phi (nom de code "Knight's Landing") devraient eux aussi se démarquer nettement des versions actuelles...

Datacentres (suite) : l'Europe se bouge, un peu...

Fort de leurs infrastructures hors-normes, les géants du Web sont en train de réinventer le Datacenter. Facebook avec son **Open Compute Project**, Google avec son organisation **The Green Grid** travaillent main dans la main avec les industriels pour définir les normes et les technologies qui feront les datacenters de demain. Clairement, le centre de gravité de l'innovation dans ce domaine a clairement basculé Outre-Atlantique.

Le message semble avoir été reçu par la Commission Européenne, qui vient d'octroyer un budget de 1,7 M€ à **PEDCA** (Pan-European Data Centre Academy) pour remettre le vieux continent dans la course. PEDCA fédère 13 partenaires aux Pays-Bas, en Allemagne et au Royaume-Uni, dont des collectivités locales, des centres de recherche, des universités et quelques industriels. Le projet vise à appliquer "*un plan d'action agissant sur de multiples leviers*", comme on dit à Bruxelles. En clair, il s'agit de financer non seulement des travaux de recherche et de standardisation technologique, mais aussi la mise en place de cursus de formation, de l'investissement dans les réseaux européens, des plans de lutte contre le cybercrime... le tout en 18 mois à peine ! Rendez-vous le 31 décembre 2014 pour voir si, compte tenu de cette manne généreuse, les grands équilibres ont bougé.

Barcelone continue d'explorer la piste de l'hybride ARM / GPU



Des calculateurs à base de CPU ARM ? L'intérêt que suscite la question ne se dément pas, efficacité énergétique oblige. C'est une voie qu'explore avec brio le BSC (Barcelona Supercomputing Center) depuis quelques années, avec une première machine déployée dès octobre 2011. Tibidabo, tel est son nom, tenait plus du POC que du calculateur à haut rendement scientifique, mais elle a permis de valider l'approche. Et de paver à voie à Pedraforca, calculateur "full-hybrid" dont la nouvelle version, en cours d'assemblage, se composera de CPU NVIDIA Tegra 3 (Cortex A-9 quadricores), de GPU NVIDIA Tesla K20 et de composants InfiniBand QDR.

Principale caractéristique de Pedraforca 2 : l'implémentation de connexions GPU à GPU et réseau en RDMA sur ARM. Le but est d'atteindre une consommation électrique minimale pour les applications qui s'exécutent principalement sur GPU, et de maximiser la bande passante InfiniBand QDR pour l'*offloading* distant des tâches massivement parallèles.

Dixit Alex Ramirez, responsable du projet à BSC, "*construire des prototypes de ce genre est indispensable pour faire avancer les développements logiciels, tant au niveau système qu'applicatif. Les multiples innovations de Pedraforca contribuent à l'enrichissement de la pile logicielle ARM, ce qui devrait donner naissance à des plateformes d'une bien meilleure efficacité pour les applications GPU-centriques, quelle que soit leur échelle de déploiement.*" Cofinancé par PRACE, Pedraforca 2 est coordonné au niveau technique par Bull, l'italien E4 Computer Engineering ayant été choisi pour la fourniture des nœuds. **AR**

Kalray : plus d'outils pour les développeurs

Que serait un processeur innovant sans un environnement de développement multifonctions digne de ce nom ? C'est pour offrir aux intégrateurs un contrôle total sur les vitesses d'exécution et la consommation énergétique que Kalray, architecte *fabless* du MPPA-256, annonce MMPA AccessCore, son IDE dédié, en version 1.0. Parmi les principales nouveautés, citons deux nouveaux modèles de programmation C/C++. Le premier se destine au déploiement automatique des applications sur les 256 cœurs disponibles. Le second autorise une gestion plus fine des ressources matérielles sous-jacentes. Il complète en cela les fonctions natives de la puce, qui sait réduire à la demande sa vitesse de traitement et stopper totalement les cœurs inoccupés. Kalray devrait d'ailleurs prochainement ajouter au tableau des possibilités d'introspection énergétique permettant aux applications de mesurer et d'adapter en temps réel leur consommation. Enfin, plusieurs outils de débogage, de profilage et d'optimisation font leur apparition, sans lesquels le difficile métier de développeur devient franchement pénible ! Pour mémoire, les processeurs Kalray MPPA-256 délivrent 230 Gflops pour 5 à 10 Watts mesurés, soit un rendement sensiblement supérieur aux architectures concurrentes. Et - cocorico - il s'agit d'une technologie française !

Colt dévoile ses ambitions HPC dans le Cloud

Les acteurs majeurs de l'hébergement et du Cloud sont de plus en plus nombreux à se positionner sur le segment du calcul intensif à la demande. Rien de surprenant, donc, à ce que Colt affiche à son tour ses prétentions. Allié à Intel et HP, le britannique s'est également associé au cabinet de conseil Aneo, dont les ingénieurs prendront en charge la gestion des grilles de calcul des clients HPC. Les derniers TechDays Microsoft de Paris ont d'ailleurs été l'occasion pour Aneo de démontrer une certaine maîtrise des architecture HPC hybrides basées sur Microsoft HPC Server 2012 et Azure.

Sans grande surprise, Colt a fait le choix de configurations HP Proliant SL / Intel Xeon E5 de densité élevée, gérées à partir de la solution de management HP CMU (Cluster Management Utility). Comme souvent, c'est le marketing qui va réellement faire la différence. Il faut en effet convaincre les clients de "monter" dans le cloud, mais aussi se positionner face aux géants du secteur, Microsoft et Amazon en tête, ainsi aux spécialistes tels qu'OVH/Oxalya et Bull Extreme Factory. Une chose est claire : dans le secteur, les grandes manœuvres ont commencé. Si tout se passe bien, Ce sont les tarifs qui devraient en bénéficier ! **AR**

MapR : du stockage à 100k IO/s plus ou moins gratuit...

Le californien MapR vient d'annoncer une nouvelle solution de stockage Big Data dans le Cloud Amazon. Le point intéressant, c'est que ses offres M3, M5 et M7, dédiées aux très grosses volumétries et aux besoins en faible latence, sont basées sur la solution Open Source HBase. Ce nom ne vous dit rien ? HBase a par exemple été retenu par Facebook pour sa messagerie instantanée *Messages*. Rien qu'avec elle, HBase absorbe chaque jour plus de 8 milliards de nouveaux échanges, soit 2 Po de données en ligne et plus de 75 milliards d'opérations de lecture/écriture quotidiennement. Pour monter son offre commerciale et gagner en performance, MapR a élagué HBase, ce qui lui permet d'en proposer plusieurs déclinaisons. L'entrée de gamme, M3, est gratuite ; M5 et M7, quant à elles, présentent des fonctions plus évoluées.

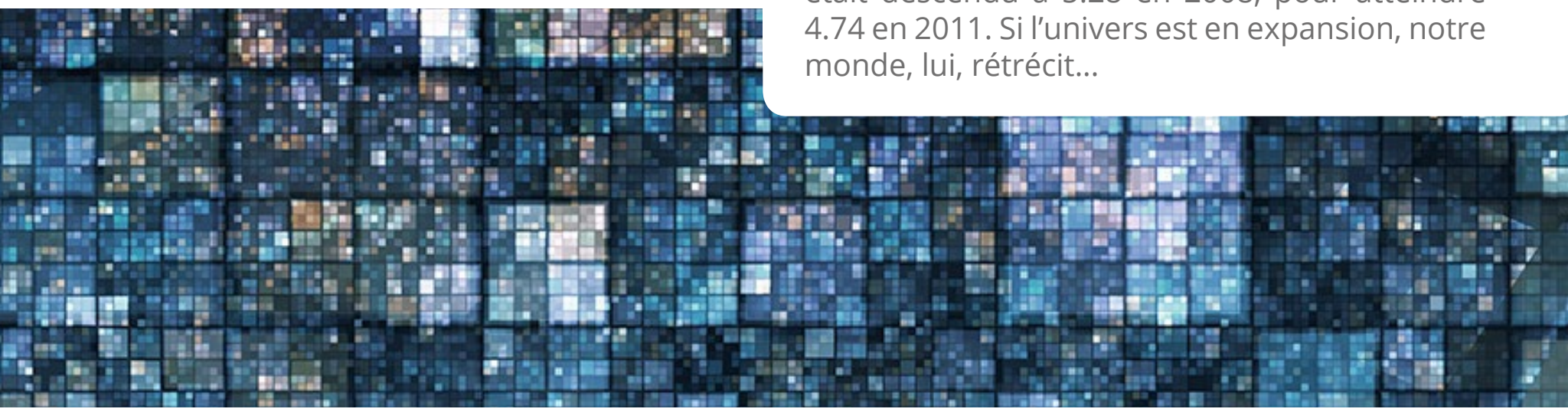
En portant son offre sur Amazon Elastic, MapR démontre une stratégie de prospection intelligente, qui propose des alternatives clés en main aux bancs de test applicatifs et autres *proofs of concept* déployés localement. L'éditeur souligne que, pour peu que l'on choisisse des instances Amazon EC2 haute performance avec stockage SSD ("High-I/O On-Demand Instances"), on peut obtenir des débits d'entrées/sortie assez élevés, de l'ordre de 100.000 opérations par seconde et par nœud sur le benchmark YCSB. L'idéal pour des besoins ponctuels de gros trafics destinés à devenir durables.... AR

/actus

"Data deluge" : de pire en pire !

On le sait, vaguement, mais mettre des chiffres sur le phénomène le rend toujours aussi étourdissant, même quand on est spécialiste. La dernière conférence TEDxGeorgeMasonU a été l'occasion pour Kirk Borne, professeur à l'Ecole de Physique, Astronomie et Informatique de ladite université, de partager quelques statistiques que nous ne pouvions pas ne pas rapporter dans ces colonnes. Il estime ainsi que, depuis la nuit des temps jusqu'en 2002, l'humanité a produit environ 5 Eo (10^18) de données. C'est alors qu'arrivèrent les premières applications Big Data. "Grâce" à elles, il ne fallait plus que 2 jours en 2011 pour produire le même volume d'information, une infinité qui se réduit à 10 minutes en 2013.

Pour Borne, "*le plus inquiétant est que le rythme s'accélère de façon très sensible*". A partir de la même métrologie, il prévoit que ces 5 Eo correspondront au volume de données produites en quelques secondes, tous les jours, à horizon 2020. Voilà qui explique pourquoi on compte aujourd'hui environ 100 offres d'emploi pour chaque candidat potentiel présentant un profil de Data scientist - un problème pour lequel les grands acteurs du secteur commencent à se mobiliser (voir notre actu ci-dessous). A titre anecdotique, cette tendance impacte aussi les "Six Degrees of Kevin Bacon", théorie selon laquelle tout être humain n'est distant de tout autre que par six degrés de séparation sociale. Avec l'avènement des Facebook, Twitter etc., le chiffre était descendu à 5.28 en 2008, pour atteindre 4.74 en 2011. Si l'univers est en expansion, notre monde, lui, rétrécit...



Comment former des experts Big Data ?

Compte tenu du manque de compétences dans les domaines de l'analyse de données à haute performance et plus globalement du Big Data, IBM multiplie les partenariats institutionnels pour créer des cursus de formation *ad hoc*. Plusieurs universités américaines ont déjà répondu à l'appel, notamment George Washington et Georgetown, ainsi que plusieurs autres à l'international comme la National University of Singapore, la Dublin City University en Europe et d'autres en Inde ou encore aux Philippines. Aucune université française ne figure pour l'instant sur la liste.

Aux termes de ces partenariats, IBM fournit des licences de ses logiciels les plus concernés, notamment les bases de données, mais il est aussi question de bourses sonnantes et trébuchantes pouvant aller jusqu'à 100.000 USD pour accompagner les étudiants dans leur cursus. Des dispositifs déjà testés avec succès par le géant américain ces dernières années pour former des experts mainframe.

Les choses se mettent donc en place pour alimenter un marché de l'emploi actuellement très demandeur de 'Data scientists' au profil à la fois statistique, informatique et métier. Pour eux, l'avenir s'annonce radieux : le Bureau of Labor Statistics, rattaché au département du travail US, estime la croissance de la demande pour ces postes à 24 % par an sur les 8 prochaines années. Voilà qui donnerait presque envie de retourner à la fac...

AR

DataScale : France *vs* Big Data

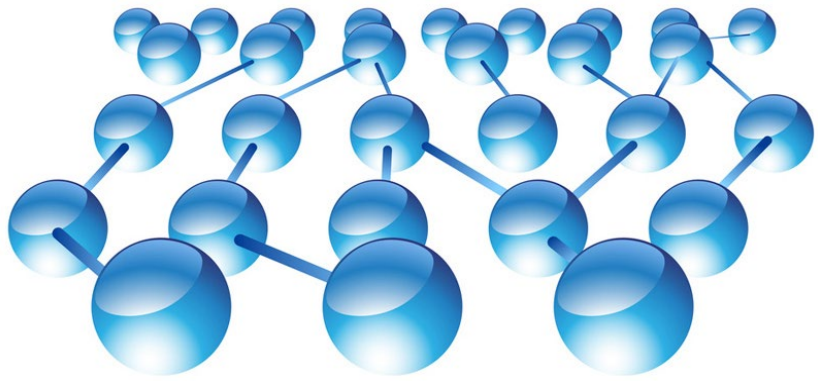
Démarré après Teratec, le projet DataScale a pour ambition d'apporter des réponses technologiques aux défis soulevés par l'explosion du Big Data. Indépendamment de l'aspect purement volumétrie, le terme recouvre en effet des problématiques liées aux types de données (homogènes ou non, structurées ou non, globales ou atomiques...) et aux algorithmes de traitement (archéologie de données, traitements à la volée, consolidation...). Face à cela, DataScale s'est donné pour mission de créer des briques technologiques *ad hoc* et de les démontrer à partir d'une palette d'applicatifs et de leur passage à l'échelle. En ligne de mire, la transition programmée vers le cloud, qui nécessite des middlewares adaptés et de nouvelles avancées dans l'architecture des bases de données.

L'effort est à saluer. D'abord parce qu'il regroupe autour de Bull un éventail d'acteurs (dont le CEA, l'Inria, l'IPGP, ActiveEon, Armadillo...) constitués d'institutionnels, de grandes entreprises et de PME. Ensuite parce que priorité est données aux résultats. La durée de l'initiative est en effet fixée à 24 mois, avec une liste de livrables très précise. Enfin parce qu'il se focalisera sur trois grands domaines d'application Big Data à partir desquels il sera possible de dériver les progrès accomplis. Primo, un code sismique de type "e-science", domaine où la multiplication des capteurs et l'apparition de nouveaux instruments expérimentaux aboutit à la création de nouvelles masses de données. Secundo, une application dédiée à la surveillance d'une grande infrastructure technique, avec monitoring d'un calculateur HPC. Tertio, une plateforme e-commerce et multimédia incluant en particulier des analyses poussées d'opinions et d'avis.

Une fois abouti, le projet DataScale devrait engendrer des outils qui profiteront à bon nombre de piles logicielles. On en attend notamment l'intégration de fonctions de gestion hiérarchique dans Lustre (avec DLM - *Data Lifecycle Management*), une alternative à MapReduce avec interfaçage Lustre et une architecture de SGBD distribuée incluent des index ensemblistes. Bon courage, Messieurs (et Mesdames !)...

/actus

GÉANT change de dimension



C'est à un gros chantier que s'attaque la Commission Européenne. Bruxelles veut en effet moderniser GÉANT, le réseau d'interconnexion de l'éducation et de la recherche qui maille l'ensemble du vieux continent. Les débits devraient ainsi passer de 100 Gbit/s à 2 Tbit/s d'ici 2020, une entreprise titanesque dans le mesure où le backbone dessert une centaine de réseaux nationaux jusqu'au Royaume-Uni et en Irlande, soit environ 10 000 institutions. Il faut rappeler de plus de 1 000 To de données circulent quotidiennement via GÉANT dont celles des grands centres HPC européens, CERN inclus. C'est DANTE (Delivery of Advanced Network Technology to Europe), opérateur fédérateur du réseau en complément des entités nationales, qui mène cette modernisation, avec pour objectif le renouvellement de 50 000 km de câble et de fibre. On s'en était fait l'écho, son équipe de recherche a récemment établi une connexion à 26,02 Tbits entre Amsterdam et Hambourg via des équipements optiques Infinera. Ce sont donc logiquement les matériels de ce dernier, en l'occurrence des DTN-X, qui ont été retenus pour la mise à jour, ainsi qu'une variété de commutateurs Juniper.

Résultats de notre sondage-express de juillet-août

Question : *Avez-vous des projets sur architecture ARM ?*

Oui : 26,8 % Non : 73,2 %

Opinions exprimées : 211

Merci d'avoir voté
sur www.hpcmagazine.fr

Qui veut gagner ~~un~~ 2 accélérateurs AMD ?

Les gagnants de l'été !

Ont obtenu la faveur du sort, de la destinée, du dieu des jeux-concours :

- **François RUE (Inria), qui gagne l'accélérateur serveur FirePro S9000***
- **Pascal TREMBLAY (Michelin), qui gagne l'accélérateur serveur FirePro S7000*.**

Nos félicitations aux heureux gagnants - et bon développements à eux, en OpenCL !

Les prix leur seront envoyés par La Poste à l'adresse qu'ils ont indiquée.

(1) 2,23 Tflops SP, 806 Gflops DP, 6 Go GDDR5 ECC, 264 Go/s, 235 W, PCIe 16X, 1 DisplayPort, Compatible OpenCL 1.2, OpenGL 4.2, DirectX 11.

(2) 2,4 Tflops SP, 152 Gflops DP, 4 Go GDDR5, 154 Go/s, 150 W, PCIe 16X, 1 DisplayPort, Compatible OpenCL 1.2, OpenGL 4.2, DirectX 11.

FIREPRO GRAPHICS AMD

Des milliers de cœurs... rien que pour mon projet !

Disposer de tout un centre de calcul pour soi ? Grâce aux "méso-challenges", certains chercheurs en ont fait l'expérience...

Les dix centres de calcul ayant récemment bénéficié d'un financement Equip@Meso / Investissements D'avenir ont profité de la livraison de leurs nouveaux calculateurs pour offrir à plusieurs équipes de recherche des conditions de calcul exceptionnelles, avec pour objectif la réalisation de simulations inédites. Avant l'ouverture des machines en mode partagé à l'ensemble de leurs utilisateurs, celles-ci ont été réservées dans leur intégralité à un ou plusieurs projets, avec tout l'accompagnement technique et scientifique nécessaire.

Les différents centres impliqués ont ainsi pu proposer un accès privilégié à 1 000 à 3 000 cœurs, ce qui a conduit à la réalisation, sur une période concentrée de quelques semaines, de simulations nécessitant des millions d'heures de calcul. Ces simulations - les "méso-challenges" - ont concerné un large éventail de disciplines : astrophysique, CFD, chimie, mathématiques appliquées et expérimentations HPC.

L'initiative a permis de mettre à l'épreuve à la fois les machines, les codes et les modèles numériques, mais aussi d'obtenir des résultats scientifiques dont certains constituent de véritables avancées dans leur domaine. Les challenges ont en général été sélectionnés par des appels d'offres scientifiques dont la ligne directrice était en substance "Nous vous ouvrons tout le calculateur, soyez ambitieux".

Suivant cette directive implicite, les équipes retenues ont su parfaitement calibrer leur problème à la taille de la machine cible et ainsi utiliser au mieux la puissance de calcul offerte, qui se situe entre celle disponible au laboratoire et celle des grands centres nationaux. Toute la souplesse des mésocentres a pu être mise à profit pour adapter les couches logicielles (middleware, GPU) aux besoins des chercheurs.

Les résultats de ces challenges seront présentés le 20 Septembre 2013 à l'Institut Henri Poincaré, dans le cadre de la première journée Méso-Challenges du projet Equip@Meso. Cette journée s'articule autour de quatre sessions thématiques (voir le programme détaillé page suivante). Elle sera également l'occasion d'échanger entre scientifiques, responsables de centre et institutionnels, et se clôturera par une remise des prix. L'inscription, gratuite, est ouverte à tous : http://www.genci.fr/fr/node/378.

RD / NM / ARd

Journée Méso-challenges
Le programme complet

9h00-9h20 - Ouverture

9h20-10h15 - Session Astrophysique. Mésocentres contributeurs :
- Strasbourg
- Lyon
- Paris Sciences & Lettres.

10h15-10h40 - Pause café.

10h40-11h35 - Session Chimie Quantique / Physique Théorique. Mésocentres contributeurs :
- Toulouse (Calmip) - Physique Théorique
- Toulouse (Calmip) - Chimie Quantique
- Grenoble - Nanosciences.

11h35-12h15 - Table Ronde : Méso-challenges (animée par Michel Kern, INRIA / MdS).

12h15-13h45 - Repas offert par le projet Equip@Meso.

13h45-14h40 - Session Mathématiques / Optimisation des codes. Mésocentres contributeurs :
- Strasbourg - Mathématiques / Biosciences
- PSL - Mathématiques
- MdS - Optimisation et parallélisation de code.

14h40-15h35 - Session CFD. Mésocentres contributeurs :
- Rouen (Crihan)
- Lyon
- Grenoble.

15h35-16h00 - Pause café.

16h00 - Table Ronde : présentation des partenaires adhérents.

16h20 - Remise des prix.

GENCI à l'initiative

Equip@Meso est un projet EquipEx 2010, coordonné par GENCI, qui a reçu 10,5 millions d'euros des Investissements d'Avenir pour renforcer les équipements de calcul régionaux. Outre GENCI, en charge de la politique nationale du calcul intensif, il regroupe 10 centres régionaux et universitaires, dits mésocentres, à Rouen, Paris, Reims, Strasbourg, Lyon, Grenoble, Marseille et Toulouse. La journée Méso-challenges aura lieu le lendemain de la 6ème journée Mésocentres organisée chaque année par le Groupe Calcul.

/actus

Technical Computing : Le roadshow IBM s'arrête à Paris

Terme désormais pleinement consacré, "Technical computing" recouvre l'ensemble des technologies et applications du calcul intensif, de la simulation numérique et du Big Data. Pour mieux faire connaître ses enjeux et ses bénéfices aux entreprises soucieuses de leur compétitivité, IBM organise un roadshow européen avec une étape à Paris. Comme le précise le programme ci-dessous, cette demi-journée, organisée le 24 septembre aux Salons Etoile Wagram, mixe présentations techniques, cas d'applications concrets et témoignages clients. De quoi, pour les PME et ETI notamment, se faire une idée précise de ce que ces technologies nouvelles, aujourd'hui faciles à acquérir et à déployer, peuvent véritablement leur apporter dans l'exercice quotidien de leur activité. Bonne nouvelle, l'inscription est gratuite.

8h30 - Café d'accueil.

9h00 - Introduction.

9h10 - Présentation. Le calcul scientifique : tendances, défis et vision.

9h45 - Les 10 meilleures pratiques pour votre environnement de calcul hautes performances.

10h15 - Pourquoi vos applications d'analytique ont-elles besoin d'une grille de calcul et de données à faible latence ?

10h45 - Pause.

11h00 - Combler les lacunes du phénomène Big Data open source, Hadoop et MapReduce.

11h30 - Témoignage(s) client(s).

12h00 - Simplifier le calcul hautes performances (HPC) pour les départements et plus petites organisations.

12h30 - Comment démarrer ? les différentes étapes.

12h50 - Conclusion – Questions/Réponses.

13h00 - Cocktail déjeunatoire.

Tous
nos articles,
nos dossiers,
nos chroniques,
nos codes sources
et bien plus encore...

www.hpcmagazine.fr

DEVLOG

JDEV

Journées **Développement Logiciel** de l'Enseignement Supérieur et de la Recherche

Conception graphique : Benoît Colas, Université de Toulouse-Le Mirail / CPRS - UMS 838. Photographie : © DR.

Webcast

4 - 5 - 6 septembre 2013, **École polytechnique**

http://devlog.cnrs.fr/jdev2013

Contact : contact-jdev2013@services.cnrs.fr

cnrs · École Polytechnique ParisTech · INRA Science & Impact · Inria Informatique mathématiques · DEVLOG Réseau du développement logiciel · LAAS-CNRS · LOOPS

L'AGENDA

SEPTEMBRE 2013

JDEV 2013

Où : Polytechnique (Palaiseau)
Quand : 4-5-6 septembre
La deuxième édition des Journées nationales du Développement Logiciel aura lieu les 4-5-6 septembre 2013 à l'Ecole Polytechnique. Il s'agit d'une Action Nationale de Formation inter-établissements, soutenue par la Mission pour l'interdisciplinarité du CNRS, l'INRIA et l'INRA. Ouverte à tous les acteurs du développement de logiciels dans l'Enseignement Supérieur et la Recherche, cette manifestation se donne pour objectif d'améliorer la qualité et la pérennité de nos productions logicielles, ainsi que d'ouvrir de nouvelles perspectives avec la présentation des derniers outils et méthodes existants. Les pratiques adaptées au contexte scientifique seront privilégiées.

LAD'13

Où : Paris (Hôtel Lutetia)
Quand : 16-17 septembre
EOFS and OpenSFS are organizing the third European Lustre Workshop, called **Lustre Administrators and Developers Workshop**. It will take place in Paris, at Hotel Lutetia during 2 days, 16th and 17th of September, 2013. This will be a great opportunity for LustreTM worldwide administrators and developers to gather and exchange their experiences, developments, tools, good practices and more. Check last year LAD to see content and events of 2012 session. On Monday evening, a social event will take place on yacht, near the venue. There will be a dinner and a boat tour on Seine river, aboard the Excellence Yacht. A limited number spouses can attend too. Availabilities are on a first-come, first-served basis.

6èmes journées Mésocentres 1ère journée Mésochallenges

Où : Paris (IHP)
Quand : 19-20 septembre
Organisées respectivement par le Groupe Calcul et GENCI, deux événements d'importance pour la vie du HPC en région. Voir nos actualités.

ISC Cloud'13

Où : Heidelberg (Allemagne)
Quand : 23-24 septembre
The remarkable success of the last ISC Cloud conferences has motivated the organizers to continue with this conference series in 2013, with a profound focus on the use of Clouds for High Performance Computing (HPC), for end-users and service providers. This conference is highly valuable for members of the HPC community who want to understand, join, and benefit from this massive trend of mainstream on-demand HPC.

ISC Big Data'13

Où : Heidelberg (Allemagne)
Quand : 25-26 septembre
The intersection of big data and HPC will be a common theme at the upcoming ISC Big Data conference in Heidelberg, Germany, on September 25-26. On the second day of the event, Dr. Flavio Villanustre will deliver a keynote around this topic, offering his perspective on the merging of big data in HPC and the commercial realm. Villanustre, who is the vice president of Infrastructure & Products for the HPCC Systems group and vice president of Information Security at LexisNexis, gave us a preview of his talk, expounding on some of the technologies driving this trend.

PRACE Autumn School 2013

Où : Ljubljana, Slovenia
Quand : 23-27 septembre
The PRACE Autumn School communicates and discusses issues and perspectives of HPC targeting industrial applications arising from evaluation performance and/or design of products, e.g. equipment and processes, with a particular emphasis on the automotive, aerospace, and energy fields. The PRACE Autumn School places priority on algorithms, simulation strategies, and programming techniques for complex fluid flow simulations, structural mechanics and electro-magnetics which usually have coupled multiphysics requiring intensive use of HPC resources. In case of expertise and software availability, there will be presentations of other issues, e.g. robustness and performance analysis, a brief introduction to pre- and post-processing techniques, e.g. CAD integration, mesh generation and visualization are presented. ■

ISC events cloud computing

ISC CLOUD'13

HPC and Manufacturing Meet Cloud

September 23–24, 2013 | Heidelberg, Germany

Join research and industry experts to learn about, discuss and exchange your experience in designing, building, operating, and using compute and data-intensive cloud services

Keynote & Platinum Presentations:

HPC and Supercomputing - On our Way Towards a New Utility

Jason Stowe,Cycle Computing

Technical Computing, Mid Market, Cloud, and the Architecture for Discovery

Stephan Gillich, Intel EMEA

Leveraging the Cloud for Digital Manufacturing: Progress and Challenges

Frank Baetke & Jean-Luc Assor, HP

Highlights:

Hands-on Tutorial: Building your Cloud for HPC, here and now, in 3 hours! · Industry & Research Case Studies · Expert Panel · Digital Manufacturing as a Service, in the Cloud · Engineering Application Software in the Cloud · Cloud Service Providers · Challenges in the Cloud · SLA's & Licensing · Industry Exhibition · Full Catering · Get Together Networking Party

For more information and registration please visit:

www.isc-events.com/cloud13

Sponsors:

Platinum: CYCLECOMPUTING, hp, intel — Gold: Mellanox TECHNOLOGIES

Silver: advania data centres, ANSYS, BULL, sgi, transtec

Bronze: C12G LABS, OpenNebula.org, IDC Analyze the Future, Intersect360 RESEARCH

ISC events BIG DATA

ISC BIG DATA'13

Where Enterprise and HPC Meet

September 25 – 26, 2013 | Heidelberg, Germany

Join us for our international ISC Big Data conference, held immediately after ISC Cloud'13

Keynotes:

Big Data - Hype or Disruptive Innovation?

Felix Wortmann, University of Sankt Gallen

Big Data for Big Science: Data-intensive discovery with the co-evolution of hardware and software platforms

Girish Juneja, Intel Corporation; Stephan Gillich, Intel GmbH

HPC & Data Intensive Computing in Science and Industry

Flavio Villanustre, HPCC Systems & LexisNexis

SAS High-Performance Analytics in Action – Concepts & Examples How Big Data Changes the Way of Business

Mark Torr, SAS Institute

Highlights:

New Computational Models for Big Data · How to increase the ROI through Big Data Analytics - Real-Life Studies · Learning from Big Data Applications in Science · HPC meets Enterprise · Vendor Solutions for Big Data Challenges · Efficient & Scalable Big Data Architectures for Search & Mining · Panel Discussion: Current State and Future Opportunities for Big Data · Industry Exhibition · Full Catering · Get Together Networking Party

For more information and registration please visit:

www.isc-events.com/bigdata13

Sponsors:

Platinum: intel, sas THE POWER TO KNOW.

Gold: Mellanox TECHNOLOGIES, Quantum — Bronze: IDC Analyze the Future, Intersect360 RESEARCH

/en_couverture

4 SIMULATION NUMERIQUE : CAS D'USAGE EXEMPLAIRES

Les témoignages qui constituent ce dossier ont été sélectionnés par la rédaction pour leurs vertus didactiques. Par souci de vérité, nous avons cité les marques et produits réellement choisis par les utilisateurs pour mener à bien leurs projets mais, sauf spécificités techniques particulières, les usages décrits sont reproductibles avec des produits concurrents. Ce dossier n'a donc rien à voir avec une entreprise de publicité indirecte, déguisée ou subliminale. Qu'on ne s'y méprenne !

Alors que HPC et Technical computing s'ouvrent désormais à toutes les entreprises, ils nous a paru pertinent de vous présenter quatre cas d'usage typiques, quatre *success-stories* qui racontent en détails ce que la simulation numérique peut apporter de bénéfices concrets aux équipes qui l'adoptent…

1 - Philips Healthcare : Simulation électronique et automatisation de processus

Les systèmes d'imagerie par résonance magnétique (IRM) utilisent des fréquences radios élevées et des champs magnétiques puissants afin d'aligner systématiquement les moments magnétiques des protons des tissus humains. Mais pour qu'un système IRM soit capable de produire des images de qualité, le timing et la linéarité de la puissance de l'onde radio doivent être contrôlés avec précision. Sachant que le pic de puissance des amplificateurs RF atteint 40 kW, maintenir la précision d'un signal de phase dans un intervalle de quelques picosecondes est un véritable défi d'ingénierie.

C'est ce qu'ont réalisé les ingénieurs de Philips Healthcare en développant, grâce à la simulation, un sous-système d'alimentation RF numérique unique en son genre. A partir de plusieurs outils logiciels Mathworks et Mentor Graphics, ils sont parvenus à atteindre la précision et la linéarité requises en développant un algorithme de prédistorsion et de rétrocontrôle performant, qu'ils ont ensuite automatiquement implémenté en VHDL pour pouvoir le déployer sur des FPGA.

"*L'étape de création de code VHDL est incontournable, mais ce n'est pas une activité à laquelle nous apportons beaucoup de valeur ajoutée*", explique Mark van Helvoort, responsable du projet chez Philips Healthcare. "*Avec des outils de simulation tels que Simulink et HDL Coder, de MathWorks, nous pouvons concentrer nos ressources sur le développement d'algorithmes et la conception applicative avancée plutôt que de perdre du temps sur la syntaxe VHDL et les règles de codage.*"

Les systèmes IRM de Philips Healthcare comptent parmi les plus performants au monde. L'utilisation de la simulation numérique a permis le mise au point d'innovations telles que la multitransmission, qui élimine les ombrages diélectriques par l'utilisation de multiples sources RF et l'optimisation automatique de la puissance, de l'amplitude, de la phase et des formes d'ondes du signal.

Eviter la programmation manuelle

Comme souvent dans l'industrie, ce sont les ingénieurs hardware de Philips qui sont en charge d'implémenter en FPGA les algorithmes C développés par les architectes système. Or, cette traduction ne va jamais sans poser de problèmes - des problèmes auxquels s'ajoutent à la fois les différentes habitudes de codages des développeurs et la gestion des variations de linéarité des amplificateurs. On sait également que la programmation manuelle ralentit les itérations de conception, c'est-à-dire le débogage des algorithmes existants, les essais de nouveaux algorithmes ou le testing de nouveaux designs sur les FPGA cibles.

A tous ces niveaux, l'adoption de la modélisation et de la simulation numérique est intervenue comme un réel accélérateur de productivité , elle-même gage d'amélioration sensible de la qualité. En démultipliant les possibilités de conception, en réduisant sensiblement les délais entre itérations, Philips Healthcare est parvenu à résoudre nombre de difficultés industrielles typiques des méthodes d'ingénierie traditionnelles, dites aujourd'hui "à l'ancienne".

Modéliser de bout en bout

Concrètement, les ingénieurs ont utilisé Simulink pour modéliser un sous-système constitué d'un récepteur/comparateur RF numérique, d'un générateur d'ondes et d'un ampli. Dans cette architecture, les sorties *forward* et *reflected* sont captées et renvoyées au récepteur numérique, qui les compare au

signal d'entrée requis. Un signal cible corrigé est alors créé en utilisant la prédistortion numérique, puis envoyé vers le générateur d'ondes.

D'un point de vue technique, le modèle de générateur d'ondes RF réalisé par l'équipe comprend une porteuse NCO (Numerically-controlled oscillator), une modulation BLU et un filtre CIC (Cascaded integrator-comb), ainsi que des composants FIR et un modulateur multibande. Le récepteur/comparateur, quant à lui, inclut un convertisseur analogue-numérique, un démodulateur et des blocs de *down-conversion* numériques.

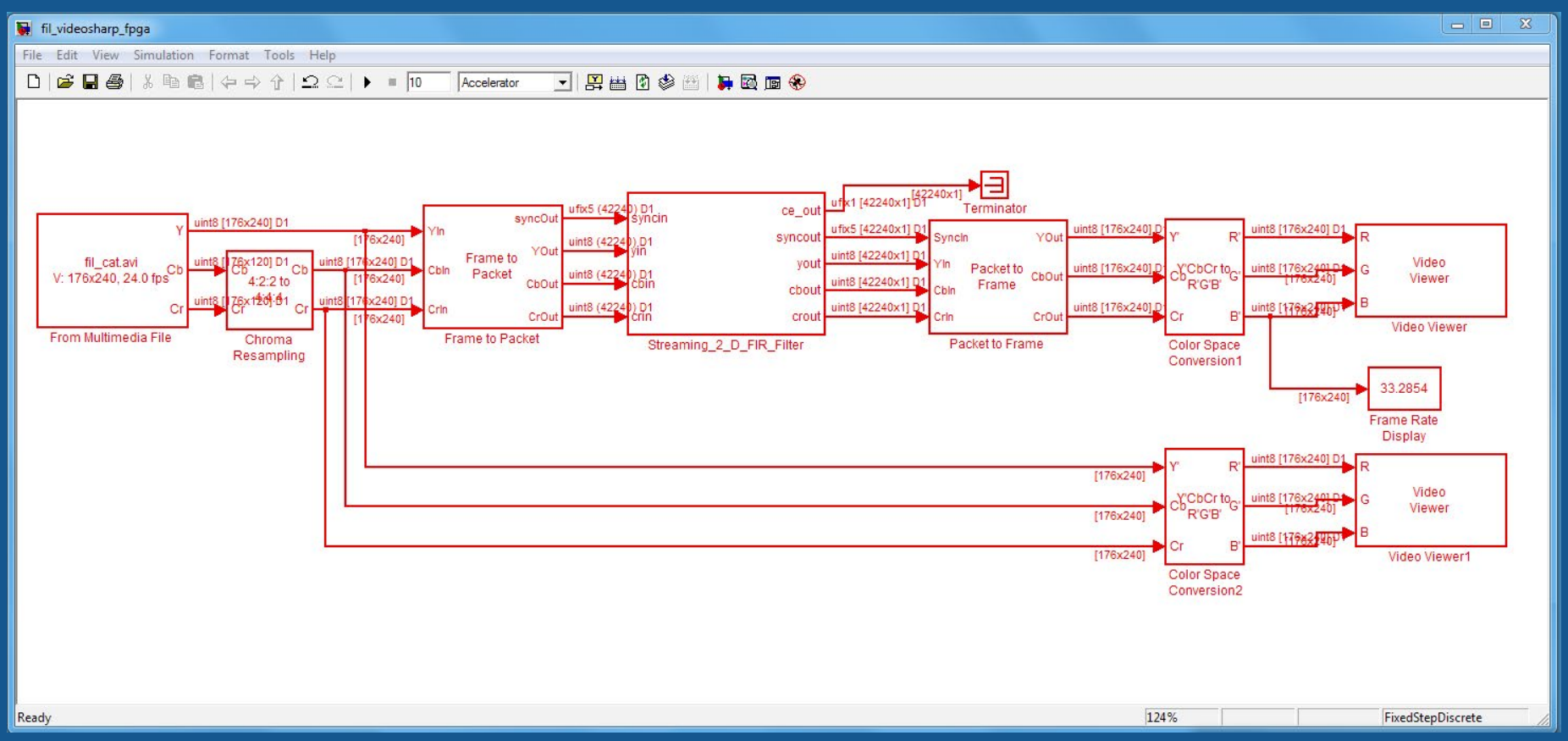
L'équipe a modélisé ces composants avec une précision de niveau virgule flottante pour obtenir une qualité de simulation lui permettant de tester, de déboguer et d'optimiser la chaîne de transmission RF avec plus de marge qu'en mode physique. Ils ont ensuite converti le modèle en précision de niveau virgule fixe, avec les vérifications supplémentaires qui s'imposent.

Une fois le modèle finalisé, il a servi de générateur de bibliothèques de code VHDL. L'importation directe de ces bibliothèques dans ModelSim, de Mentor Graphics, a permis d'intégrer l'ensemble dans le design global du FPGA. Après validation, là encore par simulation, l'ensemble pouvait être gravé dans le FPGA cible, en l'espèce un Xilinx Virtex-6. Les problèmes rencontrés pendant les tests ont pu être reproduits, diagnostiqués et résolus dans le modèle, après quoi l'équipe a repris le cycle de génération de nouvelles bibliothèques VHDL, de test et de mise à jour du silicium.

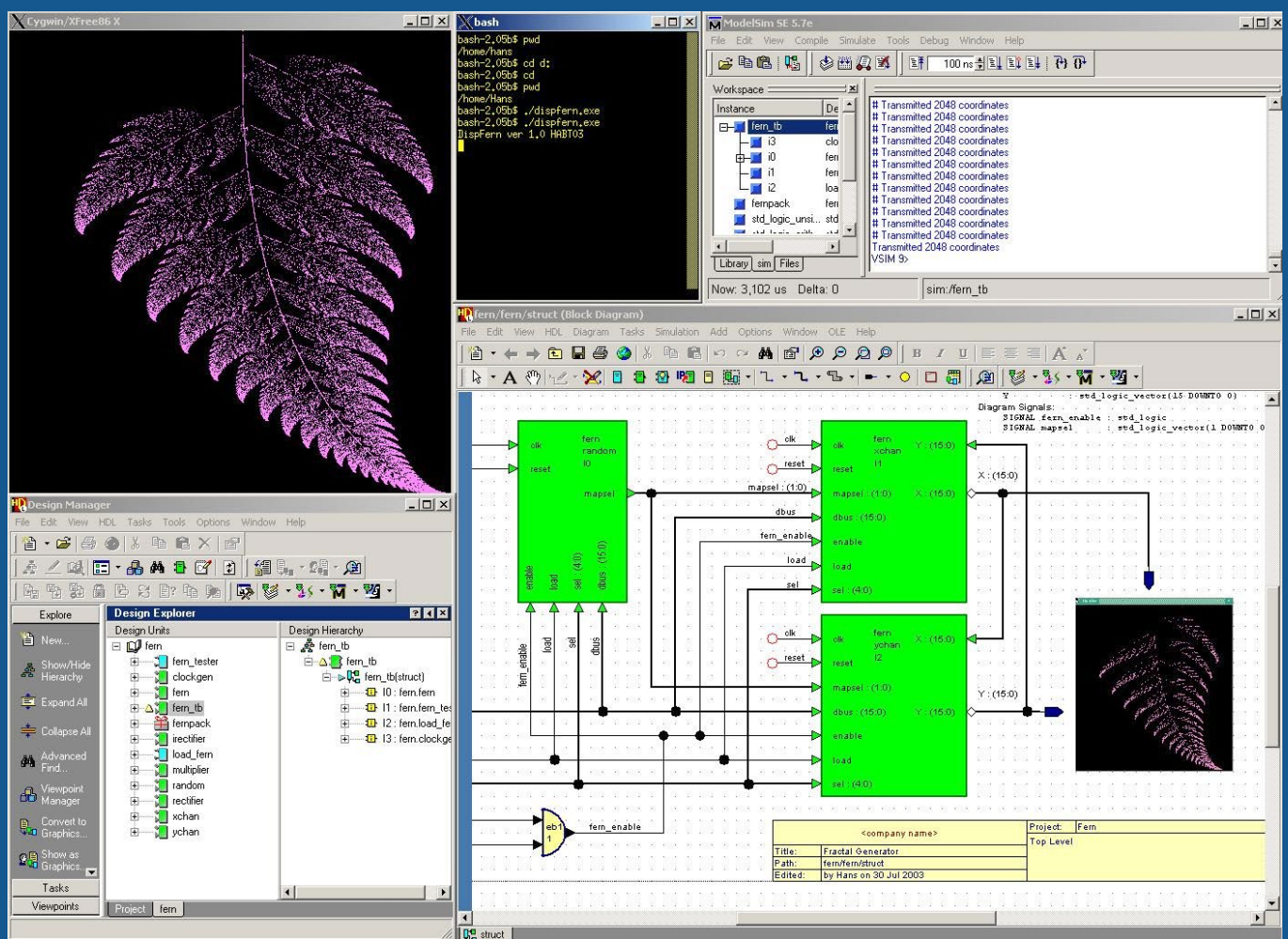
Résultats concrets

Premier bénéfice cette conception par simulation, les erreurs de conception ont été corrigées dès les premières phases de développement. "*Les simulations en virgule fixe ont révélé une fréquence parasite que notre architecte système a considéré comme inacceptable*", précise Marcel van Bakel, Electronics Designer chez Philips Healthcare. "*En utilisant notre plateforme de simulation, nous avons pu localiser le problème - une simple erreur d'arrondi - et y remédier immédiatement.*"

De plus, les options possibles ont été rapidement évaluées et implémentées. "*Un de nos premiers designs mettaient en œuvre un filtre RF 24-stage, mais nous nous sommes aperçus qu'il consommait trop de ressources FPGA*" ajoute van Bakel. "*En moins de 30 minutes, nous avons*



Exemple d'une simulation électronique modélisée via MATLAB et Simulink, codée en HDL, adaptée au FPGA cible puis optimisée pour la performance.



La technologie de single kernel generator (SKS) de ModelSim, ainsi que son débogueur unifié (VHDL, Verilog, SystemC) en font un composant prisé de la chaîne de conception électronique vers FPGA.

modélisé un filtre simplifié, l'avons intégré au design initial puis l'avons validé en utilisant un analyseur de spectre directement dans Simulink. Ce type de modification aurait pris une semaine à coder et à tester en VHDL."

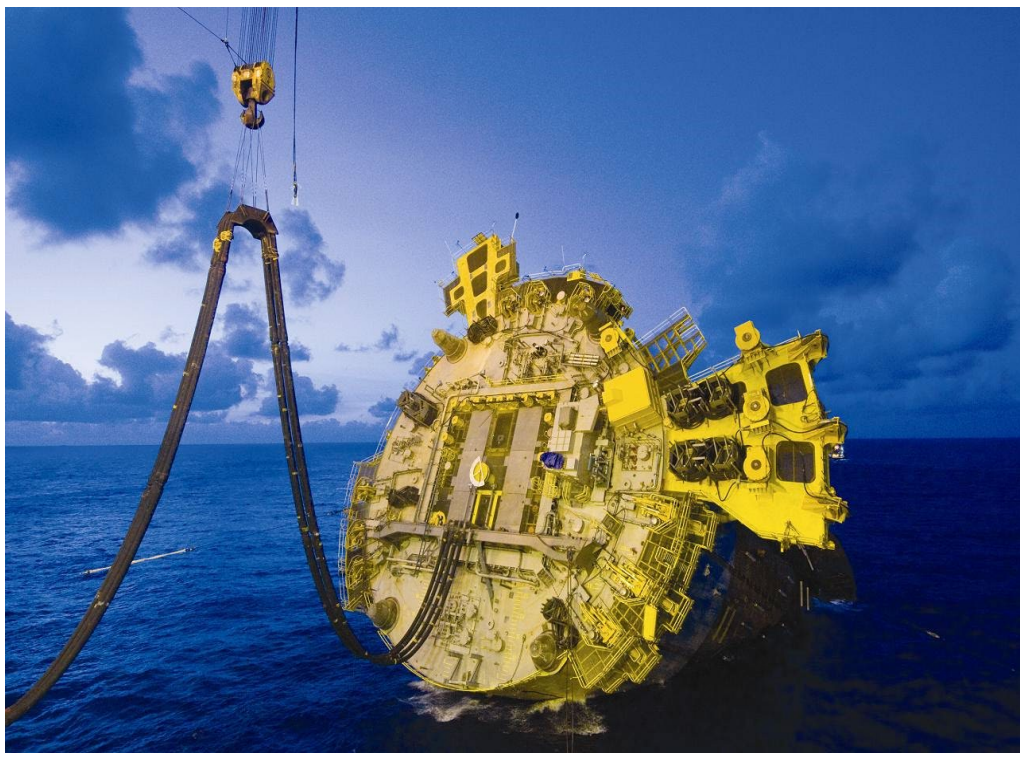
Enfin, du point de vue industriel, la cohérence des processus et la prévisibilité sont améliorés. "*Contrairement au code VHDL, typiquement conçu par plusieurs développeurs donc avec plusieurs style de codage, le code généré automatiquement est cohérent. De ce fait, les outils qui l'utilisent en aval se comportent de façon prévisible*", conclut Marcel van Bakel. "*La génération de code est non seulement plus rapide, mais elle réduit les risques d'erreur.*"

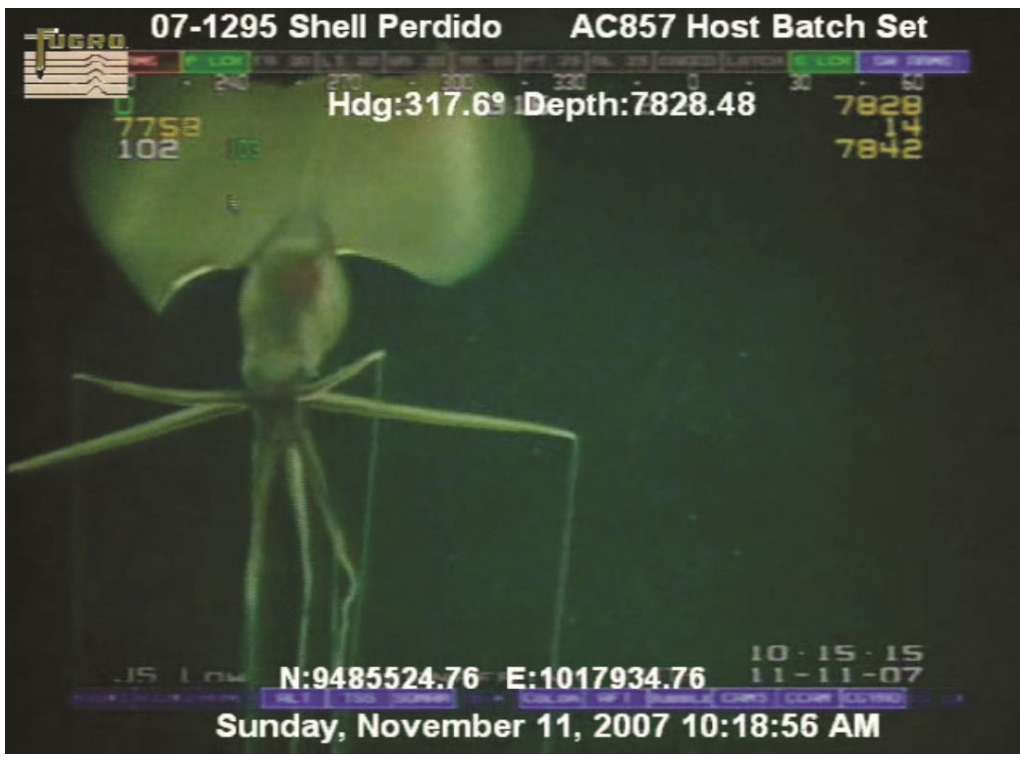
2 - Royal Dutch Shell : Le calcul intensif pour explorer plus vite, plus profond et plus propre

Le HPC est un outil critique pour les activités de **Shell**. Disposer de capacités de calcul intensif bien dimensionnées permet de trouver des hydrocarbures dans des environnements géologiques de plus en plus complexes et difficiles d'accès, de les extraire plus efficacement, et de répondre à la demande énergétique du monde en réduisant l'impact écologique de l'extraction.

Les besoins de Shell en matière de calcul augmentent régulièrement - et considérablement. Pour l'exploration en amont, les algorithmes propriétaires d'acquisition sismique génèrent des datasets de plus en plus volumineux. A cela s'ajoute des modèles de réservoir de plus en plus complexes, découlant de techniques de simulation avancées. Enfin, les programmes du département R&D et de nouvelle applications d'analyse intensive de données requièrent pour leur part également des capacités HPC de toute dernière génération.

Dans ce contexte, Shell a évidemment pour objectif de fournir à ses équipes la capacité HPC requise, au TCO le plus contenu possible. La compagnie s'appuie pour cela sur ses deux datacentres de Houston et Ams-





L'extraction d'hydrocarbures sous-marins est une activité difficile, dangereuse pour l'environnement et extrêmement coûteuse. La simulation numérique de réservoirs permet de réduire sensiblement les risques de pollution et de maximiser l'efficacité des implantations physiques.

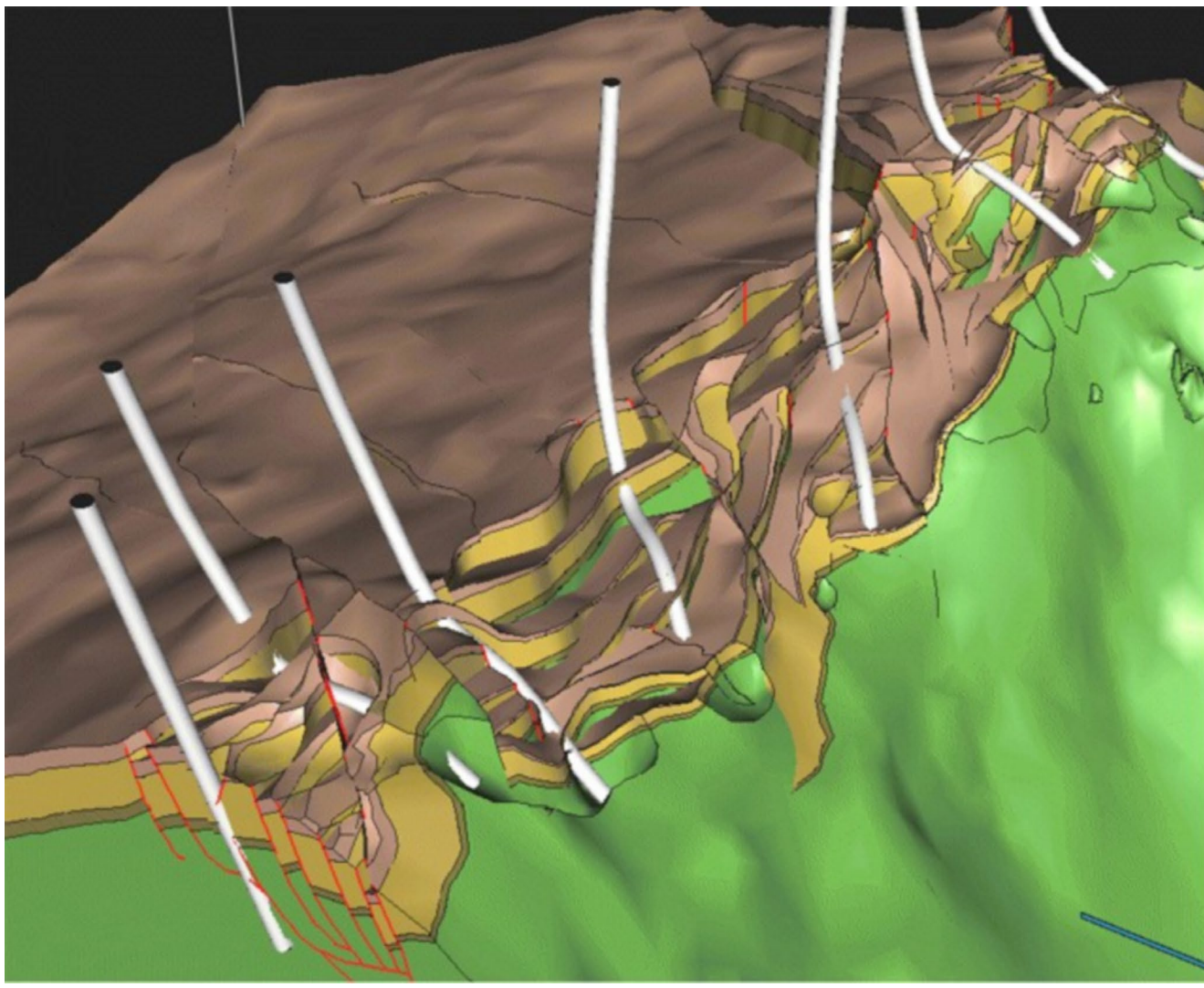
terdam, le mot d'ordre étant de concentrer les ressources de calcul sur un nombre restreint de sites. Cette approche autorise un certain nombre d'économies d'échelle, permet de maintenir un haut niveau de service et devrait s'avérer pertinente lors de la transition vers les solutions cloud, qui s'amorcera dès que les propositions techniques auront atteint un degré de maturité suffisant.

L'importance qu'a pris le HPC pour Shell l'a conduit à établir un Centre d'Excellence (CoE) dès 2009. Sa mission : multiplier par cent le rapport performance/coût. L'adoption régulière des dernières technologies et un effort constant sur les codes logiciels s'inscrivent dans cet objectif, auquel contribue également plusieurs partenariats étroits avec des fournisseurs de technologies critiques. A terme, ces efforts ont vocation à offrir à Shell un avantage concurrentiel et technique déterminant pour le devenir de l'entreprise.

Une vision globale du HPC

La philosophie de Shell consiste à penser que le HPC ne concerne pas seulement le calcul. Pour être vecteur d'une valeur ajoutée d'usage maximale, il doit inclure les applications, les données, le stockage et son intelligence, les réseaux et les technologies de visualisation. L'approche du Centre d'Excellence HPC est donc la plus globale, la plus transverse possible. Pour cela, des contrats de coopération internes ont été

instités. Avec le département Innovation, Recherche et Développement (IRD), qui fournit les technologies les plus élaborées dans le domaine géophysique. Avec le département Technique et Compétitivité (TaCIT), dont les ingénieurs logiciel optimisent continuellement les algorithmes métiers de la compagnie. Avec le service des Configurations Techniques (Technical Landscape Services), qui recette et intègre les infrastructures et services HPC afin de garantir les meilleurs niveaux de disponibilité et d'optimisation pour les workflows métier.



Modélisation type d'un réservoir, obtenue après traitement de datasets à la fois volumineux et complexes. On voit sur l'image la simulation de l'implantation des points d'extraction.

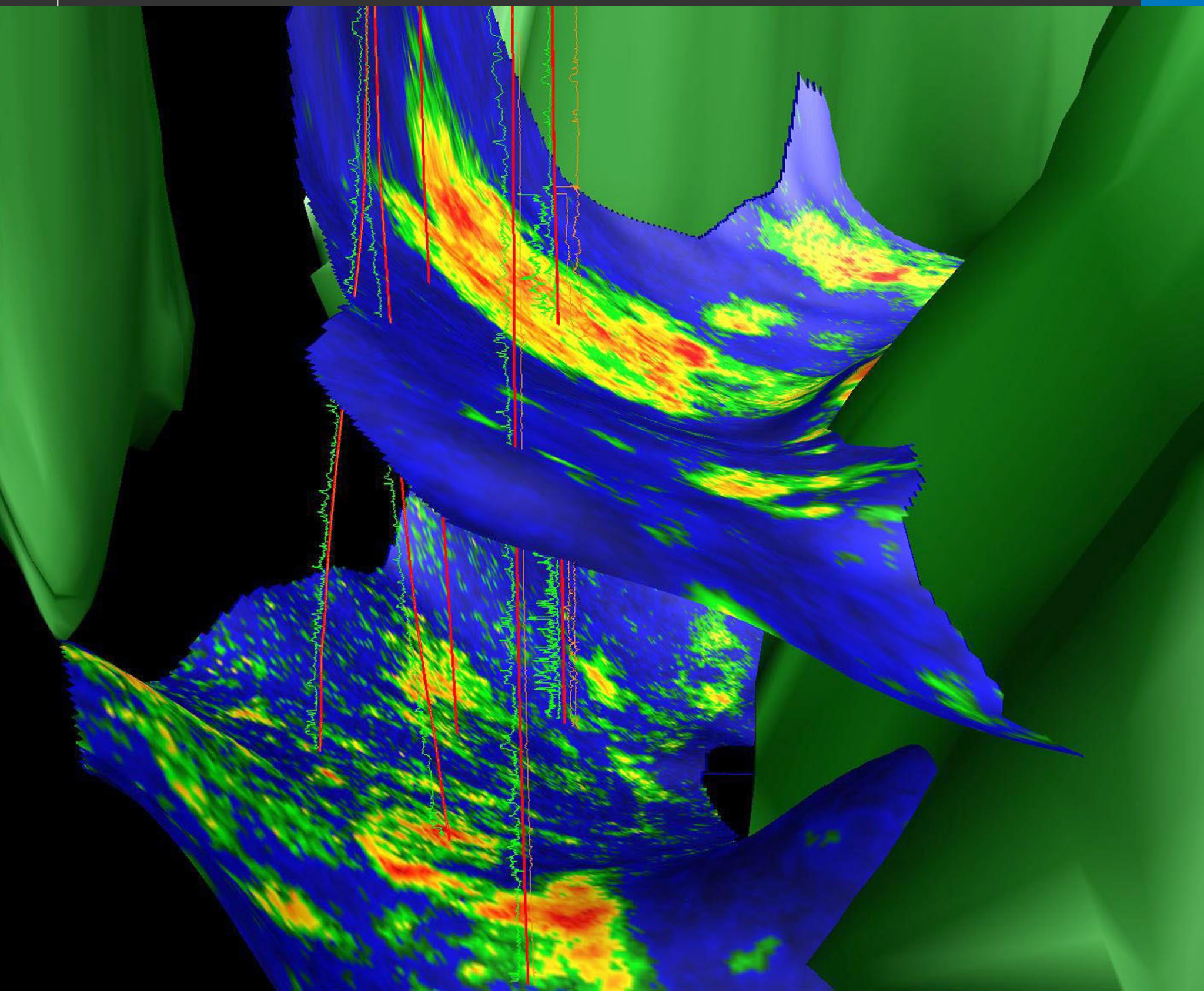
C'est en regard de cette coopération multilatérale que les partenariats le monde académique et avec l'industrie se révèlent stratégiques. Les relations qu'entretient Shell avec Intel, par exemple, lui permettent de disposer en priorité des technologies en cours de commercialisation et lui donne une vision opérationnelle de la roadmap du fondeur. En pratique, Shell a ainsi pu monter une *sandbox* de 256 CPU régulièrement mise à jour avec les Xeon les plus récents. Grâce à elle, les ingénieurs disposent d'un banc de test confortable déboguer pour les dernières itérations applicatives et optimiser à l'avance les charges de travail prioritaires.

Comme de juste, cette coopération au niveau matériel se décline au niveau logiciel. Shell utilisant les compilateurs Intel C++ et Fortran, la MKL et des optimiseurs tels que VTune, un petit groupe d'experts a été mis en place qui diffuse les meilleures pratiques de développement HPC parmi les équipes techniques. Ces efforts d'optimisation ont pu être soutenus directement par Intel, par l'entremise de correspondants techniques travaillant avec Shell aux Etats-Unis et aux Pays-Bas. Régulièrement, des réunions faisant intervenir une quarantaine d'experts Shell et Intel ont lieu pour que soient définis et planifiés les besoins et les capacités techniques des prochaines générations applicatives.

Si l'architecture globale des systèmes HPC de Shell est basée sur des CPU Xeon, cela n'empêche pas la compagnie de maintenir une veille technologique constante. Les dimensions de ratio coût/performances et de TCO global l'amènent ainsi à évaluer régulièrement des solutions alternatives telles que les accélérateurs GPU et les implémentations FPGA. Un cluster combinant CPU Xeon et GPU de dernière génération a ainsi été déployé pour tester de façon intensive les codes de visualisation en cours de développement.

Sous le fond des océans…

Sur le terrain, ces dispositions techniques impactent de façon importante les activités métier de la compagnie. L'équipe de David Schewitz, Directeur de la géophysique et de la gestion de données pour Shell Upstream Americas, en fait l'expérience quotidiennement. Lorsqu'ils partent explorer le

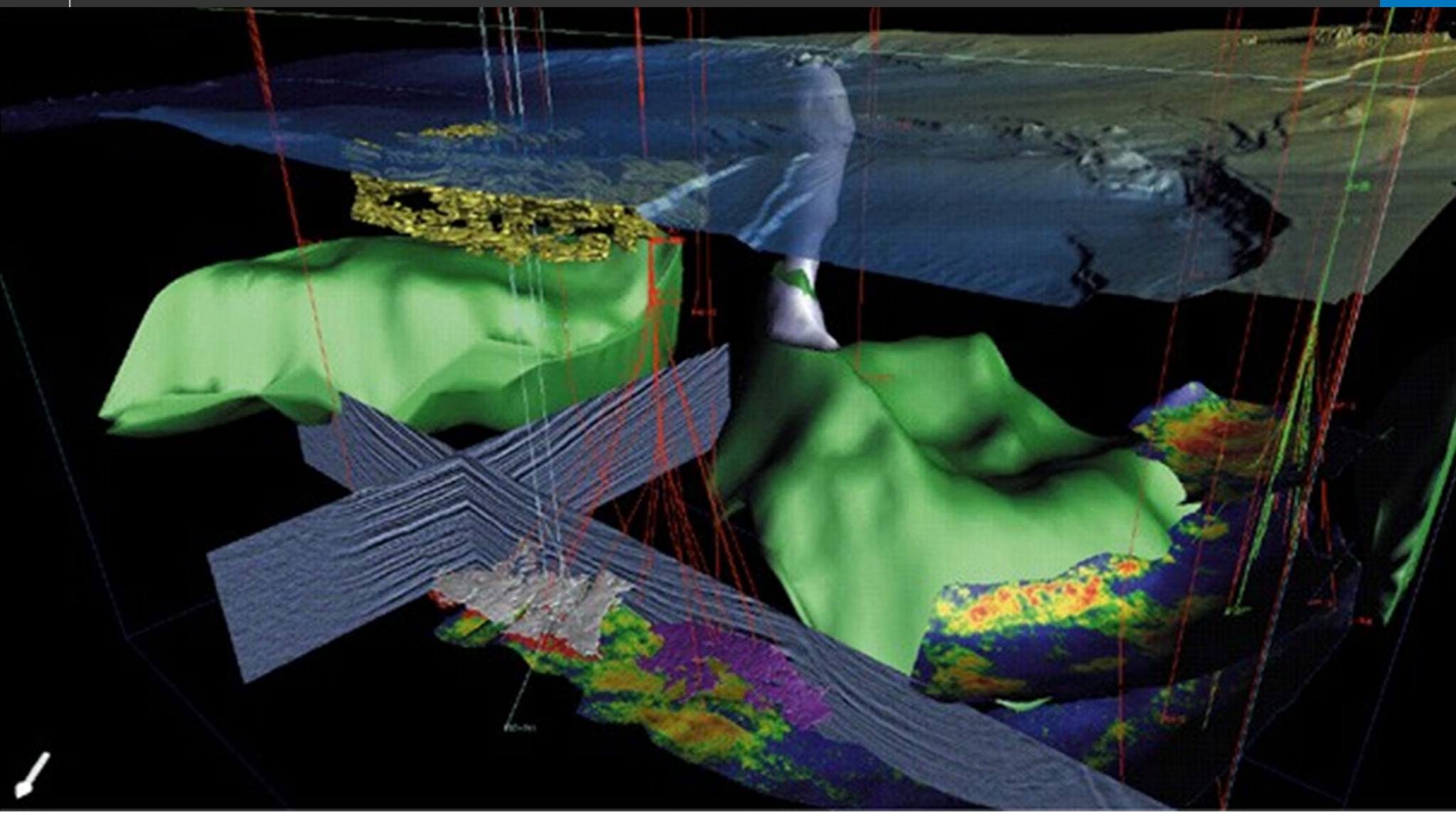


Vue sismique de la "subsurface" du champs Mars situé dans le Golfe du Mexique, au large de la Louisiane. L'image est extraite d'un des environnements de sismique virtuelle de Shell.

fond du Golfe du Mexique, ils utilisent des canons à air comprimé à pénétration profonde, grâce auxquels des impulsions contrôlées sont envoyées dans l'océan. Ces pulsations traversent la "subsurface" et réfractent sur les couches géologiques. Pour les capter, l'équipe déploie des nœuds de capteurs spécifiques sur le fond océanique et enregistre numériquement les temps de parcours et l'amplitude de l'énergie renvoyée. C'est à partir de clusters HPC qu'en sont déduites les cartographies de subsurface - avec un niveau de profondeur atteignant 16 km.

"*Dans le Golfe du Mexique*", explique David Schewitz, "*nous sommes la plupart du temps dans des prospections très profondes, sous des sels très complexes. Quand on considère que la cible, généralement assez petite, se trouve à 10 km sous la surface de la terre, il est assez prodigieux de pouvoir l'identifier et la cartographier… Sans HPC, des résultats de ce type sont impossibles à obtenir.*"

Le traitement et le stockage de données sismiques représentent pour leur part des coûts élevés, mais David Schewitz précise que ces coûts ont malgré tout été divisés par 2,5 en dix ans. "*Les économies réalisées nous permettent de traiter plus*



Modélisation et visualisation détaillées d'un dataset de "subsurface" complet. Les couleurs brillantes représentent les réservoirs de pétrole et de gaz situés sous les couches de sels. Cette visualisation représente l'état le plus abouti de l'analyse des données physiques par calcul intensif.

de données, de déployer des nouvelles technologies sismiques, de piloter simultanément plus de projets et de rendre chaque projet plus important et plus dense", ajoute-t-il.

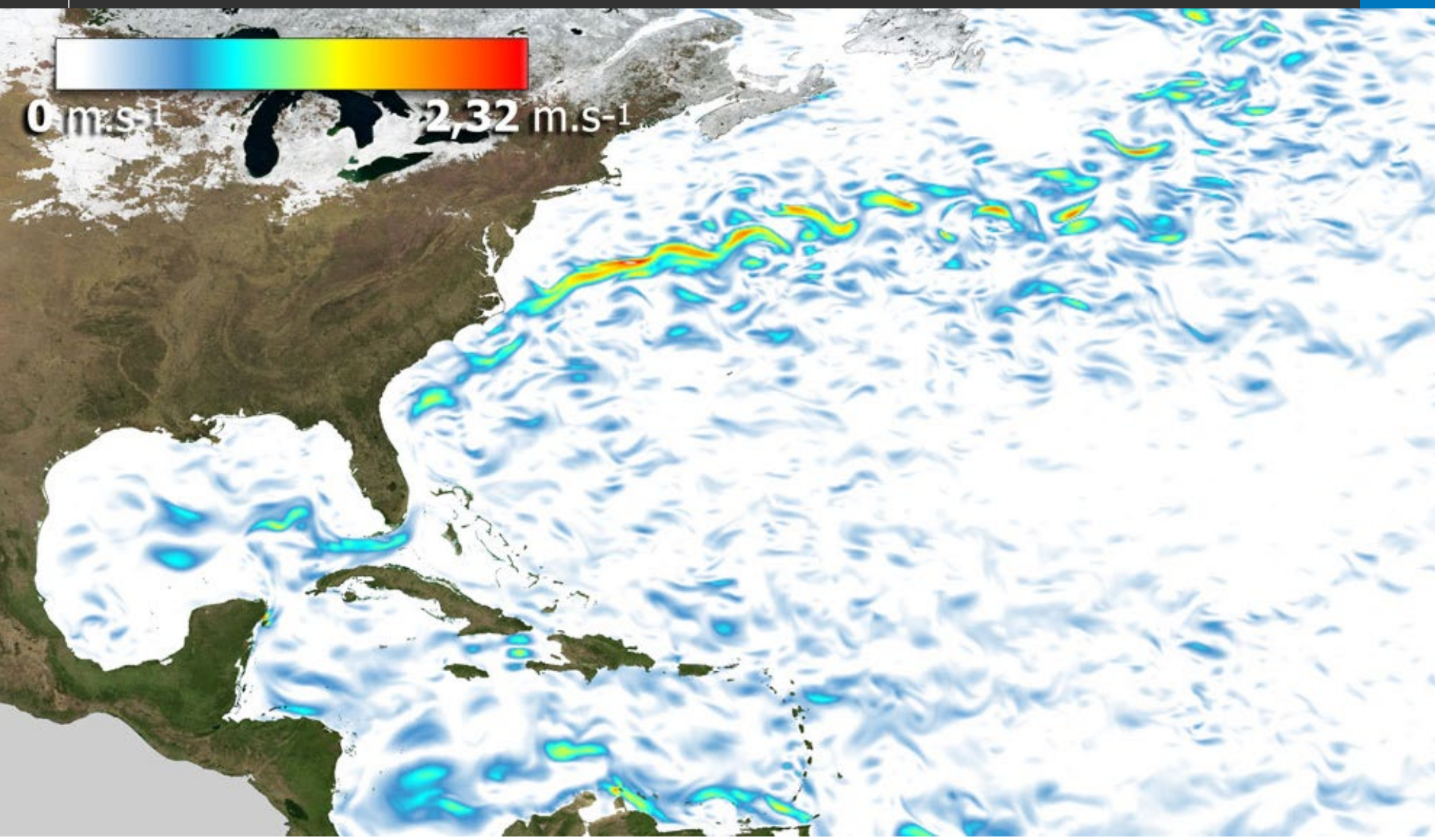
Les images sismiques plus précises, disponibles grâce à une collecte de données multiazimuts, permettent également des réductions significatives de coûts industriels, grâce notamment au meilleur positionnement des puits, à la réduction du nombre de ces puits et à des taux de récupération supérieurs. Shell a par exemple identifié d'importants gisements à proximité de ses plateformes Mars-Ursa, au Sud de la Louisiane. Mais les cartographier est difficile du fait d'une importante présence de sels. Grâce à ses programmes d'imagerie OBS (Ocean Bottom Survey) par équation d'onde anisotrope - donc grâce aux outils matériels et logiciels HPC dédiés à cette application - le risque d'erreur est sensiblement minimisé, ce qui renforce la rentabilité de l'exploitation. Au point que l'équipe de David Schewitz est aujourd'hui en train de réaliser cinq autres relevés OBS sur huit gisements différents.

Résultats concrets

Grâce à une stratégie HPC proactive, grâce à des améliorations de processus majeures, Shell améliore de façon significative la performance de ses activités à l'échelle internationale. L'équipe TaCIT a ainsi pu tripler les performances de la modélisation de réservoirs Dynamo développée en interne.

Les workflows Dynamo, quant à eux, bénéficient d'un facteur d'accélération encore plus élevé. Les efforts d'optimisation applicative ont produit des gains en efficacité de l'ordre de 2X à 12X, selon le workflow concerné, avec des augmentations comprises entre 8X à 60X si l'on prend en compte l'accès à des parcs matériels de plus grande capacité.

Enfin, les ingénieurs de technologie sismique du département IRD et de logiciel sismique de l'équipe TaCIT ont été en mesure de développer des versions hautes performances des applications métier les plus critiques de la compagnie. Avec pour conséquence un accroissement sensible de la productivité des équipes de terrain et d'analyse.



3 - Mercator Océan : 6X plus de puissance, 2X moins d'énergie

C'est presque un cas d'école. Mercator Océan, le centre français d'analyse et de prévision océaniques, est une société civile ayant pour partenaires les cinq acteurs nationaux de l'océanographie opérationnelle que sont le CNRS, l'IFREMER, l'IRD, Météo France et le SHOM (Service Hydrographique et Océanographique de la Marine). Sa mission : réaliser des prévisions tridimensionnelles de l'état physique des océans jusqu'à 14 jours. Un métier à haute valeur ajoutée, pour lequel le calcul intensif est absolument indispensable.

Tous les jours, grâce au modèle NEMO développé par le CNRS, des paramètres tels que la température, la salinité, l'état des courants de surface ou de profondeur sont recueillis sur tous les océans du globe. Il s'y ajoute des données altimétriques et des mesures issues de bouées techniques (3 000 de ces bouées circulent en mer). Les modèles utilisés sont globalement identiques à ceux de la météo. Ils intègrent des millions de données, ce qui nécessite des capacités *compute* et mémoire importantes.

Une infrastructure partagée

Afin de sécuriser au maximum les processus, toutes les chaînes opérationnelles tournent sur les ordinateurs de Météo France. Des moyens informatiques externes, tels que ceux disponibles chez Météo France, l'ECMWF ou l'IDRIS, sont par ailleurs utilisés pour la R&D. Mais Mercator Océan dispose également en propre d'un cluster de 1,23 Tflops, secondé par un système dédié au support R&D.

Basée à Toulouse, Mercator Océan fournit donc un service d'intérêt général pour la France et l'Europe. Les données océaniques produites sont utilisées par différents acteurs locaux, nationaux et internationaux dans les secteurs de l'environnement, de la sécurité maritime, de la défense, de la gestion des ressources marines et du développement durable. De plus, les programmes pilotés par Mercator Océan offrent à la communauté scientifique des

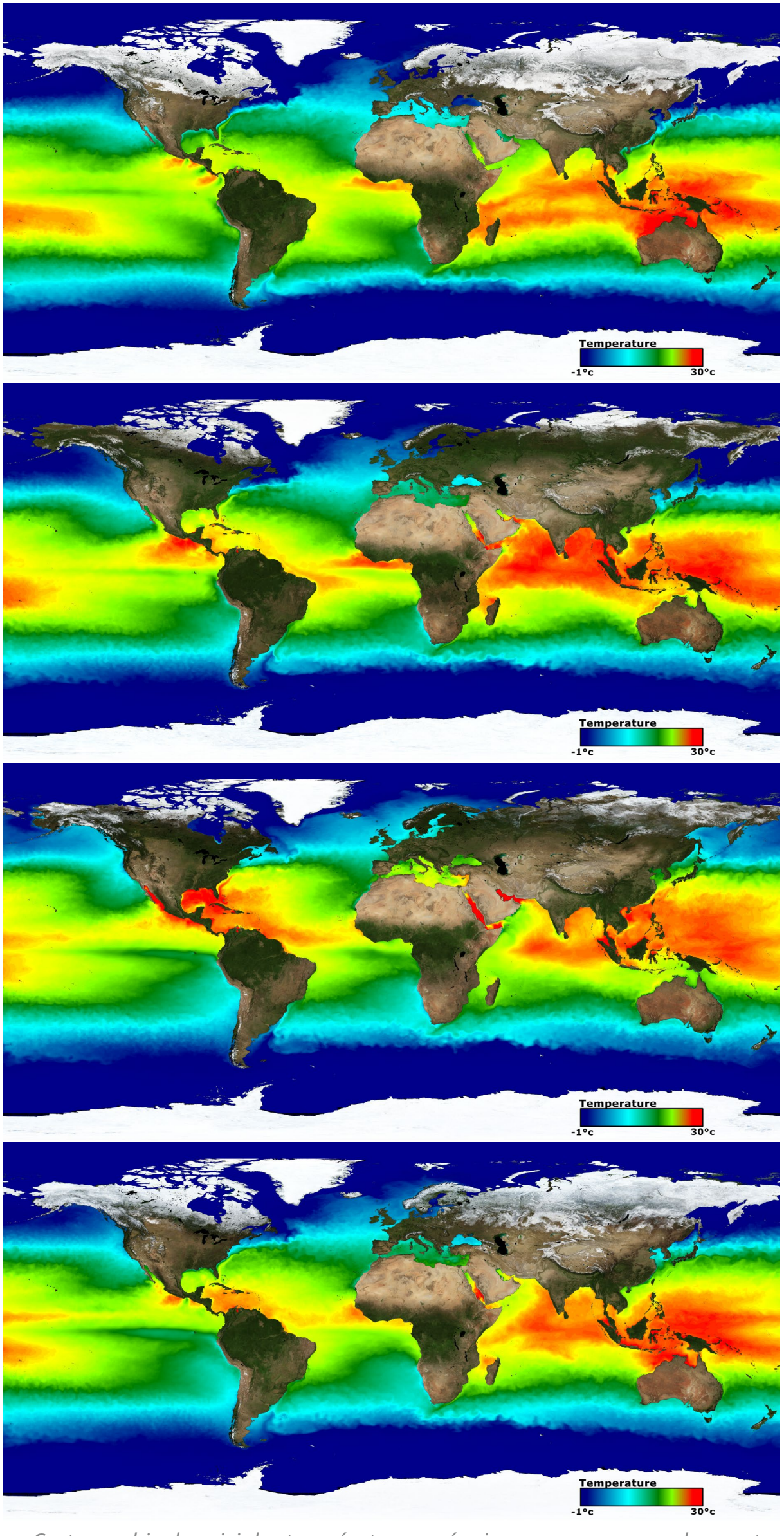
outils et des fonctions d'assistance pour la recherche océanographique. Enfin, ils contribuent au développement d'applications commerciales dans le domaine de l'océanographie et des industries qui lui sont liées.

La Commission Européenne considère l'océanographie comme l'une des trois composantes principales de son programme de surveillance de la terre et des océans (GMES – *Global Monitoring for Environment and Security*). A ce titre, afin de préparer la mise en place d'un service opérationnel au niveau européen, la Commission cofinance un projet ambitieux, MyOcean, dont Mercator Océan est chef de file depuis 2009.

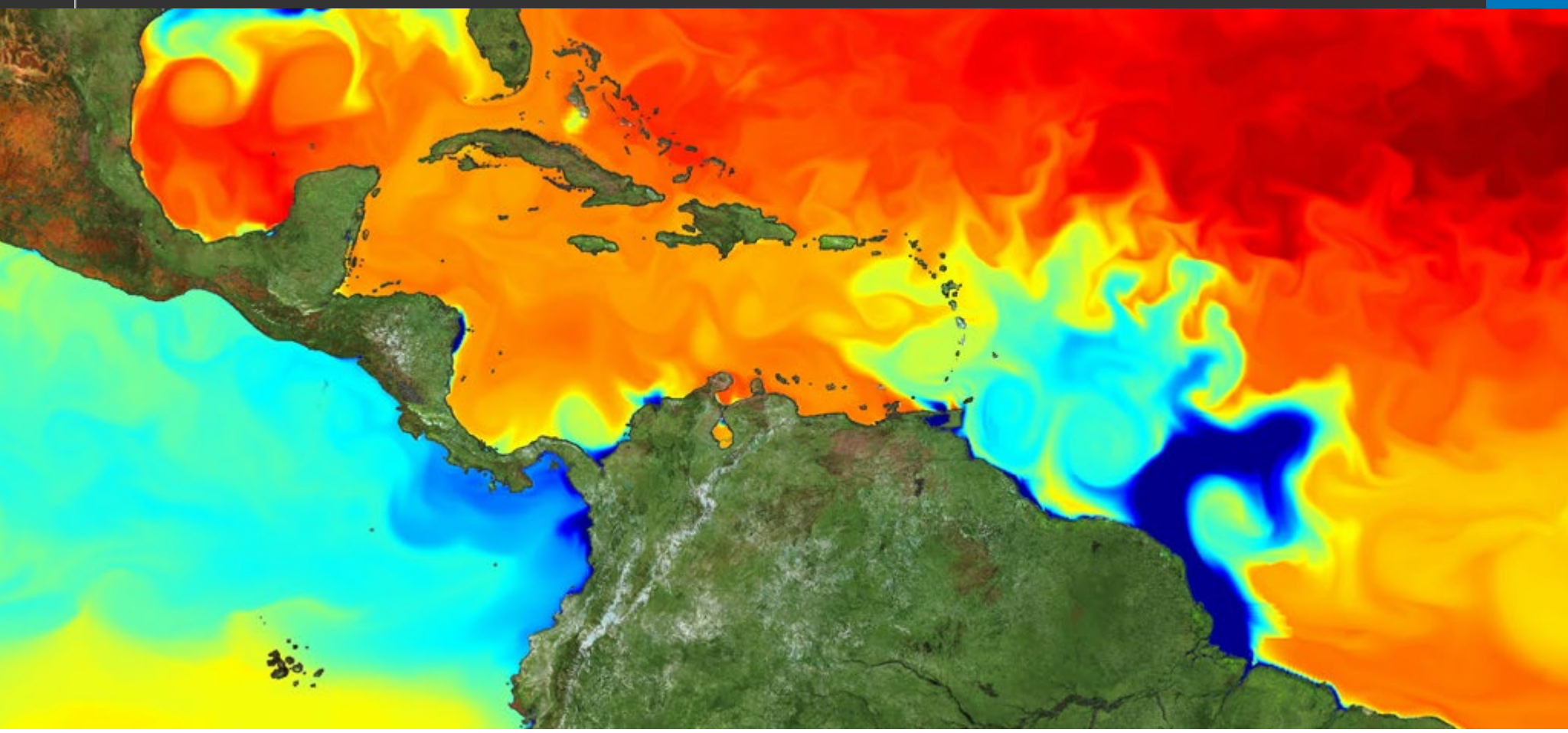
Evolution instantanée

Répondre à l'ensemble de ces besoins mobilise des modèles prévisionnels de l'état des océans qui requièrent des puissances de traitement importantes et utilisent beaucoup d'énergie. Lorsque le système de support R&D, dont le puissance ne dépassait pas 232 Gflops, est arrivé en fin de cycle, Mercator Océan a décidé de le remplacer par un cluster HPC constitué de serveurs Dell PowerEdge C6100 (4 lames bi-socket Xeon et 48 Go de RAM par unité 2U). Pour Bertrand Ferret, Responsable des systèmes d'information de Mercator Océan, "*cette mise à jour devenait nécessaire, tant en termes de performances que de réduction des coûts énergétiques.*"

Dans un tel contexte de production, une expertise "avant-



Cartographie de suivi des températures océaniques moyennes sur les quatre trimestres de 2011. Modèle global au 1/12ème.



Représentation des taux de salinité de l'Atlantique près de l'embouchure de l'Amazone. A la saison des pluies, le haut débit du fleuve les fait chuter considérablement (les zones bleues indiquent les taux les plus bas).

vente" s'imposait. Mercator Océan avait en effet besoin de tester la configuration dans son propre environnement, avec ses propres mesures. Dell lui a donc fourni un serveur d'essai puis, après validation technique, le contrat a été conclu. Deux semaines plus tard, la machine était livrée, après avoir été intégralement configurée chez le constructeur.

"*La rapidité de livraison était importante car l'augmentation de notre puissance de calcul R&D devait répondre aux besoins de nouveaux projets en cours*", précise Bertrand Ferret. "*Pour que le système entre en production, il fallait y transférer toute une section de notre suite de données. L'objectif était donc d'installer la solution en moins de 15 jours, pour limiter les interruptions de service HPC local.*" Ce sont les responsables techniques de Mercator Océan qui ont dirigé les opérations. Le support de Dell s'est ensuite déplacé sur le site de production, pour optimiser complètement certains paramètres d'exploitation.

Bénéfices concrets

Typiquement, les coûts énergétiques et les coûts de refroidissement représentent au minimum 10 % du TCO d'une infrastructure HPC. Pour les réduire, les serveurs livrés par Dell intègrent des ventilateurs à haute performance qui diminuent la consommation énergétique de la machine alors même que sa densité est élevée. A titre de comparaison, l'ancien cluster de 42U consommait environ 12 kW en fonctionnement de croisière. Son remplaçant a fait chuter cette consommation de moitié, à 6 kW, en ne mobilisant que 12U d'espace physique.

Côté performance, le facteur d'augmentation atteint pas moins de 6X. Un chiffre en ligne avec l'amélioration continue des codes de Mercator Océan et la multiplication des volumes de données à traiter. Indépendamment de la nature adaptative de ce progrès, les bénéfices en termes de productivité au quotidien sont sensibles, avec en prime une importante réduction de l'occupation d'espace au sol qui permet d'envisager sereinement différentes stratégies d'extensions prochaines.

Enfin, différents niveaux de SLA ont permis à Mercator Océan d'adapter la disponibilité de son infrastructure à la criticité de ses applications. Peu après le déploiement du nouveau cluster, un composant défectueux a quasiment détruit la carte mère d'une des lames serveurs. En quelques heures, le problème a pu être diagnostiqué et résolu sur site. Peu de temps après le retour en production, le responsable qualité du constructeur s'est déplacé pour valider définitivement l'intervention.

4 - Infiniti Red Bull Racing : Calculer pour tout gagner

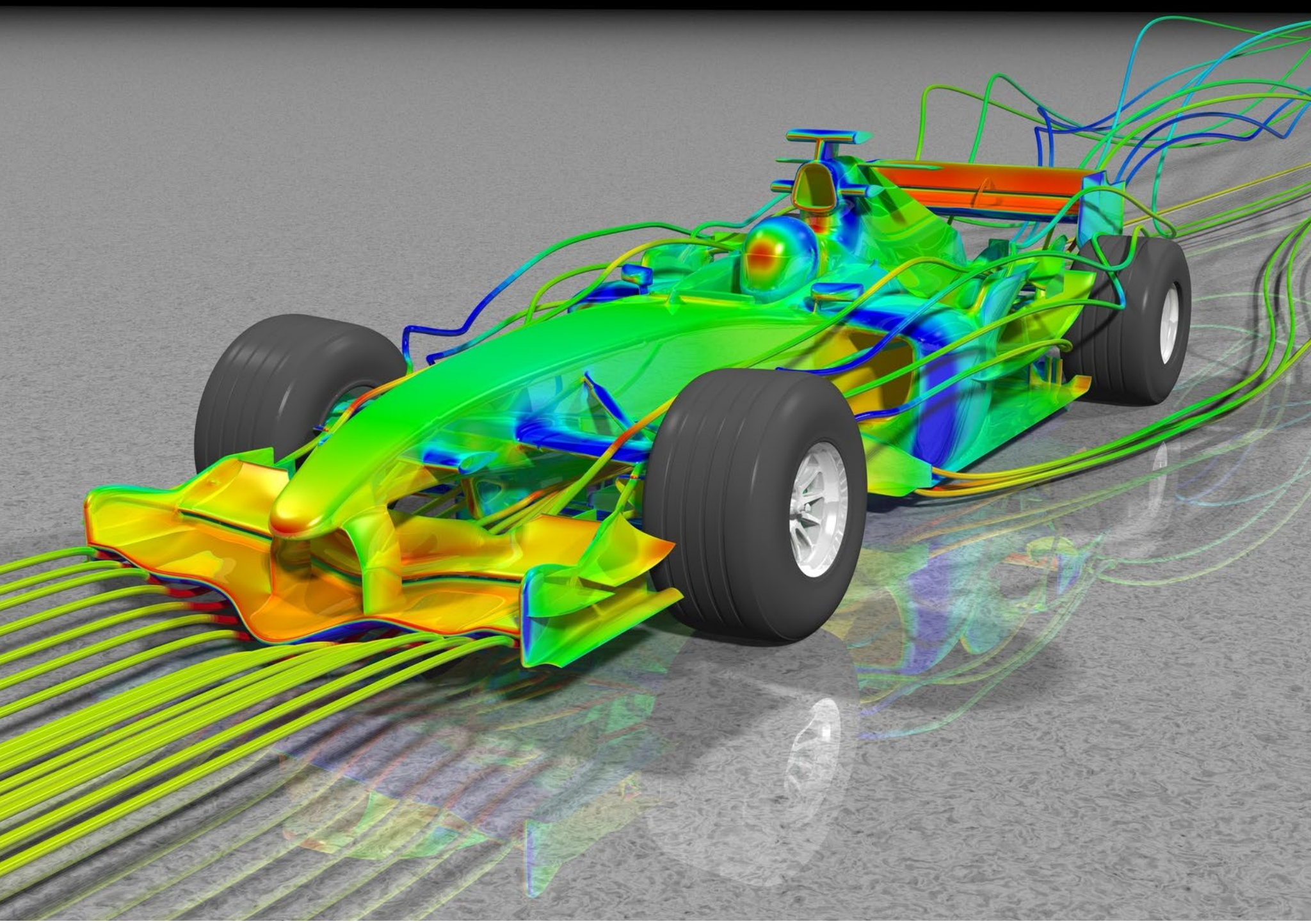
Infiniti Red Bull Racing s'est lancée en Formule 1 en 2004. Par rapport aux écuries historiques allemandes ou italiennes, cela fait d'elle une sorte de "*rookie*". Pourtant, l'ex- Team Lotus s'est rapidement imposé comme un compétiteur avec lequel il fallait compter. Ces trois dernières années, l'équipe a en effet glané trois titres de champions du monde des constructeurs et des conducteurs. Au nez et à la barbe des vétérans !

Bien que ce soit au volant que les courses se gagnent - les pilotes vous le confirmeront - construire une Formule 1 compétitive au point de remporter un Grand Prix demande des efforts continus de la part des ingénieurs, et le soutien de tous les partenaires impliqués. Des efforts en l'occurrence assez particuliers...

Toujours plus avec toujours moins

"*Nous devons faire face au défi qui consiste à développer le plus possible d'améliorations à la conception, tout en disposant de moins en moins de moyens pour le faire*". témoigne Al Peasland, Directeur des partenariats techniques d'IRBR. Il ne s'agit pas là d'un problème de ressources : "*En F1, les règlements deviennent chaque année plus restrictifs. Les régulateurs du championnat imposent par exemple des limites aux temps de test sur piste et en soufflerie. C'est une des principales raisons pour lesquelles nous utilisons systématiquement l'analyse virtuelle et la simulation, notamment dans le domaine de la CFD.*"

Ce qu'Al Peasland ne précise pas c'est que, même dans ce domaine, le règlement est contraignant. Jusqu'à la fin du championnat 2013, les équipes n'ont droit qu'à 40 Tflops d'usage en moyenne sur la saison - le contrôle s'effectuant à partir du reporting des écuries. Suite à une demande commune de la plupart des constructeurs, la limite sera portée à 80 Tflops



La CFD apporte une bien meilleure précision pour l'optimisation des voitures. Couplée aux technologies de visualisation avancées et à l'analyse haute performance (HPDA) des données de simulation, elle autorise des dizaines de tests de conception par semaine.

pour le championnat 2014, ce qui signifie que l'infrastructure HPC dont dispose chaque équipe atteint des puissances crêtes assez nettement supérieures.

Au 100ème de seconde

Quoi qu'il en soit, grâce à la CFD, Infiniti Red Bull Racing peut tester chaque innovation de design en soufflerie virtuelle, et ainsi mesurer son impact sur l'aérodynamique globale de la voiture dès les premiers cycles de conception. Ce qui est vrai dans un contexte de compétitivité de marché l'est bien plus encore dans celui de la F1, paroxysme de compétition, où chaque centième de seconde gagné fait la différence au finish. La numérisation des processus apporte plus de précision en conception, permet de tester une variété bien plus large d'idées sans entraîner de coûts supplémentaires, et cela dans des temps qui n'ont rien à voir avec ceux qu'imposent les méthodes traditionnelles.

Pour tirer le maximum de ses applications de design et d'analyse, IRBR utilise aujourd'hui un calculateur doté d'un système de fichiers parallèles GPFS IBM. Pourquoi ce choix par rapport à une architecture classique ? Pour diminuer les temps de réponse des codes tournant sur les multiples nœuds du cluster.

Dès que l'utilisation de la simulation s'est généralisée dans l'équipe, les volumes de données, donc les temps d'accès, ont eu tendance à s'envoler.

La RB9 en plein test physique d'aérodynamique. L'imposant "flow meter" situé derrière la suspension avant permet de valider les mesures, réalisées en soufflerie virtuelle, ayant servi au nouveau design des ailerons.

D'où, logiquement, une augmentation progressive des délais d'obtention des simulations et de leur analyse. Le cluster par lui-même était très largement capable de supporter une utilisation intensive. La question était plutôt : comment améliorer le flux de production du système pour que l'information soit disponible plus rapidement ?

Les responsables techniques de l'écurie ont toujours considéré l'optimisation des charges de calcul comme essentielle. Ils ont donc cherché un moyen plus efficace d'utiliser la puissance disponible en se focalisant sur l'optimisation de la planification et de l'allocation des ressources. Après avoir envisagé l'adoption d'une solution open source, l'équipe s'est tournée vers une solution commerciale, en l'occurrence IBM Platform Computing.

Grand Prix après Grand Prix, IRBR l'utilise pour gérer la charge de travail des applications HPC dans leur environnement matériel. Avec pour objectif de veiller à ce que les bonnes ressources soient automatiquement allouées aux codes qui les nécessitent, et pour contrainte de ne pas dépasser le quota de puissance autorisé.

Bénéfices concrets

Nathan Sykes, Team Leader du CFD and FEA Tools Group précise qu'avant la mise en place de ces outils, "*toutes les charges de travail étaient programmées manuellement. Nous devions donc attendre qu'un processeur ait terminé son job pour passer au suivant. Cette approche manquait d'efficacité, et s'avérait très consommatrice de ressources. Aujourd'hui, nos temps de simulation sont réduits de 20 à 30 %. Ces gains très significatifs découlent de la possibilité de concevoir automatiquement des workflows complexes et interdépendants, et de programmer cer-*

Les essais en soufflerie réelle sont à la fois coûteux, laborieux et chronophages. Ils sont l'exemple-type des bénéfices de la simulation en termes de coût et d'efficacité.

taines tâches individuellement. Une telle flexibilité nous permet de réaliser de plus en plus d'analyses, de plus en plus rapidement. Aujourd'hui, nous sommes capable de redessiner, de modéliser et de construire une nouvelle voiture en à peu près moitié moins de temps."

La solution permet à l'équipe de contrôler l'intégralité du flux des processus, jusqu'à leur exécution. Elle est utilisée en parallèle avec IBM Platform Application Center, qui autorise la soumission des charges de travail via le Web. L'ensemble forme un modèle de workflow global et assure une cohérence technique d'ensemble aux simulations lancées par les ingénieurs. Bénéfice accessoire, les licences applicatives sont également gérées par le système, qui garantit ainsi une utilisation optimale de ces ressources généralement très coûteuses.

Plus vite encore avec les analyses en temps réel

Compte tenu des premiers retours sur investissement, Infiniti Red Bull Racing a récemment décidé d'aller plus loin en déployant une plateforme d'analyse et de visualisation de données à haute performance. Objectif : valoriser son trésor de données d'analyse en identifiant de nouvelles pistes d'améliorations techniques. Pour maintenir la cohérence technique évoquée précédemment, l'équipe a choisi un produit du même fournisseur, à savoir IBM Platform Analytics.

Car la conception numérique en amont n'est qu'un aspect du travail des designers techniques. Ce qui se passe sur le circuit est tout aussi déterminant pour la victoire, d'où l'importance des prises de décision en temps réel. Choix des pneumatiques, stratégies de gestion des ressources moteurs, IRBR doit pouvoir analyser toutes les données issues des capteurs placés dans la voiture pour transmettre aux mécaniciens et aux pilotes les recommandations les plus efficaces.

"*La rapidité et le volume de nos analyses sont en augmentation constante*" explique Christian Horner, Team Principal, "*ce qui impacte très positivement la performance de nos voitures. Là où certaines simulations nécessitaient des heures, jusqu'à parfois plusieurs jours, l'ensemble du processus ne prend plus aujourd'hui que quelques minutes. Nous sommes donc capables d'utiliser les résultats directement en course, ce qui nous permet d'ajuster en permanence un nombre toujours plus élevé de paramètres, c'est-à-dire d'être réellement plus efficaces, week-end après week-end…*" ■

Call for Participation

The International Conference for High Performance Computing, Networking, Storage and Analysis

SC13
Denver, CO | 2013

Conference Dates:
Nov. 17 - 22, 2013
Exhibition Dates:
Nov. 18 - 21, 2013

Why participate in SC13?

Not only is the exhibit space larger, the technical program innovative, and the city "Mile High," you'll have the opportunity to participate in new technologies, see next year's breakthroughs, and be "dared to see Denver in a whole new way."

SC's continuing goal is to provide an informative, high-quality technical program that meets the highest academic standards, with every aspect of the SC13 technical program being rigorously peer reviewed. Moreover, we are streamlining the opportunities for participants to immerse themselves in the complex world of HPC, such as earlier acceptance deadlines for workshop program improvement and new programs to highlight innovative technologies.

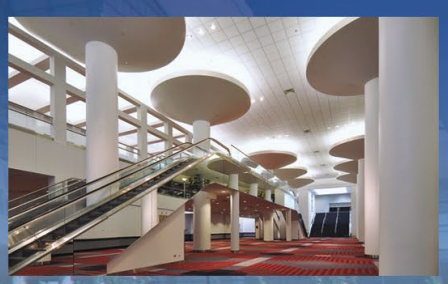
It's not the Denver you remember – this great city has countless restaurants, hotels, and more, all within walking distance of the convention center, making it all the more convenient to collaborate and connect.

Visit sc13.supercomputing.org for more information on programs and deadlines.

With HPC everywhere, why not participate?

Sponsors: IEEE computer society, acm Association for Computing Machinery

Follow us on Facebook and Twitter



Verbatim

© CADAM2013

MARC SIMON

DIRECTEUR TECHNIQUE, SGI FRANCE

A l'ordre du jour :

- Retour sur la conception technique de Pangea
- Très haute densité et efficacité énergétique
- L'adéquation entre Big Data et mémoire partagée
- La vraie valeur ajoutée d'Hadoop…
- La place du HPC dans l'IT métier

PROPOS RECUEILLIS PAR ALEX ROUSSEL

HPC Magazine : SGI et Total ont récemment défrayé la chronique HPC en annonçant la mise en production prochaine de Pangea, le calculateur privé officiellement le plus puissant au monde. Quelles étaient les spécificités et les contraintes de l'appel d'offres ?

Marc Simon : Total souhaitait augmenter ses capacités en termes de performance. Quand Total parle de performance, c'est de performance applicative sur son propre code qu'il s'agit. Le processus a donc consisté, d'une part, à ce que Total établisse des niveaux mi-

> *« PANGEA, C'EST 6 912 NŒUDS DE CALCUL, SOIT 110 592 CŒURS, ET 8 PO DE STOCKAGE SCRATCH... POUR UNE CONSOMMATION TOTALE DE 2,3 MW, REFROIDISSEMENT INCLUS.*

nimum à atteindre, et d'autre part à ce que nous obtenions ces niveaux à partir de codes applicatifs fournis. De cela, on a déduit la taille de la machine. C'est l'une des particularités les plus intéressantes de Total : ne pas travailler sur des performances crêtes théoriques mais réellement sur des codes maison. Total dispose des compétences internes pour développer ses propres applications, principalement dans les domaines de la sismique et de la simulation de réservoirs. Ce sont ces codes-là qui ont servi d'étalon de configuration. Cette approche nous a laissé une très grande latitude pour travailler sur l'optimisation et au final obtenir la configuration la plus efficace possible.

Pour vous donner quelques chiffres, Pangea, c'est 6 912 nœuds de calcul, soit 13 824 processeurs Intel Xeon Sandy Bridge, des E5 de la série 2600 cadencés à 2,6 GHz. Ca représente un total de 110 592 cœurs de calcul. En termes de stockage, on a un peu plus de 8 Po scratch attachés directement au calculateur. Côté compétition, on était en face des acteurs habituels, notamment IBM, Bull et HP - les équipes que l'on a l'habitude de rencontrer sur tous les gros appels d'offres.

Outre le niveau de performance requis, quelles étaient les contraintes liées au datacentre de Pau et à l'alimentation électrique du calculateur par lui-même ?

Parmi les autres particularités de cet appel d'offres, il y avait effectivement de fortes contraintes environnementales. La taille de la salle machine étant limitée, on devait prévoir une extension pour un doublement de la puissance de la plateforme. C'était une contrainte assez forte, de même que la puissance électrique, avec un maximum de 3 MW à respecter pour la totalité de la configuration, stockage compris. Ces contraintes fortes ont éliminé un certain nombre de candidats, car ils n'avaient pas la densité nécessaire pour pouvoir s'intégrer à la salle en question. Autre point important de l'appel d'offres : à la fourniture de l'infrastructure devait s'ajouter la capacité humaine à superviser et gérer la totalité des systèmes HPC en place à Pau. C'est une offre de service associé non négligeable, qui correspond à la mise en place d'un guichet unique pour tous les aspects HPC de Total et s'étend au-delà de Pangea - et de l'ancien système Rostand - pour couvrir également des systèmes qui ne sont pas d'origine SGI.

Pour revenir sur la phase dimensionnement de la machine et benchmarking, que vous a fourni Total pour les tests ?

Total nous a livré des codes accompagnés de vrais jeux de données, correspondants à des études réelles. Un bloc de test, composé du code et des données, représente environ 4 téraoctets. Il y avait une demi-douzaine de codes différents avec des objectifs à tenir sur chacun. L'un d'eux, le RTM (Reverse Time Migration), était prioritaire. C'est un code critique pour Total et plus globalement pour l'ensemble du secteur pétrolier. Très vite, la question s'est posée de savoir si nous testerions ces codes sur des

CPU généralistes uniquement ou si nous allions utiliser des accélérateurs, sachant que plusieurs de ces codes existaient déjà en version GPU. On avait toute liberté à ce niveau ; d'ailleurs, un de nos concurrents a répondu avec une version full-GPU. C'était le seul moyen pour eux d'être suffisamment dense pour tenir dans la salle et répondre aux exigences de performance évoquées précédemment. La salle complète peut accueillir entre 100 et 120 racks. Avec notre solution, on en remplit 48, alors que la plupart de nos concurrents étaient dans les 100. Et donc, pour 48 racks de calcul, on arrive à un total de 2,3 Pétaflops crête. Le LINPack, publié en juin est à 2,1 Pflops , ce qui montre un ratio entre pic théorique et pic LINPack soutenu assez satisfaisant.

Comment expliquez vous cette différence de densité entre votre offre et celles de vos concurrents ?

La machine a été conçue par rapport à une puissance applicative et un nombre typique de jobs à tourner par jour. On en est arrivé à une solution construite autour de notre machine ICE-X, dans une configuration cellulaire appelle M-Cell qui nous permet d'être extrêmement dense. Dans cette architecture, chaque rack de calcul fournit 48 Tflops. En gros, on réunit dans un rack deux fois la densité que l'on peut obtenir d'un rack traditionnel. Il ne s'agit pas de racks standards, car notre système M-Cell a pour particularité de confiner l'air qui circule à l'intérieur des armoires dans une cellule complètement contenue, en ne rejetant aucun air chaud dans le datacentre. Grâce à notre système de refroidissement liquide, on peut fonctionner avec des températures d'échange relativement élevées. Total a voulu des températures de fonctionnement situées entre 25° et 30°, ce qui permet de s'appuyer sur des chaînes de refroidissement beaucoup plus efficaces que ce qui était déjà en place. Pour info, les calculateurs SGI ICE fonctionnent avec des températures d'eau de l'ordre d'une dizaine de degrés. Ca coûte beaucoup moins cher en énergie et, au final, la consommation de la solution complète, stockage inclus, est inférieure à 2,3 MW, là où Total nous laissait aller jusqu'à 3 MW. Or, comme vous le savez, l'efficacité du refroidissement est un aspect important de la consommation globale…

> *POUR LES BESOINS D'ANALYSES DE GRAPHES, QUI CONDITIONNENT LA DÉCOUVERTE DE CONNAISSANCES, RIEN NE VAUT LES ARCHITECTURES À MÉMOIRE PARTAGÉE OÙ LA TOTALITÉ DES DONNÉES EST MONTÉE EN MÉMOIRE...*

Vous n'avez donc eu recours ni aux accélérateurs Intel ni aux GPU dans la configuration actuelle. C'est quelque chose qui pourrait évoluer sur les prochains upgrades prévus à court ou moyen terme ?

Ni les accélérateurs x86, ni les accélérateurs GPU n'ont été retenus pour cette première phase, mais il n'est pas dit que ce soit encore le cas sur les prochaines. Ca fait partie des études en cours sur l'évolution du calculateur à moyen terme, sachant que les prochaines configurations ne sont pas encore gelées. On a une première évolution prévue d'ici 12-18 mois et, en fonction des technologies disponibles, on proposera telle ou telle solution.

Seront évalués les CPU du moment, les accélérateurs Intel et les GPU NVIDIA. Ce sont des projets d'étude qui sont menés avec les supports des trois fournisseurs concernés. On connaît les roadmaps de chacun, on connaît les targets, mais Total s'intéresse surtout à la performance applicative. Or, il est difficile de parier aujourd'hui sur la performance applicative que l'on pourra obtenir avec les prochaines générations de composants. Donc, pour l'instant, tout reste ouvert. SGI a la particularité de pouvoir proposer un mix de ces technologies au sein d'une même machine ICE X. Notre meilleur exemple sur ce plan, c'est notre client historique : la NASA. Ils ont fait l'acquisition d'une première version d'ICE, en 2007. C'était le début de la machine Pleiades et, depuis 2007, ils upgradent cette machine 2 fois par an. On est justement en train d'installer chez eux la toute dernière version ICE X. C'est la sixième génération qu'ils voient arriver dans leur datacenter - une machine à laquelle on peut ajouter de la puissance de calcul sans arrêter la production en cours.

SGI est aujourd'hui OEM Intel sur les processeurs, au même titre que Bull, Cray ou même des constructeurs généralistes comme Dell et HP. Quels sont vos principaux différenciateurs ?

Pour Intel, on est epsilonesques comparés à HP, IBM ou Dell. On ne fait pas du tout les mêmes volumes qu'eux. Ce qui nous pose très clairement problème quand on est face à ces grands qui ont des coûts d'acquisition très différents des nôtres, mais on se bat avec nos armes. C'est-à-dire principalement notre savoir-faire. Aujourd'hui, dans le monde HPC et plus largement Technical Computing, si vous ne répondez pas avec du x86, principalement de l'Intel, du Linux et de l'interconnect standard, se placer est très difficile. La différenciation ne se fait pas sur les composants de base, elle se fait sur l'efficacité de leur intégration à grande échelle.

S'agissant de performances pures, on est toujours parmi les plus efficaces, y compris sur de très grosses configurations, comme vous pouvez le voir régulièrement dans le TOP500. On détient aussi quelques records comme le SPEC MPI2007 - un benchmark standard sur lequel l'ICE est seul à offrir ce niveau de compétitivité. En termes d'efficacité énergétique, les deux machines Total (car Pangea a une petite sœur à Houston, Texas) sont en tête des calculateurs généralistes du Green500. Cela signifie que des machines utilisant du Xeon standard et du Linux standard peuvent être très, très efficaces énergétiquement. Il y a aussi notre savoir-faire dans l'intégration des systèmes de cooling innovants. Quand nos clients font évoluer les machines, la facture énergétique devient critique pour eux.

Mais à processeur égal, comment atteindre une densité plus élevée ?

On fait pas mal de coengineering avec Intel sur notre gamme UV. C'est une plateforme unique sur le marché : elle utilise des composants standards Intel, mais pousse leur scalabilité, c'est-à-dire le nombre de composants que l'on est capable de mettre au sein d'une seule et même machine, bien plus haut que le design standard Intel. Aujourd'hui, avec Sandy Bridge, le modèle permet de monter jusqu'à 4 sockets. Nous sommes les seuls à pouvoir fournir du Sandy Bridge jusqu'à 256 sockets. On développe nos composants depuis 1996, ce qui a fait de nous les premiers à proposer des calculateurs CC-NUMA, c'est-à-dire des architectures à mémoire partagée, pour un large nombre de processeurs. A l'époque, cette architecture était 100 % propriétaire en termes d'interconnect, d'OS, de processeurs MIPS. Aujourd'hui, c'est beaucoup plus ouvert, avec du CPU Xeon standard, du Linux standard SUSE ou Red Hat, et de l'interconnect où, si le protocole est propriétaire, les cartes physiques sont des InfiniBand.

Votre catalogue propose des configurations à base d'accélérateurs x86 ou GPU, au choix du client. Pourquoi cet œcuménisme ?

On n'a pas à dire que l'un est meilleur que l'autre. C'est une hérésie. Il y a des domaines dans lesquels on préférera proposer du GPU, d'autres domaines où on préférera proposer du Phi. On se doit d'être, de ce côté-là, le plus ouvert possible par rapport aux attentes des clients. La machine de la NASA est un exemple, celle de Total en est un autre, mais toutes nos configurations présentes au TOP500 font de la "vraie science". Regardez, on vient d'arrêter la première machine pétaflopique, Roadrunner, qui était installée au laboratoire américain de Los Alamos. Cette machine IBM était construite à base d'accélérateurs Cell (c'est d'ailleurs à peu près la seule). Roadrunner a été conçu pour être la première machine pétaflopique, point ! Les efforts d'optimisation sur les codes étaient tels que ça n'a pas servi à grand-chose derrière. Chez SGI, ce n'est pas l'optique. Roadrunner a été arrêté et démantelé alors la machine de la NASA date de la même époque et continue d'être upgradée. Elle est toujours plus productive et constitue pour les opérateurs un environnement qui s'améliore au fil du temps, dans une réelle continuité.

Quelles sont les dates clés de la roadmap SGI dans les mois à venir ?

Nous avons trois grandes gammes de serveurs : la gamme UV à mémoire partagée, les clusters ICE X orientés HPC et calculs parallèles, et la série

> *SI LES CONSTRUCTEURS AUTOMOBILES FRANÇAIS VEULENT POUVOIR SE COMPARER AUX CONSTRUCTEURS ALLEMANDS, S'ILS VEULENT CHANGER DE SEGMENT, PEUT-ÊTRE VONT-ILS DEVOIR CHANGER D'APPROCHE AU NIVEAU DÉVELOPPEMENT PRODUIT…*

Rackable, plus orientée Web et Big Data. Sur cette dernière, nous allons faire une annonce importante le mois prochain, avec la création d'une nouvelle gamme basée sur le succès rencontré depuis 18 mois. En fait, les serveurs Rackable étaient produits principalement pour trois gros clients américains. Très bientôt, ils vont former une gamme complète, qui va nous permettre de proposer une offre complémentaire et différenciée dans ces domaines en forte croissance.

Quels sont ces trois clients ?

Ce n'est pas un secret : pour le Web, on travaille beaucoup avec Amazon, Microsoft et Yahoo. Côté Big Data, SGI est pas mal implanté dans certaines agences gouvernementales américaines. Je ne peux pas donner beaucoup de détails, pour des raisons que vous comprendrez, mais ce sont des gammes de machines qui sont en production chez nous depuis plus d'un an et qui correspondent à des besoins que l'on voit aujourd'hui croître un peu partout, surtout sur le marché Technical Computing élargi, qui intègre désormais certains besoins de type Big Data.

Le Big Data est-il déjà un marché significatif pour SGI ? Quelle est la pertinence de la marque sur ce marché où l'on a encore souvent recours à du serveur banalisé ?

Si on regarde aujourd'hui ce qui est fait en Europe sur Hadoop, c'est extrêmement light : quelques dizaines de serveurs, des *proofs of concept*, des phases 0 ou 1 de projets. C'est du serveur et du cluster lambda. L'expérience que l'on a aux Etats-Unis, notamment au niveau gouvernemental, c'est des déploiements de plusieurs milliers de nœuds, qui nécessitent une infrastructure matérielle, une intégration réseau et une infrastructure logicielle différentes.

Ce que l'on voit avec les architectures Big Data de type Hadoop, c'est une adéquation parfaite avec le monde HPC traditionnel. Les infrastructures qu'on développe depuis des années pour le HPC vont correspondre de plus en plus au monde Big Data avec l'augmentation des volumes de données et des tailles de projets analytiques en mode distribué. L'autre aspect du Big Data sur lequel on a une vraie valeur ajoutée, c'est la performance. Le terme "Big Data" ne recouvre pas uniquement Hadoop et le traitement de larges volumétries. Il y a aussi un problème de rapidité, avec des données filtrées en temps réel, des flux Internet, satellitaires, etc. De plus en plus de clients ont aussi des besoins d'analyses de graphes, qui conditionnent la découverte de connaissances. Pour cela, rien ne vaut les architectures à mémoire partagée où la totalité des données est montée en mémoire. Là, on peut mettre en œuvre des processus très simples et extrêmement efficaces. C'est une fois les éléments remarquables identifiés que l'on peut traiter les données via Hadoop.

On est vraiment à la croisée des chemins en termes d'architectures, d'infrastructures et d'évolution de ce marché Big Data qui est encore essentiellement un marché logiciel. Sur le nombre de nœuds Hadoop installés dans le monde aujourd'hui, selon un cabinet américain, SGI détient 8 % du total alors que sur le marché des serveurs généralistes, on est plutôt dans les dixièmes de pourcents… Très clairement, les déploiements Hadoop à

grande échelle tels qu'on va les voir en Europe nous permettrons de pérenniser nos infrastructures.

Quels sont les marchés les plus dynamiques au niveau mondial pour SGI ?

L'académique et la recherche restent un axe important, mais le marché le plus dynamique est celui des sciences de la vie. Beaucoup de choses se passent dans ce domaine - et cela pas uniquement dans le domaine du calcul. Le stockage, également, est concerné. C'est un marché très dynamique pour nous, d'autant que notre UV 2000 répond bien aux contraintes de la plupart des appels d'offre. Aux Etats-Unis, le Big Data est quelque chose qui est bien amorcé…

IDC montre que si le marché des grands calculateur est très dynamique avec plus de 25 % de croissance, c'est loin d'être le cas des calculateurs de taille intermédiaire. Est-ce également le cas pour SGI ?

Oui, c'est quelque chose qu'on observe depuis quelque temps déjà. Deux phénomènes peuvent l'expliquer. Pour prendre l'exemple de la France, il y a le fait que beaucoup d'investissement ont été faits au niveau des grands centres nationaux, et qu'il faut utiliser ces machines-là en priorité. Un certain nombre de laboratoires sont encouragés à ne plus faire d'acquisitions locales et à se tourner vers les centres de type T1. C'est le cas au sein du CEA par exemple. Toutes les entités sont invitées à exploiter les ressources centrales même si ce n'est ni plus simple ni plus performant. Quand vous êtes à l'autre bout de la France et qu'il vous faut travailler sur des téraoctets de données, le temps de calcul devient négligeable par rapports aux temps de transfert réseau. Autre phénomène, pour beaucoup d'industriels français, les moyens de calculs sont considérés au même titre que les moyens IT généraux. Or, ce qu'on demande aujourd'hui à un DSI, c'est de réduire les coûts.

Quand je discute avec mes collègues ingénieurs automobiles américains ou allemands, chez qui le phénomène est encore plus flagrant, je vois que, pour eux, le calcul, la modélisation et la simulation font partie de la branche produit, du métier. Pour produire une nouvelle plateforme automobile, ils ont besoin de simulation et la simulation est chez eux. Ils considèrent que l'investissement fait sur ces moyens de calcul est quelque chose qui va donner plus de valeur à leur produit. Cet investissement n'est donc pas géré par l'IT traditionnelle. C'est une autre approche, qui explique peut-être les différences que l'on constate entre les automobiles françaises et allemandes. Si les constructeurs français veulent pouvoir se comparer aux constructeurs allemands, s'ils veulent changer de segment, peut-être vont-ils devoir changer d'approche au niveau développement produit… Aujourd'hui, leurs investissements HPC ont pris des ordres de grandeurs de retard par rapport à ce que font certains constructeurs européens.

Le premier supercalculateur exaflopique sera-t-il estampillé SGI ?

C'est difficile à dire, mais ce n'est pas quelque chose que l'on cible spécifiquement. Comme je vous l'ai dit, on cherche plutôt à fournir des centaines de machines à 100 Pétaflops qu'une seule machine exaflopique. Cela correspond plus à une utilisation en production avec des impératifs en retour sur investissement. On sait très bien qu'avec le budget de consommation électrique d'une année du premier calculateur exa, vous pourrez acquérir un calculateur de 10 Pétaflops… Cela ne veut pas dire qu'on n'est pas impliqués, au contraire. SGI contribue à beaucoup de projets de recherche. On est actuellement dans des appels d'offres des grands labos américains pour des machines pré-exaflopiques, des machines aux alentours de 300 Pflops prévues pour 2016. Des configurations qui nous rapprocheront réellement de l'exascale, dont la plupart des roadmaps montrent qu'on y atteindra d'ici la fin de la décennie… ■

NVIDIA Quadro K6000,

le GPU le plus rapide et le plus avancé jamais conçu

Plus d'informations sur produits, visitez www.pny.eu/quadro
et le configurateur PNY http://pny.quadrok-selector.com/

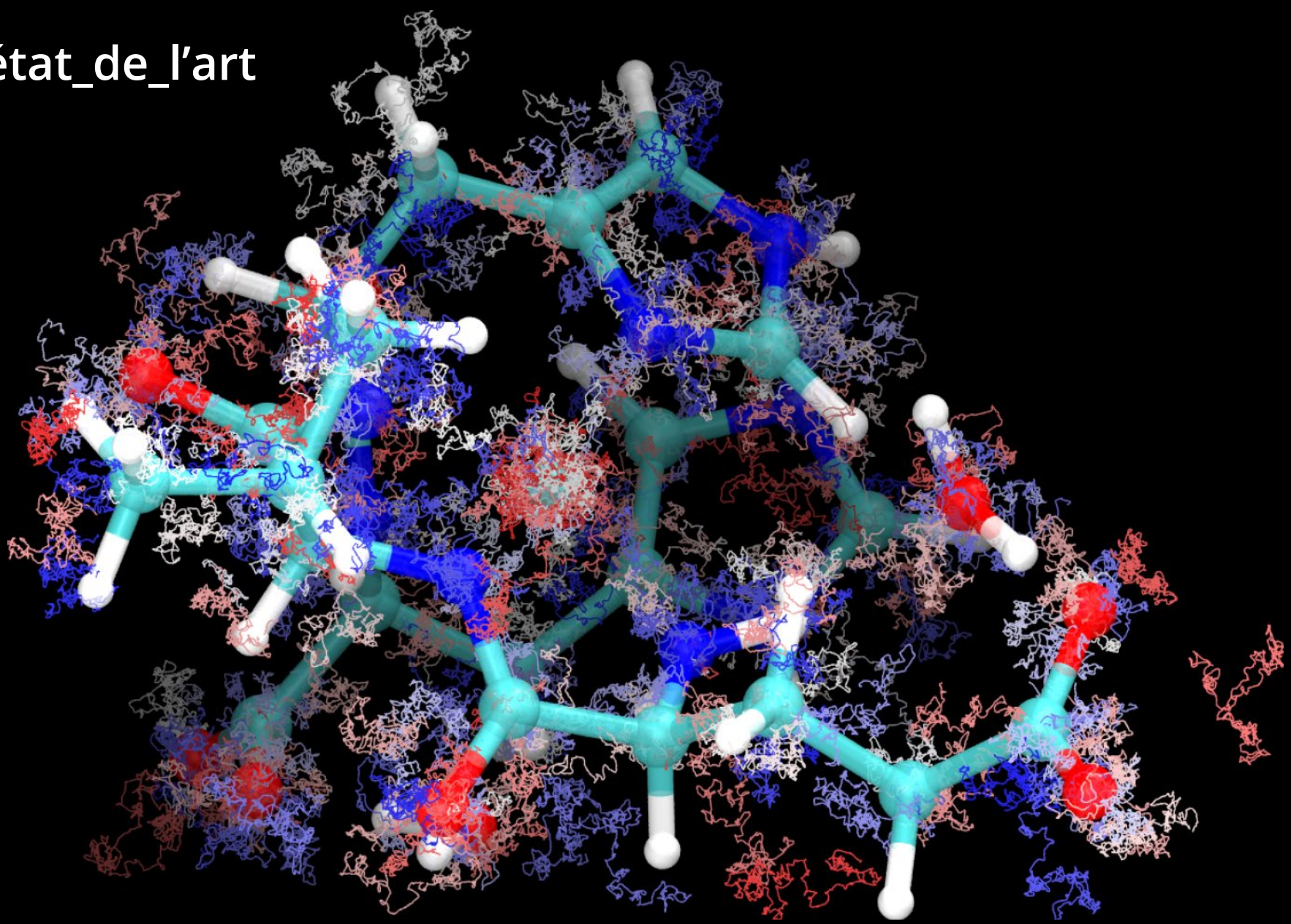
Suivez-nous: @PNYProFR

PNY®

© 2013 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA Quadro, Tesla, and CUDA are trademarks and/or registered trademarks of NVIDIA Corporation. All company and product names are trademarks or registered trademarks of the respective owners with which they are associated.

NVIDIA.

/l'état_de_l'art



SIMULATIONS EN CHIMIE :

LES BENEFICES DES METHODES MONTE-CARLO QUANTIQUE

La question ne fait maintenant plus débat : le passage à l'échelle commande de nouvelles stratégies algorithmiques. En chimie, une approche originale - la méthode Monte-Carlo quantique ou QMC - exploite massivement le parallélisme intrinsèque des méthodes probabilistes. Une implémentation judicieuse peut même y ajouter une vraie résilience aux pannes matérielles…

MICHEL CAFFAREL*
ANTHONY SCEMAMA*

L'apparition récente des calculateurs massivement parallèles donne naissance à un nouveau paradigme et ouvre une voie nouvelle à la simulation numérique. Plutôt que de s'efforcer à paralléliser toujours plus des algorithmes qui ne s'y prêtent pas toujours facilement, il peut être profitable de s'écarter des approches classiques pour se tourner vers d'autres méthodes. Des méthodes intrinsèquement peu efficaces sur des ordinateurs à quelques milliers de processeurs mais dont la structure algorithmique permet d'exploiter sans difficulté un nombre de processeurs potentiellement très large.

On sait que sur une architecture monoprocesseur, l'algorithme optimal est celui qui permet de calculer une quantité donnée en un nombre d'opérations élémentaires le

* CNRS - Laboratoire de Chimie et Physique Quantiques, Univ. Paul Sabatier, Toulouse.

plus petit possible (par souci de simplicité on ignore ici d'éventuelles contraintes liées aux I/O et/ou à la mémoire centrale). Le temps de restitution, c'est à dire le temps qu'attend l'utilisateur pour obtenir le résultat souhaité, et le temps d'exécution sont alors essentiellement identiques et proportionnels au nombre d'opérations à effectuer.

Dans une architecture parallèle, on sait que la notion de nombre d'opérations à effectuer devient secondaire, l'objectif étant de réduire le temps de restitution grâce au calcul simultané, quitte à effectuer au total un bien plus grand nombre d'opérations élémentaires. Quand le temps de restitution peut être rendu inversement proportionnel au nombre de cœurs de calcul utilisés, on est alors dans une situation de parallélisme optimal. Quand ce régime peut être étendu à des nombres arbitrairement grands de processeurs, on peut alors parler de méthodes qui "passent à l'échelle".

Avec le développement des supercalculateurs au nombre de processeurs toujours plus élevé, disposer d'un algorithme qui passe à l'échelle devient donc véritablement crucial. Un tel algorithme garantit qu'il existera toujours un nombre minimal de processeurs pour lequel il sera supérieur à tout autre qui ne passe pas à l'échelle, et cela quelles que soient les performances intrinsèques (monoprocesseur) de ce dernier. C'est ce à quoi travaillent bon nombre de nos collègues scientifiques aujourd'hui.

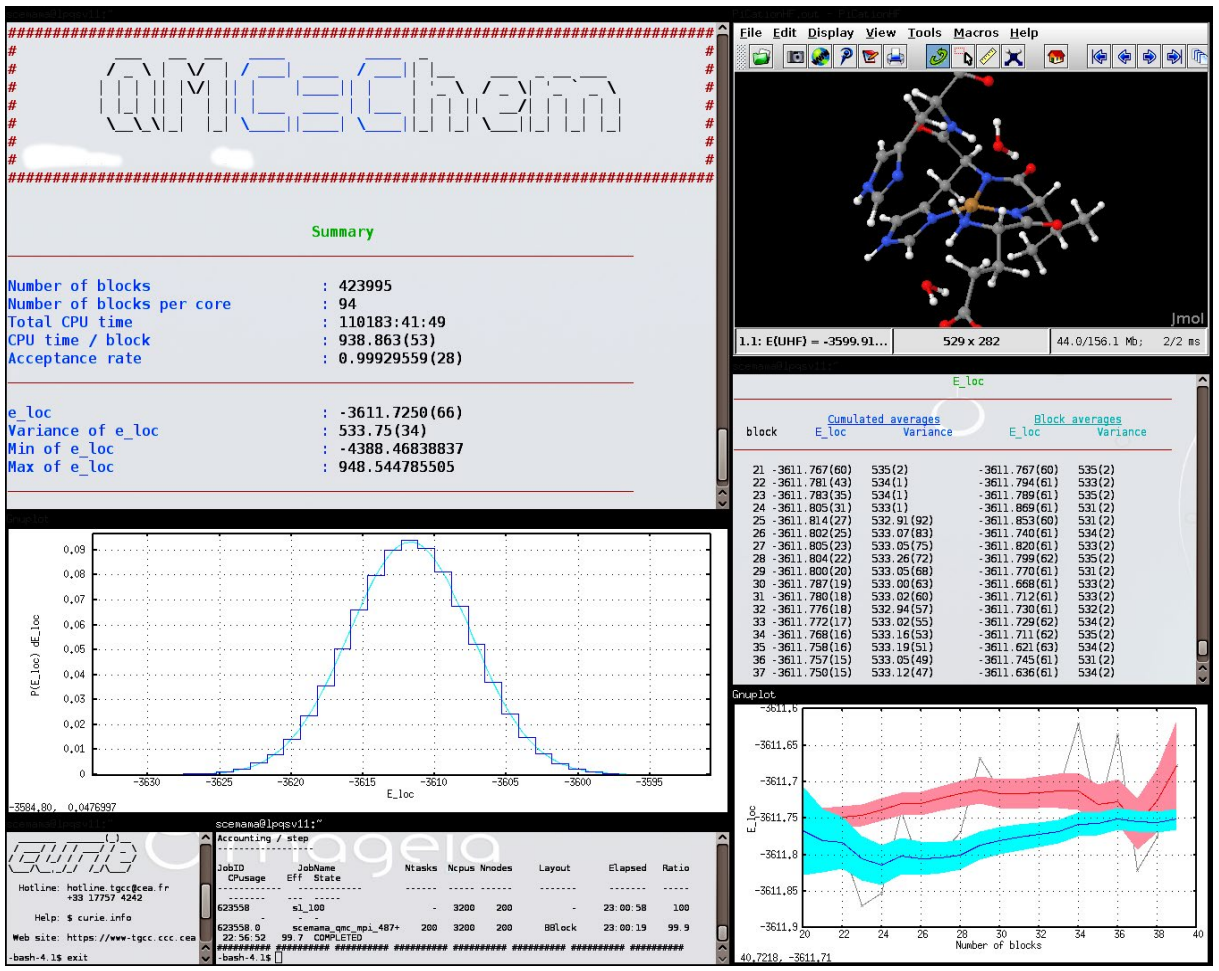


Fig. 1 - Une simulation QMC type.

Chimie virtuelle

Cette problématique s'illustre parfaitement dans son application à la chimie, discipline extrêmement gourmande en simulations numériques. La chimie est au cœur de notre vie quotidienne : développement des médicaments (*drug design*), nouveaux matériaux (nanosciences), énergies nouvelles, etc. Être capable de comprendre, de prédire et d'innover dans ce domaine est un enjeu de société considérable.

En parallèle à la chimie traditionnelle en laboratoire se développe une véritable chimie virtuelle où les processus sont simulés sur ordinateur à partir des équations microscopiques de la matière. Avec cette nouvelle "chimie sur ordinateur", les scientifiques cherchent à reconstituer le plus fidèlement possible les échanges électroniques complexes à l'origine des liaisons chimiques, des interactions entre atomes et finalement des diverses propriétés chimiques recherchées.

Ce problème constitue un formidable défi mathématique et informatique puisqu'il met en jeu la célèbre équation de Schrödinger de la mécanique quantique dont la solution recherchée - la fonction d'onde - est une fonction particulièrement complexe de l'ensemble des positions de tous les électrons (des milliers, voire plus !) et des noyaux des atomes.

Durant ces cinquante dernières années, et toujours en relation très étroite avec le développement des caractéristiques matérielles et logicielles des ordinateurs, plusieurs méthodes ont émergé. D'un point de vue

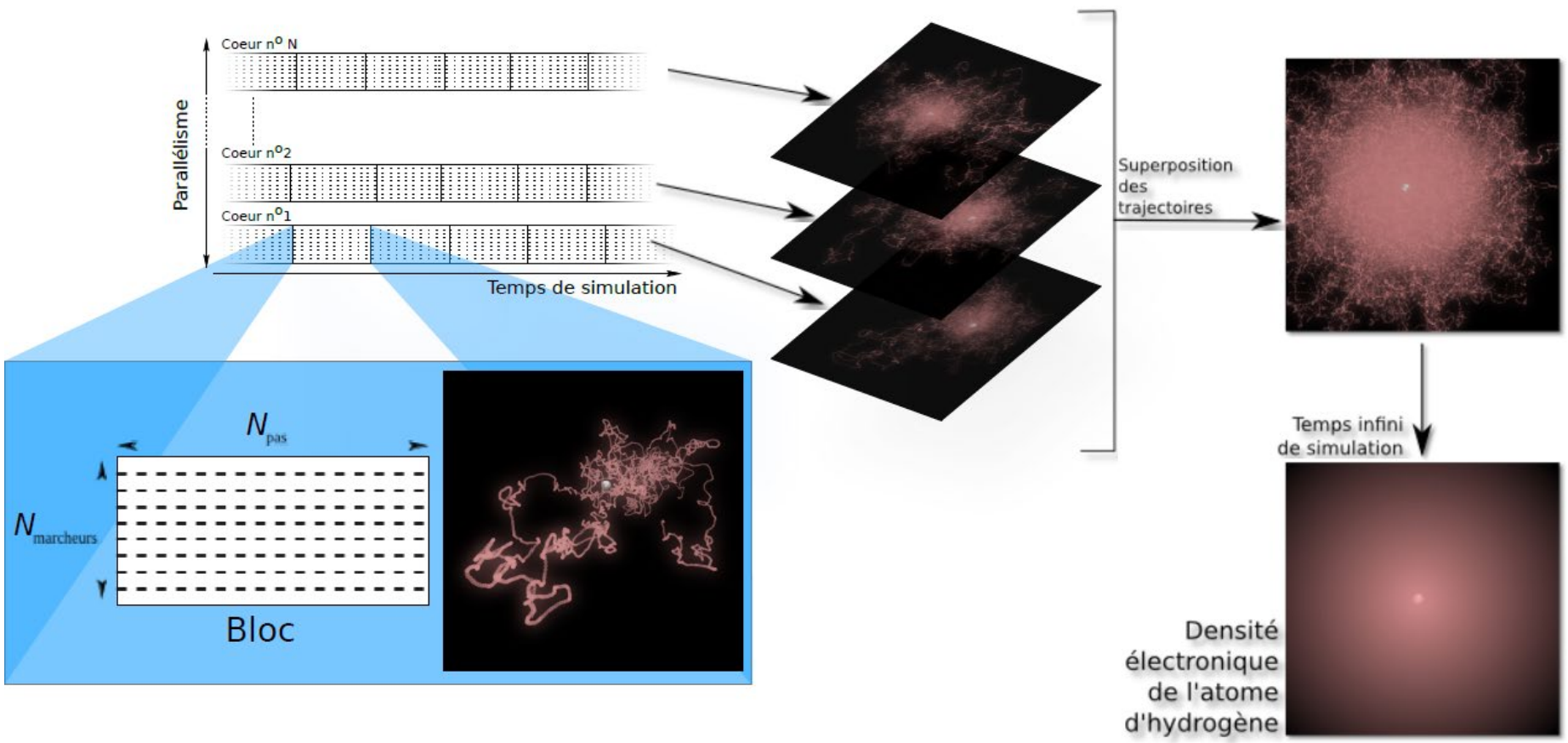


Fig. 2 - Application de la méthode QMC à la simulation de l'atome d'hydrogène.

algorithmique, elles s'appuient principalement sur des schémas itératifs de résolution de systèmes linéaires de très grandes tailles nécessitant des volumes de calculs et des besoins en I/O et/ou en mémoire centrale considérables. Malheureusement, parce qu'elles sont basées sur la manipulation et le traitement de très grandes matrices, ces approches ne se prêtent pas facilement au calcul massivement parallèle.

Nous développons dans notre groupe une approche alternative pour la chimie, très différente des approches usuelles mais qui passe à l'échelle naturellement. Basée sur une interprétation originale des probabilités de la mécanique quantique, cette approche connue sous le nom de "méthode Monte-Carlo quantique" (quantum Monte Carlo ou QMC) propose de simuler le monde quantique réel, où les électrons possèdent un caractère délocalisé, par un monde fictif où les électrons suivent des trajectoires "classiques" comme les planètes autour du soleil.

Afin d'introduire la délocalisation quantique absente de ce schéma, une composante aléatoire est ajoutée au mouvement des électrons. C'est ce caractère aléatoire qui donne son nom à la méthode : à chaque déplacement d'électron un tirage au sort est effectué comme à la roulette du casino de la fameuse station balnéaire.

Chaque trajectoire ou groupe de trajectoires aléatoires peut être distribué à volonté sur un nombre arbitrairement grand de cœurs de calcul, chacune de ces trajectoires évoluant indépendamment des autres (pas de communications entre elles). La situation est donc idéale du point de vue du calcul parallèle.

La **Figure 2** illustre la méthode QMC appliquée au cas de la simulation de l'atome d'hydrogène, le plus simple des atomes (un électron "gravitant" autour d'un proton).

L'image insérée dans le bloc bleu représente quelques milliers de pas d'une trajectoire de l'électron se déplaçant aléatoirement autour du noyau fixe (en blanc) de l'atome. Les trajectoires obtenues sur des cœurs de calcul indépendants peuvent être superposées à volonté (partie droite de la figure) et la densité électronique exacte (probabilité de présence de l'électron) est reconstruite dans le cas-limite d'un nombre infini de trajectoires.

Aspects computationnels

Déployer efficacement des simulations où des milliers d'électrons effectuent des milliards de pas nécessite d'optimiser au mieux les aspects algorithmiques et informatiques les plus critiques. Détaillons maintenant les principales stratégies que nous avons employées.

Nous définissons un *marcheur* comme l'ensemble des coordonnées *x*, *y* et *z* de chacun des N électrons du système (un vecteur à 3 *N* dimensions). Pour réaliser une trajectoire, un marcheur va effectuer une marche aléatoire dans l'espace à 3 *N* dimensions, de façon à reconstruire la densité de probabilité correspondant au carré de la fonction d'onde.

Nous définissons un *bloc* (**Figure 2**) comme un ensemble de trajectoires courtes réalisées par un groupe de marcheurs pendant un certain nombre de pas (de l'ordre de 10 000). Pour chaque bloc, on calcule les valeurs moyennes des propriétés (énergie, moment dipolaire, etc.) à partir des positions successives de chacun des marcheurs le long de toutes les trajectoires du bloc.

Si les blocs sont suffisamment longs, les positions finales des marcheurs sont complètement décorrélées des positions initiales, de sorte qu'on peut considérer les blocs comme étant tous indépendants. Dans ce cas, les valeurs moyennes calculées sur les blocs ont une distribution statistique obéissant à une loi Gaussienne. Chaque moyenne de bloc est un échantillon que l'on utilise pour calculer statistiquement les valeurs moyennes des propriétés.

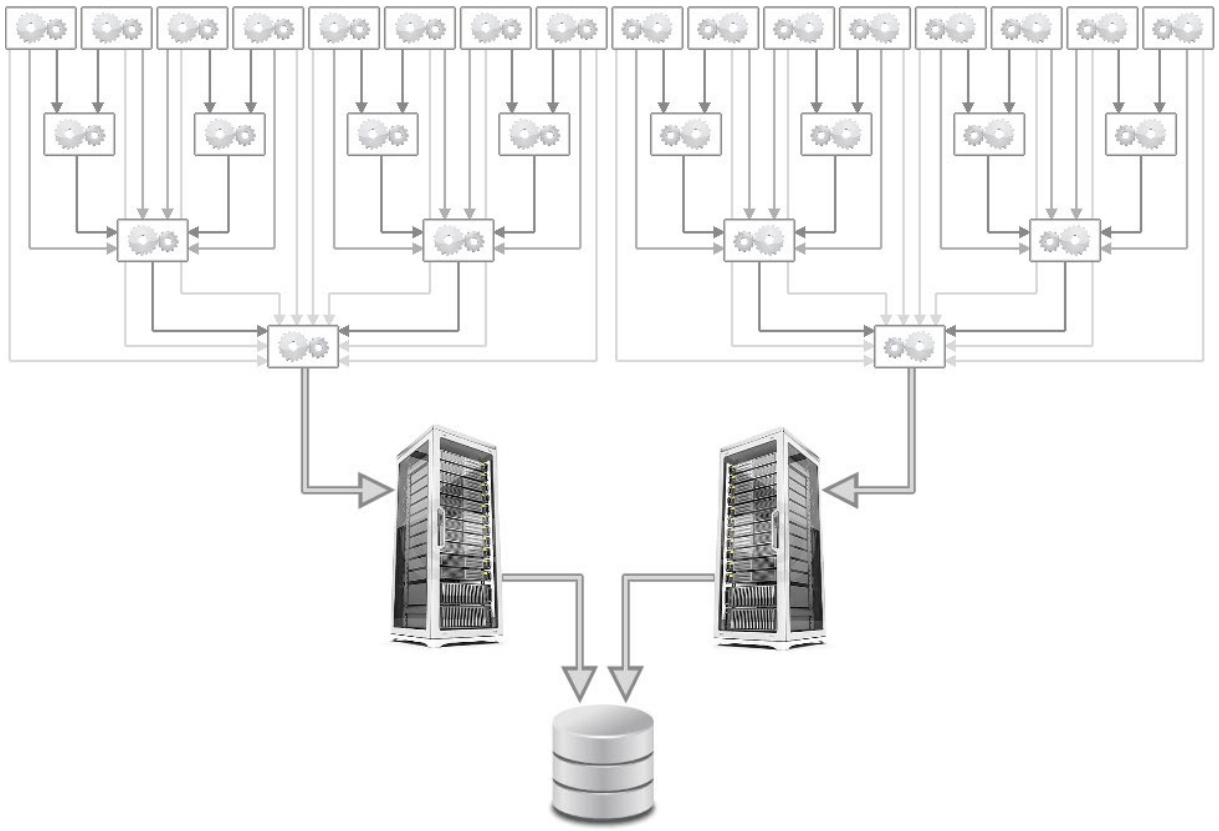


Fig. 3 - Les clients sont organisés en arbre binaire afin d'alléger les communications réseau : un client envoie son tampon de résultats à son parent dans l'arbre. Lorsque le parent reçoit des données provenant d'un de ses fils, il les ajoute à son tampon de résultats, qui sera envoyé plus tard à son parent (le grand-père), etc. Ainsi, le serveur reçoit par le réseau un petit nombre de gros paquets de résultats au lieu d'un grand nombre de petits paquets. Résilience : si un parent ne répond pas, le client va envoyer son tampon de résultats à son grand-père. Si le grand-père ne répond pas non-plus, le client va l'envoyer à l'arrière grand-père, etc., jusqu'à ce que le paquet finisse par être directement envoyé au serveur de données.

Le but est alors de calculer le plus de blocs possible le plus rapidement possible. Pour cela nous avons adopté deux stratégies complémentaires. La première met en place un système de parallélisation efficace. La seconde réalise la réduction du temps de calcul de chaque bloc en travaillant sur l'optimisation mono-cœur.

Gérer la résilience

Lorsque l'on utilise des dizaines de milliers de cœurs, la tolérance aux pannes devient critique. Si un nœud de calcul tombe en panne en moyenne au bout de cinq ans, un calcul mobilisant 2 000 nœuds simultanément ne pourra pas travailler 24 heures sans défaillance. Nous devons donc disposer d'un système qui survive coûte que coûte.

Dans le cas d'algorithmes déterministes le calcul est découpé en tâches, l'accomplissement de chacune d'elles étant indispensable pour obtenir un résultat correct. En cas de panne, certaines tâches ne pourront pas être effectuées.

Les bibliothèques MPI sont bien adaptées aux calculs déterministes. En effet, lorsqu'un client MPI ne répond plus, tout le reste de la simulation est tué car cer-

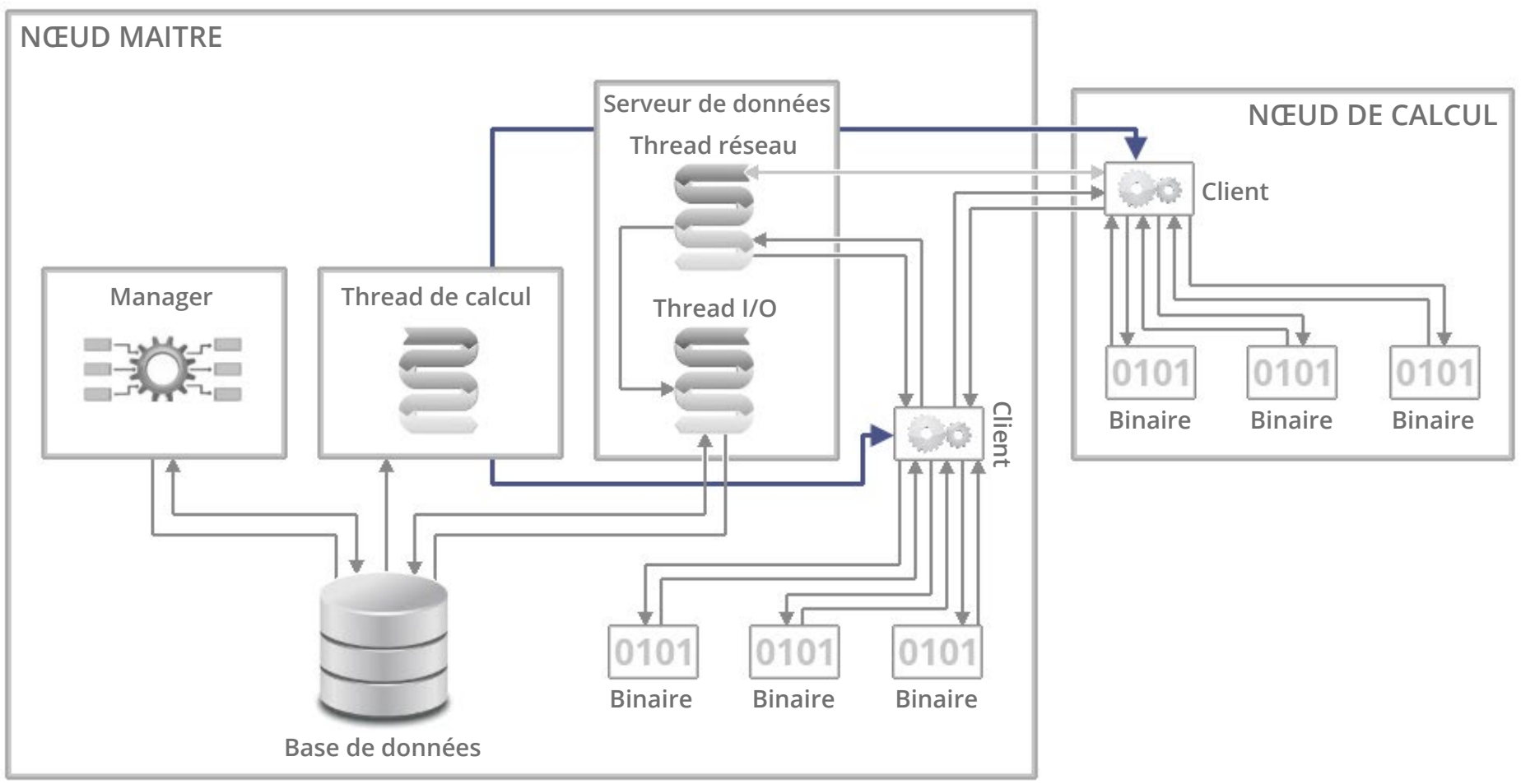


Fig. 4 - Schéma fonctionnel du modèle algorithmique client-serveur.

taines tâches ne pourront pas être réalisées. Dans notre modèle, perdre des blocs au cours de la simulation ne change pas les moyennes, mais influe seulement sur la barre d'erreur du résultat. Nous avons donc choisi de ne pas utiliser MPI pour la parallélisation, mais nous avons implémenté un modèle client/serveur, où les clients calculent des blocs et envoient les résultats au serveur qui les stocke dans une base de données. Si un client meurt, tant pis, les autres clients continuent à calculer.

Les clients et les serveurs sont des scripts Python multithreadés qui communiquent par sockets TCP (**Figure 4**). Notons que les communications étant non-bloquantes et peu intensives, les latences dues à la couche TCP ne posent pas de problème. Il y a un seul client par nœud de calcul. Chaque client lance autant de programmes de calcul qu'il y a de cœurs physiques sur le nœud. Le programme de calcul est un binaire Fortran monothread connecté au client par un *pipe* Unix. Dès que le résultat est communiqué au client, le programme de calcul commence immédiatement à calculer le bloc suivant. Simultanément, lorsque le client est inactif, il transmet ses résultats à un autre client dans le but de les acheminer vers le serveur.

Le serveur comporte deux threads principales : une thread *réseau* et une thread *I/O*. Le serveur reçoit des données par la thread réseau, les traite et les place dans une file d'attente. Simultanément, la thread I/O vide la file d'attente pour mettre les résultats sur disque. Les clients n'accèdent jamais au disque dur local ou au système de fichiers partagé mais plutôt à un disque virtuel en RAM (/dev/shm) afin d'échapper aux pannes liées au stockage temporaire (disque plein, panne physique, etc.).

Facteur d'accélération en fonction du nombre de nœuds

Puisque les blocs sont indépendants, aucune synchronisation n'est nécessaire entre les clients, et cela permet d'obtenir un facteur d'accélération presque idéal en fonction du nombre de nœuds. Cependant, il y a deux étapes critiques qui vont nuire à l'idéalité du facteur d'accélération : l'initialisation et la terminaison.

Pour avoir une initialisation rapide, il faut démarrer les clients aussi vite que possible sur les nœuds de calcul. La mé-

thode qui nous paraît la plus rapide est l'utilisation d'un lanceur MPI qui va envoyer par *MPI_broadcast* un fichier tar contenant le binaire statique Fortran, les scripts Python et le fichier d'*input* sur tous les nœuds. Lorsqu'un nœud reçoit l'archive, il la décompresse et lance le client qui démarre immédiatement les programmes de calcul. Lorsque tous les clients ont démarré, le lanceur MPI se termine. Ainsi, chaque nœud de calcul démarre dès que possible. Pour un calcul sur un millier de nœuds, nous avons mesuré un temps d'initialisation de l'ordre d'une vingtaine de secondes.

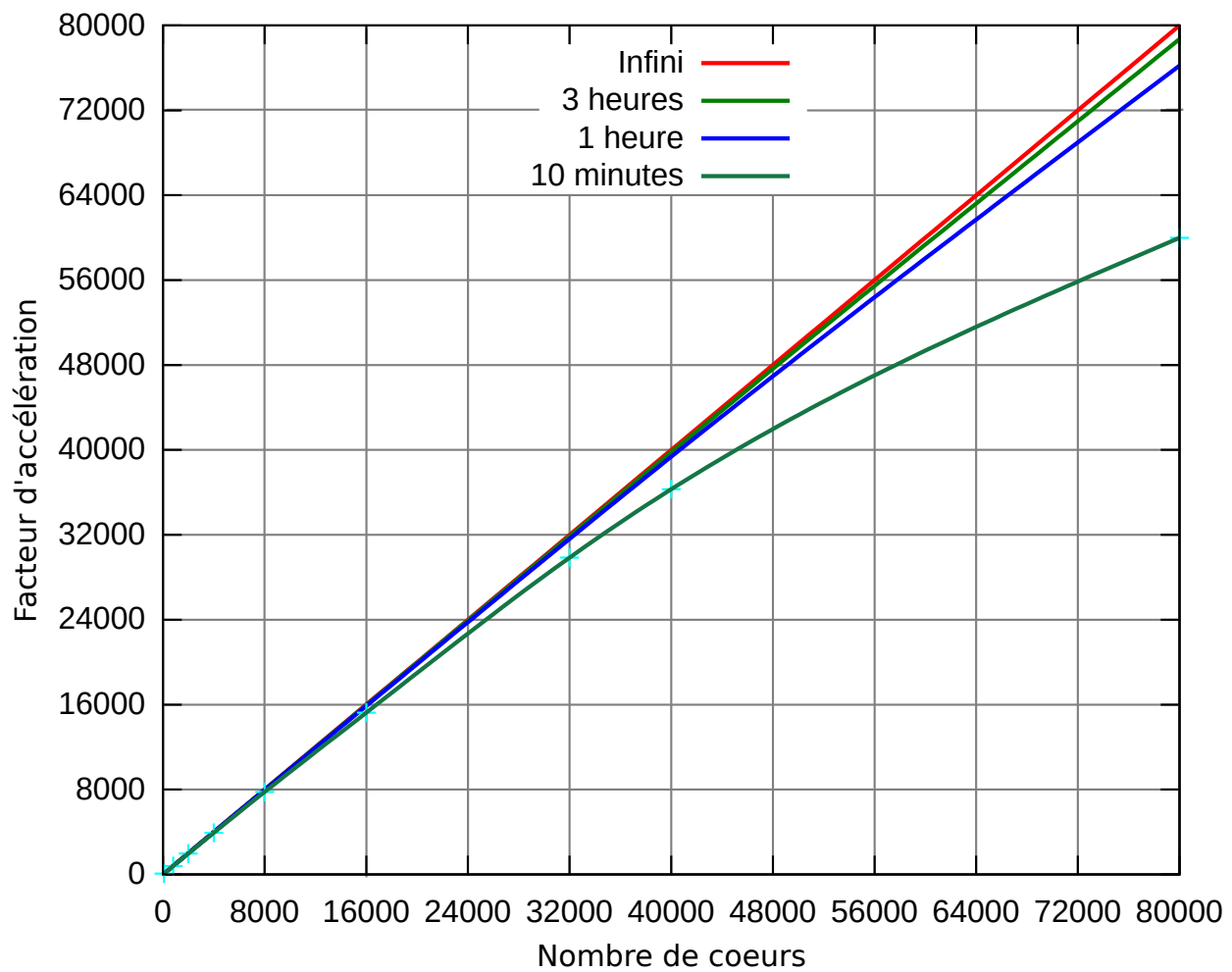


Fig. 5 - Mesure de l'efficacité parallèle résultant des différentes stratégies combinées.

Pour la terminaison, sachant que tous les nœuds de calcul sont désynchronisés, il faudrait attendre que chacun des nœuds ait fini de calculer son bloc. Il pourrait en résulter une latence importante, beaucoup de nœuds attendant que les derniers aient terminé. On ne peut pas non plus tuer violemment tous les clients, car les résultats des blocs en cours seraient perdus.

Nous avons donc permis aux programmes d'intercepter le signal SIGTERM afin d'écourter le bloc en cours et de transmettre le résultat de ce bloc au client. Ainsi, la terminaison d'un client est presque immédiate et aucune seconde de calcul n'est perdue. Pour un calcul d'un millier de nœuds, nous avons mesuré une terminaison de l'ordre de 10 secondes. Ces éléments combinés permettent d'obtenir la courbe de la **Figure 5**. À nombre constant de nœuds, le temps d'initialisation et de terminaison est constant, donc plus le temps de calcul est long, plus l'efficacité parallèle est bonne.

Optimisations mono-cœur

Chaque seconde gagnée dans le calcul se répercutera sur la totalité de la simulation car aucune synchronisation n'est bloquante. Ainsi, nous avons entrepris un travail important d'optimisation mono-cœur de notre programme **QMC=Chem** en profitant de l'expérience et des outils développés par le groupe du **Pr Jalby** au Laboratoire Exascale Computing Research (Intel-CEA-UVSQ-GENCI) à l'Université de Versailles-Saint-Quentin-En-Yvelines.

Dans notre algorithme, nous devons évaluer à chaque pas Monte-Carlo (plusieurs milliards) la fonction d'onde, ses dérivées par rapport à chacune des coordonnées électroniques et son Laplacien. Ces opérations font intervenir des produits de petites matrices (< 1000 x 1000) dont l'une est dense et l'autre creuse.

L'outil d'analyse statique MAQAO développé par nos collègues de Versailles nous a aidé à écrire une routine de produit matrice dense x vecteur creux dont les boucles les plus internes atteignent théoriquement les 16 flops/cycle en simple précision sur les processeurs Intel Sandy Bridge du calculateur CURIE (TGCC-GENCI-CEA). Pour cela, nous avons effectué les modifications suivantes dans le but de favoriser la vectorisation :

• Tous les accès à la mémoire sont consécutifs ;

• Tous les tableaux sont alignés sur 32 octets en utilisant des directives du compilateur ;

• La dimension la plus interne des tableaux multidimensionnels est toujours un multiple de 8 éléments, afin que chaque colonne du tableau soit alignée sur 32 octets ;

• Déroulage de la boucle externe (*unroll and jam*) pour réduire le nombre d'écritures en mémoire ;

• Distribution des boucles pour ne pas dépasser les 16 registres et donc réduire les accès au cache L1.

En pratique, étant donné que nous faisons de l'ordre de N^2 opérations (N, nombre d'électrons) pour N^2 accès à la mémoire, le calcul est inévitablement limité par les accès à la mémoire. Nous avons mesuré jusqu'à 61% de la performance crête du processeur dans cette routine de produit matrice x vecteur.

Lors de l'installation de CURIE en décembre 2011, et grâce au soutien des ingénieurs de BULL, nous avons pu réaliser notre tout premier calcul QMC grandeur nature sur l'ensemble des ≈ 80 000 cœurs de la machine. Celle-ci étant en cours d'installation, les calculs ont subi des interruptions et des pannes. Ce qui initialement nous avait apparu comme une gêne s'est finalement révélé une chance : le fait que nos simulations aient pu être menées à bien malgré ces difficultés a validé en pratique la grande robustesse de notre schéma.

Simulations pétaflopiques pour la chimie d'Alzheimer

Grâce aux moyens exceptionnels mis à notre disposition ces deux dernières années par GENCI et le consortium européen PRACE, notre groupe a pu démontrer la faisabilité pratique de nos simulations pour des systèmes difficiles à atteindre par les méthodes usuelles du domaine. Les applications abordées ont concerné la chimie de la maladie d'Alzheimer, qui met en jeu des interactions particulièrement délicates à reproduire.

On sait que cette maladie est associée à une dégénérescence du cerveau liée à l'apparition de plaques composées d'agrégats de molécules (peptides) dites β-amyloïde. Comprendre pourquoi et comment ces molécules s'agrègent est un des enjeux majeurs de la compréhension des mécanismes les plus fondamentaux à l'origine de l'apparition de la maladie.

En collaboration avec le groupe de biochimie expérimentale du Pr Faller (LCC, Toulouse), nous avons pu démontrer que la précision chimique nécessaire pour décrire quantitativement ces systèmes pouvait être atteinte en utilisant le calculateur pétaflopique CURIE mentionné plus haut. La représentation graphique qui ouvre ce dossier illustre comment les trajectoires aléatoires des électrons se déploient dans le cas des systèmes chimiques simulés.

Lors de ces simulations, un débit de calcul réel de presque 1 Pétaflops soutenu pendant plusieurs heures (précisément, 960 Tflops/s en simple et double précision) a pu être mesuré.

Même si beaucoup reste à faire, nos premières simulations probabilistes sur une architecture massivement parallèle ouvrent des perspectives prometteuses.

Bien que les volumes de calcul nécessaires - qui représentent des millions d'heures CPU monoprocesseur - soient encore beaucoup trop importants pour en faire des méthodes de routine, l'arrivée d'ici 2020 des machines exaflopiques mille fois plus puissantes permet d'espérer transformer ces techniques émergentes en un outil de simulation accessible à la communauté scientifique la plus large. ■

Les auteurs remercient l'ANR pour son soutien à ce travail (projet QMC=Chem ANR 2011 BS08004 01), ainsi que CALMIP (Toulouse), GENCI et PRACE pour les moyens informatiques importants mis à leur disposition.

VOUS SOUHAITEZ FAIRE CONNAITRE VOS TRAVAUX ?

Contactez la rédaction !

PC EXPERT

Entreprise
Débit, intrus
Sécurisez votre réseau

Développement
Créez un lecteur vidéo
en HTML5 et jQuery

e-Réputation
défendez-vous
CE QU'IL FAUT SAVOIR
COMMENT LA GÉRER

+ Un mois avec Windows 8.1
Le verdict de la rédaction

Télécharger dans l'App Store

NOUVEAU
MAGAZINE TACTILE POUR INFORMATIQUE AGILE
www.pcexpertlemag.fr

/découvrir

FIN DE ROUTE POUR ROADRUNNER

Il nous avait fait rêvé. Il est désormais hors service. Première machine à passer, dès 2008, la barre du Pétaflops (10^15), le voilà maintenant hors circuit, depuis le 31 mars dernier, prêt à livrer ses derniers secrets à travers ce qui s'apparente déjà à une autopsie technique complète.

Installé au laboratoire de Los Alamos, Roadrunner faisait partie du programme Advanced Simulation and Computing (ASC) de la National Nuclear Security Administration américaine. Pendant ses cinq années de service actif, il a été en charge des plus grosses simulations du programme *Stockpile Stewardship* (maintenance et test de fiabilité des armements nucléaires Outre-Atlantique), une mission qui a évité à la planète bien des essais réels. Véritable montre de puissance à l'époque, il a par ailleurs contribué à une certain nombre de projets scientifiques non classés Défense pendant sa phase de construction et de testing.

Chris Deeney, Administrateur adjoint de la NNSA pour le programme en question, témoigne de ce que "*Roadrunner était un exemple : une équipe de haut niveau intégrant des codes complexes dans des architectures de pointe afin d'assurer une force de dissuasion sécurisée et efficace sans recourir aux essais souterrains. Roadrunner et ses succès nous ont mis permis d'avancer de façon décisive dans le domaine du HPC. Il a été un formidable vecteur de modernisation technologique pour notre armée et, plus globalement, pour nos recherches scientifiques.*"

Si la conception de Roadrunner était unique, elle a aussi été controversée. La machine intégrait en effet deux types de processeurs, ce qui a longtemps fait d'elle le plus gros calculateur hybride au monde. On y dénombrait 6 563 CPU génériques (des AMD Opteron dual-core), chaque cœur étant

par ailleurs relié à un GPU IBM PowerXCell 8i (qu'on a vite appelé tout simplement "Cell"). La presse s'est largement fait l'écho de ce choix technologique surprenant, le Cell étant une version améliorée (8 VPU, 100 Gflops par puce) d'un GPU conçu à l'origine pour la Playstation 3, de Sony. Cette spécificité, ajoutée à la roadmap pour le moins incertaine de PowerXCell, faisait douter beaucoup de spécialistes de sa pertinence technique. Mais peu importe. Pour le laboratoire de Los Alamos et pour IBM, Roadrunner était à la fois une vitrine et un porte-drapeau.

La vocation de Roadrunner était donc le calcul intensif dédié à la simulation. Il a ainsi permis de répondre à un certain nombre de questions relatives à la gestion du flux énergétique des armes atomiques et à la relation entre ce flux et le rendement desdites armes. Les progrès issus de son utilisation ont été décisifs dans ce domaine. Tout comme le furent, pendant sa période d'installation, les expériences qui lui ont été confiées sur le comportement des nanofils, la reconnexion magnétique, la rétrodiffusion laser, la phylogénétique du VIH - sans oublier une simulation de l'univers à une échelle de 70 milliards de particules. Des voix se sont élevées pour critiquer les coûts de portage applicatifs mais, là encore, les acquis technique en matière de stratégie de développement sont bel et bien là.

"*Roadrunner était vraiment une idée innovante*" explique Gary Grider, du département HPC du Laboratoire. "*Il a transformé définitivement notre façon de penser, de construire et d'utiliser les calculateurs. Grâce à lui, l'intégration de processeurs spécialisés et d'accélérateurs dans les systèmes est maintenant*

monnaie courante. En ce sens, Roadrunner était en quelque sorte une démonstration de faisabilité technologique. Mêmes les plus sceptiques a priori ont été obligés d'y prêter attention."

C'est en effet de son architecture hybride que provenait sa puissance, encore classée 22ème au TOP500 de novembre dernier après avoir été n° 1 deux ans durant. Les deux processeurs se partageaient le traitement des codes mais ce sont feu les Cell qui prenaient la main dès que les choses devenaient sérieuses, comme le font aujourd'hui les accélérateurs Tesla ou Xeon Phi. Cet équilibre CPU / GPU a également montré la voie en matière d'efficacité énergétique, Roadrunner délivrant son potentiel avec "seulement" 2,3 MW de consommation moyenne relevée. Dans ce domaine également, il y a eu un avant et un après Roadrunner.

Malgré ces états de service remarquables, c'est sans cérémonie qu'il a été débranché, comme un projet - ce qu'il était d'un point de vue politique - arrivé à terme. Avant qu'il ne soit complètement démonté, les spécialistes IT de Los Alamos ont eu un mois pour réaliser un certain nombre d'expériences, notamment de compression mémoire sur les OS et d'optimisation de routage des données. Autant de recherches qui n'auraient pu être menées "de son vivant".

Ensuite, Roadrunner a été examiné sous ses coutures les plus intimes : intégrité de ses composants (mémoire notamment), efficacité de son système de refroidissement après cinq ans, degré d'usage de son armature mécanique... Ces analyses sur l'état d'usage d'un système à haut rendement (informatique et thermique) n'ont officiellement pas filtré, mais il apparaît qu'une durée de vie de cinq ans sans gros incident est déjà un beau parcours. Roadrunner aurait-il pu tenir deux années de plus ? Probablement pas sans une maintenance lourde, semble-t-il.

Reste donc un espace vide de près de 600 m², soit environ 15 % de plus que la place au sol occupée par Titan. La nature en général et les laboratoires de calcul intensif en particulier ayant horreur du vide, il se pourrait que les lieux soient prochainement investis par une machine "commissionnée" pour dépasser les 100 Pflops. De quoi réjouir les *exascale optimists*, qui y voient un point d'inflexion dans la course au prochain passage à l'échelle. En tous cas, s'il ne fallait qu'une mesure des progrès accomplis ces cinq dernières années par la communauté HPC... ■

CARRI Systems

Architecte technologique depuis 1992

Intel Technology Provider Platinum 2013

NVIDIA PARTNERFORCE TESLA PREFERRED PROVIDER

Concepteur d'Architectures Rationnelles pour la Recherche et l'Industrie

«Créée en 1992 par des passionnés, CARRI Systems est le dernier fabricant français informatique haut de gamme sur-mesure. Basée à Noisy-le-Sec, près de PARIS, la société est, depuis quelques années, devenu un acteur majeur sur le marché du GPU Computing et du calcul intensif. CARRI Systems est membre du Pôle européen de compétence en simulation numérique haute performance (Ter@tec) depuis 2011.»

Franck Darmon
CEO CARRI Systems

intel NVIDIA AMD

FABRIQUÉ EN FRANCE

Découvrez toutes nos solutions XLR sur

www.carri.com

/développer

PROGRAMMER XEON PHI

OPTIMISER LES IMPLEMENTATIONS MPI [2ÈME PARTIE]

Souvent ignorées des scientifiques, les techniques d'optimisation des couches MPI sous-jacentes aux applications peuvent s'avérer décisives en termes de performances. Ce mois-ci, nous nous penchons sur les stratégies d'interaction avec les transports InfiniBand, avec à la clé un bug dans MVAPICH2 et un appel à témoins…

JEROME VIENNE, PHD*

Si les bibliothèques MPI (MPICH, Open MPI, MVAPICH2, Intel MPI, etc.) suivent un standard, elles se caractérisent aussi par un ensemble de spécificités à bien connaître si l'on veut optimiser finement l'exécution des applications qui les utilisent. Dans notre article précédent, qui traitait la question du placement des processus MPI et du choix des algorithmes utilisés pour les communications collectives, nous avons vu que les choix techniques des développeurs des bibliothèques étaient parfois très différents, et que les ignorer pouvait nous priver d'importants facteurs d'accélération.

Ce mois-ci, nous allons nous pencher sur les possibilités offertes par les implémentations MPI lorsque que l'on exécute une application à grande échelle sur des connexions Infi-

* HPC Software Tools Group
Texas Advanced Computing Center (TACC)

niBand. Comme chacun sait, le nombre de cœurs présents au sein des clusters du TOP500 est en perpétuelle augmentation. Pour les scientifiques, c'est évidemment tout bénéfice : ces énormes ressources permettent d'attaquer des problèmes de plus en plus complexes, avec un effet boule de neige d'un point de vue interdisciplinaire. Mais utiliser plus de cœurs augmente aussi la consommation mémoire pour chaque processus MPI...

Dans un tel contexte, réduire l'espace mémoire occupé par les processus devient critique. Mon expérience à TACC m'a permis de constater que beaucoup d'utilisateurs de nos grappes de calcul ignorent que le standard InfiniBand et son implémentation Mellanox proposent différents moyens pour aider à réduire cette consommation mémoire. Nous allons donc en regarder deux de plus près : l'eXtended Reliable Connection (XRC) et l'Unreliable Datagram (UD).

Au cœur d'InfiniBand

Année après année, InfiniBand s'est fait une place de choix dans la communauté HPC. En juin 2003, une seule configuration TOP500 l'utilisait ; en ce moment, elle équipe 205 systèmes, soit 41 % du classement. Il s'agit donc, sans conteste, de la principale architecture réseau haut performance disponible aujourd'hui.

En pratique, la rapidité des connexions InfiniBand réduit la durée des phases de communication et, par conséquent, limite les temps d'attente au niveau des cœurs. Un réseau InfiniBand se distingue en effet par deux caractéristiques essentielles : une faible latence et une large bande passante, synonyme de débits élevés. Pour mémoire, la latence correspond au temps nécessaire à l'envoi et à la réception du premier octet, tandis que la bande passante mesure le débit maximal obtenu via l'envoi de messages de grandes tailles.

D'un point de vue technique, l'architecture InfiniBand définit quatre types de couches transport : Reliable Connection (RC), Reliable Datagram (RD), Unreliable Datagram (UD) et Unreliable Connection (UC). Les couches RC et UD étant les plus à même d'être utilisées par les applications MPI, c'est à elles que nous allons nous intéresser aujourd'hui.

Nous négligerons volontairement la couche RD, qui n'est implémentée par aucun matériel (contrairement aux trois autres) et n'est pas utilisable avec MPI. UC, quant à elle, offre un mode connecté mais, outre que ce mode n'est pas

Tous
nos articles,
nos dossiers,
nos chroniques,
nos codes sources
et bien plus encore...

www.hpcmagazine.fr

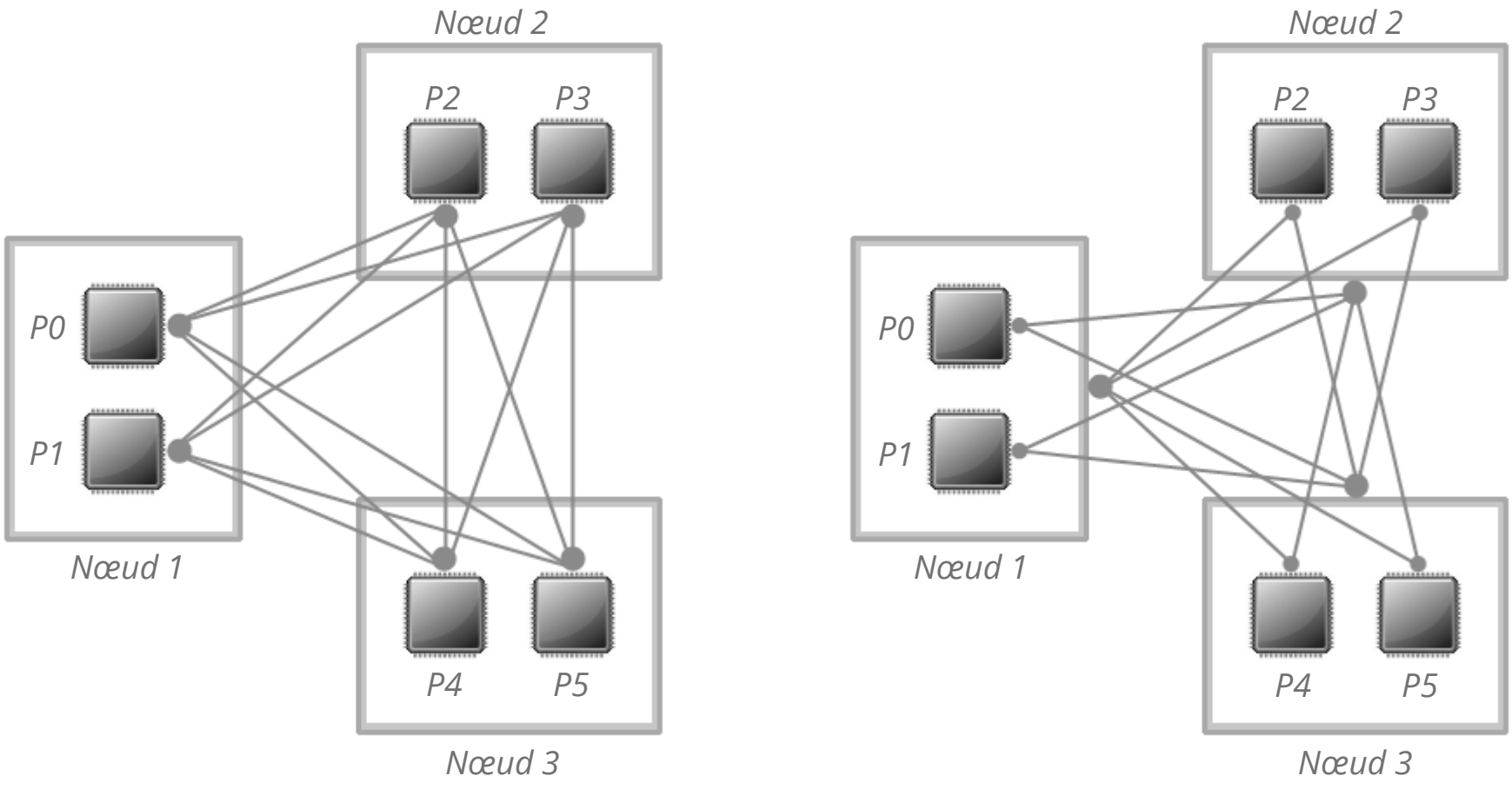


Fig. 1 - Comparaison des connexions entre RC (à gauche) et XRC (à droite).

d'une grande fiabilité, l'ordre d'arrivée des paquets n'est pas garanti, ce qui n'est pas simple à gérer. Bien qu'elle supporte les accès RDMA (Remote Direct Memory Access), la quantité mémoire nécessaire pour l'utiliser est identique à celle de RC, si bien qu'elle est finalement peu intéressante pour les bibliothèques MPI.

En termes de proportion d'usage, RC se taille la part du lion. Pour les développeurs, le choix de RC se justifie par sa fiabilité, par le fait qu'elle fonctionne en mode connecté et parce qu'elle est le transport qui offre le plus de possibilités sur InfiniBand (RDMA, opération atomique, etc.). Cela étant précisé, lorsque le nombre de processus MPI augmente, RC est aussi le mode de transport le plus coûteux en ressources. Pourquoi ? L'explication est simple : comme on est en mode connecté, chaque partenaire d'une communication doit établir une connexion dédiée nécessitant un espace mémoire de quelque Ko. C'est peu, mais lorsque l'on exécute des expériences à grande échelle, l'espace mémoire combiné peut très vite atteindre de l'ordre du Go par processus.

Si l'on considère *N* nœuds chacun avec *C* cœurs, chaque processus MPI devra établir *(N-1) x C* connexions pour pouvoir communiquer avec les autres processus MPI. Pour faire face à ce problème, compte tenu de l'augmentation régulière du nombre de cœur par socket, Mellanox a développé l'eXtended Reliable Connection sur ses cartes ConnectX. XRC est donc disponible sur ces cartes et les suivantes, pour peu que l'on utilise la bibliothèque **OFED** (voir **figure 1** page précédente) en version 1.3 ou postérieure.

Priorité à l'eXtended Reliable Connection (XRC)

Dans les transports InfiniBand, il n'y a pas de distinction entre une connexion au niveau cœur ou au niveau nœud. L'idée d'XRC est donc d'autoriser une connexion unique entre un processus et un nœud entier, ce qui réduit le nombre de connexions nécessaires pour chaque processus MPI au nombre total des nœuds. En utilisant XRC, chaque processus MPI devra donc établir *(N-1)* connexions pour pouvoir communiquer avec tous les autres processus MPI. La **figure 1** permet d'apprécier cette différence. Au passage,

il faut noter qu'XRC peut être utilisée avec la plupart des bibliothèques MPI dont bien sûr Open MPI, MVAPICH2 et Intel MPI. Attention, dans ce dernier cas, elle ne fonctionne qu'avec les ressources OpenFabrics Alliance et non avec **DAPL**.

Au niveau performance, toutefois, les bénéfices ne sont pas flagrants. Si quelques gains peuvent être observés sur certaines applications dans la mesure où la quantité mémoire disponible est supérieure, nous ne pourrons malheureusement pas les quantifier de façon précise. Nous avions en effet prévu un certain nombre de tests basées sur MVAPICH2 1.9, mais il s'avère qu'XRC présente un bug (en tout cas sur le cluster Stampede) qui rendrait les résultats inutilisables. La rédaction de cette article nous aura au moins permis de le détecter et de le reporter à l'équipe de développeurs.

Cela étant, d'expérience, XRC reste coûteux sur les applications à très grande échelle impliquant un nombre élevé de cœurs ou de nœuds. Dans de telles conditions, et sous certaines réserves, il peut être recommandé d'utiliser la couche Unreliable Datagram, comme nous allons le voir maintenant.

L'Unreliable Datagram (UD)

Contrairement aux couches RC et XRC, UD est un transport non connecté, ce qui réduit le besoin en ressources mémoire pour les processus MPI. L'inconvénient, c'est qu'UD est beaucoup moins riche en fonctionnalités. Par exemple, RDMA n'est pas supporté et la taille des messages se limite à la valeur de la MTU. La bibliothèque MPI a donc plus de travail à faire, notamment pour gérer la segmentation, l'ordre et la retransmission des messages. En toute logique, le coût en performance n'est pas négligeable dans les contextes applicatifs mobilisant un petit nombre de nœuds et fonctionnant à partir de messages de grande taille.

Au contraire, pour les codes utilisant principalement des messages courts, la quantité mémoire requise pour chaque processus MPI reste quasiment constante lorsque l'on augmente la taille du système. Par ailleurs, sachant qu'il n'est plus nécessaire de perdre du temps à établir les connexions, UD permet d'obtenir de meilleurs performance lorsque l'on utilise beaucoup de nœuds. Pour benchmarker cela, nous avons utilisé l'application **SMG2000** compilée avec Intel MPI sur le cluster Stampede, en utilisant 4 096 cœurs et les paramètres suivants :

```
mpiexec.hydra –n 4096 …
./smg2000  -n 120 120 120 -c 2.0
3.0 40 -P 16 16 16
```

Activer UD avec Intel MPI implique de définir deux variables d'environnement :

```
export I_MPI_DAPL_UD_
PROVIDER=ofa-v2-mlx4_0-1u
export I_MPI_DAPL_UD=enable
```

Pour rappel, Intel MPI supporte différents types de protocole de communication, mais elle est principalement optimisée pour DAPL.

Solveur multigrid parallèle, SMG2000 fait partie des "ASC Purple Benchmarks", une série de tests développée par le Lawrence Livermore National Laboratory (LLNL). SMG2000 est structuré en trois étapes : "Struct interface", "Setup phase" et "Solve phase". La **figure 2** page suivante montre la moyenne des temps d'exécution de chacune de ces étapes sur dix exécutions avec et sans UD.

Comme vous pouvez le constater, les temps d'exécution relevés pour la première et la dernière phase sont *grosso modo* identiques. En revanche, la différence est importante concernant la phase Setup. L'explication est simple : cette phase est principalement constituée de communications non bloquantes point-à-point de petite taille entre les différents processus MPI.

UD ne nécessitant pas de connexion, l'envoi de messages est beaucoup plus rapide. Quand on sait que plus de 1 060 841 238 messages sont échangés, on peut comprendre que le léger gain de temps pour chaque communication se traduise au final par une accélération globale sensible.

Autre avantage lié à UD : en règle générale, le temps de démarrage des programmes MPI est beaucoup plus court. Un simple Hello World avec 16K cœurs peut prendre plus d'une 1 minutes 30 avec DAPL alors qu'il démarre en moins de 30 secondes avec DAPL_UD.

Qu'en est-il pour les applications utilisant des messages de

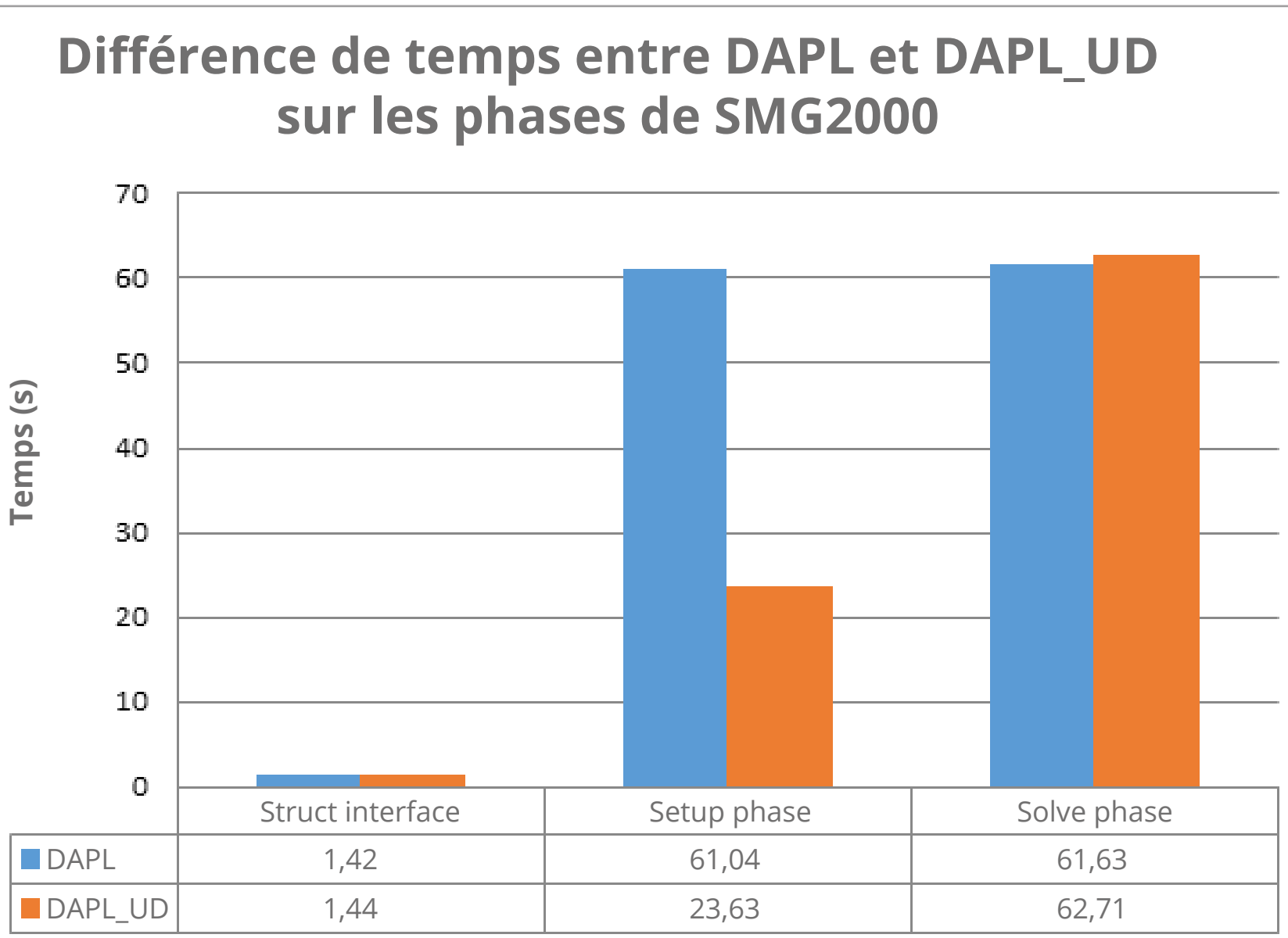


Fig. 2 - Moyenne des temps d'exécution pour chacune des étapes de SMG2000, avec et sans UD.

grandes tailles ? La musique est tout autre. Les envois sur UD ne pouvant dépasser la taille MTU, les messages longs nécessitent plus de paquets, d'où un risque supérieur de perte de ces paquets. Dans un tel contexte, la consommation mémoire peut s'avérer, au cas par cas, nettement plus importante.

Conclusion temporaire

Cet article se proposait d'examiner deux approches différentes destinées à optimiser la scalabilité des applications sur réseau InfiniBand. Laquelle est la meilleure ? On l'a vu, tout dépend de votre bibliothèque MPI et des particularités de votre application. Si celle-ci requiert le passage de larges messages, XRC peut se révéler plus profitable à grande échelle, bien que nous n'ayons pu le tester valablement pour les raisons techniques mentionnées plus haut. En revanche, si les messages de votre applications restent de taille contenue, c'est-à-dire s'ils sont inférieurs à la taille de la MTU, UD peut se révéler meilleur, sous réserve que votre bibliothèque MPI le supporte. Et pour l'avenir ?

La nouvelle famille d'interconnexions Connect-IB de Mellanox propose un nouveau protocole appelé "Dynamically-connected transport", dont l'objectif est précisément de réduire la consommation mémoire. Que donne ce protocole comparé aux deux autres dans les cas que nous avons évoqués ? Sachant que notre laboratoire n'a pas encore déployé les cartes *ad hoc*, difficile de le mesurer de façon définitive. Promis, dès que les conditions du test seront réunies, nous y reviendrons en détail.

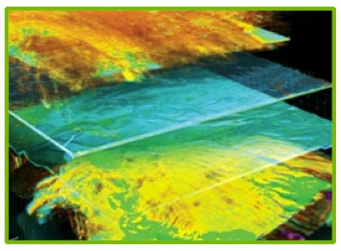
Dans l'intervalle, si vous-même êtes en mesure de benchmarker l'ensemble, nous sommes volontiers preneurs des résultats. Cette rubrique n'a-t-elle pas vocation à être un effort collaboratif visant à partager les meilleures pratiques de développement HPC ?

Bons développements !

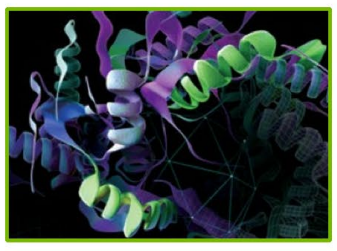
CAPS

The Many-Core Programming Company

x5,6 en 5 jours
Comprendre les origines de l'univers
Panos Labropoulos, Université de Groningen et ASTRON

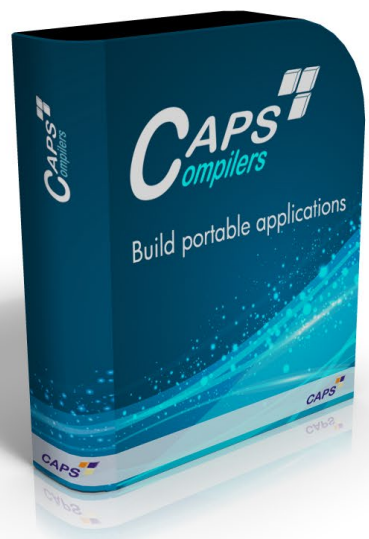


x3 en 7 jours
Trouver du pétrole dans les plus grands réservoirs du monde
Global Oil Company



x5 en 1 jour
Simuler des solvants dans les protéines
Bharat Medasani, Université du Texas à San Antonio

NVIDIA®



Simple.

Puissant.

Ouvert.

Profitez d'une évaluation gratuite de 30 jours de

CAPS OpenACC Compilers

En savoir plus *

* http://www.nvidia.eu/openacc

/développer

```
cudaStream_t lstream, rstream;
cudaStreamCreateWithFlags(&lstream, cudaStreamNonBlocking);
cudaStreamCreateWithFlags(&rstream, cudaStreamNonBlocking);

// Begin by freeing our atomicData storage. It's better for the ringbuffer algorithm
// if we free when we're done, rather than re-using (makes for less fragmentation).
ringbufFree<qsortAtomicData>(atomicDataStack, atomicData);

// Exceptional case: if "lt_len" is zero, then all values in the batch
// are equal. We are then done (may need to copy into correct buffer, though)
if(lt_len == 0)
{
  if(source_is_indata)
    cudaMemcpyAsync(indata+offset, outdata+offset, gt_len*sizeof(unsigned),
      cudaMemcpyDeviceToDevice, lstream);
  return;
}

// Start with lower half first
if(lt_len > BITONICSORT_LEN)
{
```

L'ATELIER CUDA

COPPERHEAD : UN PYTHON PARALLÉLISÉ POUR CUDA

Allier la productivité de Python et l'efficacité de la programmation parallèle sur GPU, telle est la vocation de Copperhead. Compte tenu de l'intérêt que suscite l'un et l'autre chez les scientifiques programmeurs, un petit tour de piste s'impose…

BRIAN CATANZARO, PHD*

Les langages tels que C et Fortran ont un gros avantage : ils permettent de tirer le maximum des ressources matérielles sous-jacentes. En contrepartie, ils nécessitent une bonne connaissance de ces ressources, notamment dans leur architecture de bas niveau. Orientés résultats, ils sont irremplaçables lorsque la performance prime. Mais les choses changent. Plutôt orientés productivité, les langages de plus haut niveau tels que Python ou Ruby peuvent afficher des performances élevées en calcul intensif lorsqu'on les utilise intelligemment. Ce faisant, ils libèrent le programmeur de contraintes techniques liées à la parallélisation - en tous cas au niveau matériel - et produisent des codes qui peuvent souvent être réutilisés sur des plateformes différentes.

C'est l'objectif de Copperhead, un projet expérimental de NVIDIA Research, que d'offrir à chacun ces multiples avantages. Les codes Copperhead permettent d'utiliser des fonctions primitives connues telle que `map` et `reduce`, et s'exécutent en parallèle sur les GPU CUDA et les CPU multicœurs. D'où l'attention que nous lui portons ce mois-ci…

axpy, le Hello World parallèle

Plongeons tout de suite au cœur du sujet, et commençons par le commencement ! Le **listing 1** page suivante montre

* Research scientist, NVIDIA

une version Copperhead d'**axpy**, la forme générique de **saxpy** ("Single-Precision A.X Plus Y"), qui est un peu le Hello World des programmes parallèles. On le voit, les fonctions Copperhead sont intégrées à des programmes Python. Elles n'utilisent d'ailleurs qu'un sous-ensemble bien délimité de ses instructions. Pour marquer ces fonctions, on utilise tout simplement le décorateur Python `@cu`.

A l'exécution, les programmes se déroulent normalement, jusqu'à ce que l'interpréteur Python rencontre un premier appel à une fonction décorée Copperhead. Là, l'exécution transite par le runtime Copperhead puis retourne à l'interpréteur Python, en reprenant au point d'entrée de la fonction d'origine. Les données sont alors ramassées par le garbage collector standard de Python.

Dans notre programme **axpy**, une opération par éléments est lancée sur les deux tableaux d'entrée, via une list comprehension. C'est elle qui sera exécutée en parallèle. Notez qu'on aurait également pu coder **axpy** en utilisant `map` et une fonction anonyme lambda, comme au **listing 2**, ou une fonction nommée imbriquée, comme au **listing 3**. Les trois implémentations fonctionnent de la même façon et donnent les mêmes résultats. Libre au programmeur d'utiliser celle qu'il préfère.

Le `map` de Python exécute les opérations élément par élément sur les séquences, en passant les éléments des séquences d'entrée à la fonction invoquée. Que se passe-t-il pour les données auxiliaires qui ne sont pas traitées par élément, mais qui doivent être utilisées dans un calcul par éléments ?

Les programmes Copperhead utilisent des protections pour pouvoir diffuser ces données ; ainsi, dans notre exemple, vous pouvez observer que la fonction par éléments utilise la scalaire `a` de portée parente. C'est une pratique idiomatique courante en Python, et c'est de cette façon que les programmes Copperhead transmettent les données aux opérations par éléments.

Autre point intéressant dans cet exemple, l'interaction avec l'incontournable bibliothèque **numpy** pour la constitution de tableaux homogènes et bien typés. **axpy** est appelé avec des valeurs 64-bit en virgule flottante (ce qui le transforme en **daxpy**) car il s'agit du type par défaut pour les littéraux fp de Python (`a = 2.0`) mais également pour les tableaux

Listing 1 - axpy, "le Hello World" parallèle, en version Copperhead.

```
from copperhead import *
import numpy as np

@cu
def axpy(a, x, y):
    return [a * xi + yi for xi, yi in zip(x, y)]

n = 1000000
a = 2.0
x = np.random.rand(n)
y = np.random.rand(n)
with places.gpu0:
    gpu_result = axpy(a, x, y)
with places.openmp:
    cpu_result = axpy(a, x, y)
```

Listing 2 - axpy codé avec map et une fonction lambda anonyme.

```
@cu
def axpy(a, x, y):
    return map(lambda xi, yi: a * xi + yi, x, y)
```

Listing 3 - axpy codé avec une fonction nommée imbriquée.

```
@cu
def axpy(a, x, y):
    def el(xi, yi):
        return a * xi + yi
    return map(el, x, y)
```

aléatoires qui utilisent `numpy.random.rand`. Pour appeler ce programme avec des valeurs fp 32-bit (ce qui en fait un **saxpy**), il faut utiliser les mécanismes numpy standards comme au **listing 4**.

Listing 4 - Mécanismes numpy standards d'appel d'un saxpy avec des valeurs fp 32-bit.

```
a = np.float32(2.0)
x = np.array(np.random.rand(n), dtype=np.float32)
y = np.array(np.random.rand(n), dtype=np.float32)
```

Enfin, remarquez que la fonction ne contient pas de boucle. En principe, Copperhead ne supporte pas les effets de bord comme les modification de valeur des variables. C'est une restriction importante, qui a pour avantage de sécuriser la parallélisation des programmes.

Listing 5 - Implémentation Python d'axpy écrasant une des variables en entrée via une boucle.

```
def side_effect_axpy(a, x, y):
    for i in range(len(y)):
        y[i] = a * x[i] + y[i]
```

Le **listing 5** montre une implémentation de **axpy** dans laquelle une boucle écrase une des données en entrée. Ce style de codage n'a pas cours avec Copperhead. Sécuriser les boucles de parallélisation telles que celle-ci est pose donc un problème, sachant que certaines boucles ne peuvent tout simplement pas être parallélisées. Alors, plutôt que des boucles, les codes Copperhead utilisent des primitives conçues pour être toujours parallélisables. Le runtime peut ainsi planifier les calculs en fonction de la plateforme cible. Et dans les cas où une itération séquentielle est nécessaire, Copperhead passe par une classique récursion. Compte tenu de ce qui précède, tous les branchements vers Copperhead doivent se terminer par une instruction `return`.

Listing 6 - Equivalence entre les notations map et list comprehension.

```
z = [fn(xi, yi) for xi in zip(x, y)]
z = map(fn, x, y)
```

Les primitives Copperhead

Les programmes Copperhead utilisent un jeu de primitives parallèles claires. Partout où cela est possible, ces primitives exposent les mêmes interfaces que les primitives natives de Python. Voici les plus importantes :

- `map(f, ...)` applique la fonction `f` sur un certain nombre de séquences. Son arité doit être égale au nombre d'arguments, et les séquences doivent avoir la même longueur. Pour des questions pratiques, cette arité est limitée à 10. Et pour le confort du programmeur, les list comprehensions sont traitées de façon équivalente à `map`. Les deux lignes du **listing 6** sont par conséquent équivalentes.

- `reduce(fn, x, init)` applique cumulativement une fonction binaire `fn` aux éléments de `x` pour les réduire à une seule valeur. La fonction `fn` doit nécessairement être cumulative et commutative et, contrairement au `reduce` de Python, la sémantique de parallélisation ne garantit pas que les éléments seront réduits de gauche à droite. Attention, c'est un piège fréquent.

- `filter(fn, x)` retourne une séquence contenant les éléments de `x` pour lesquels `fn(x)` renvoie `True`. L'ordre des éléments de la séquence `x` est préservé.

- `gather(x, indices)` retourne la séquence `[x[i] for i in indices]`.

- `scatter(src, indices, dst)` crée une copie de `dst` et la met à jour en distribuant chaque `src[i]` vers `indices[i]` dans la copie. Si un indice quelconque est dupliqué, l'une des valeurs correspondantes de `src` est choisie arbitrairement et placée dans le résultat. C'est alors la copie mise à jour qui est renvoyée.

• **`scan(fn, x)`** retourne le scan inclusif (appelé également "prefix sum") de **`fn`** sur la séquence **`x`**. Il en va de même pour **`rscan`**, **`exclusive_scan`**, et **`exclusive_rscan`**.

• **`zip(...)`** retourne une séquence de tuples. Toutes les séquences en entrée doivent avoir la même longueur.

• **`unzip(x)`** retourne un tuple de séquences.

• **`sort(fn, x)`** retourne une copie triée de **`x`** à partir d'un comparateur binaire **`fn(a,b)`** tel que **`(a < b)`** est **`True`**.

• **`indices(x)`** retourne une séquence contenant tous les indices des éléments contenus dans **`x`**.

• **`replicate(x, n)`**, où **`x`** est un scalaire, retourne une séquence contenant **`n`** copies de **`x`**.

• **`range(n)`** retourne une séquence contenant **`[0, n)`**.

• **`bounded_range(a, b)`** retourne une séquence contenant **`[a, b)`**.

Listing 7 - Utilisation des tuples (ici avec une comparaison d'éléments). Notez les deux notations différentes pour les tuples.

```
@cu
def compare(kv0, (k1, v1)):
    k0, v0 = kv0
    if k0 < k1:
      return kv0
    else:
      return k1, v1
```

Copperhead et le typage

Les programmes Python se basent sur le *duck-typing* : si ça marche comme un canard, si ça nage comme un canard et si ça fait coin-coin comme un canard, alors c'est un canard. Les listes, par exemple, peuvent contenir des éléments de tous types. C'est plus facile à programmer mais l'introspection nécessaire ralentit significativement les codes. Ainsi, une implémentation Python native de **axpy** nécessiterait que soient vérifiés tous les éléments des vecteurs d'entrée, ne serait-ce que pour savoir quel type d'addition leur appliquer. Pour contourner le problème, Copperhead utilise l'inférence de type. Le runtime applique un *type-check* à toutes les procédures marquées du décorateur **`@cu`**. Celles qui ne passent pas l'examen d'inférence sont rejetées.

Le système de types de Copperhead est assez simple : il se limite aux scalaires, aux tuples et aux séquences. Les classes et les autres types définis par l'utilisateur ne sont pas supportés.

Mais rappelez-vous, le système interagit avec la bibliothèque **numpy**. C'est elle qui fournit les types numériques et les tableaux de types homogènes. Ainsi, Copperhead utilise les cinq types scalaires suivants :

• **`np.float32`** : nombre à virgule flottante 32-bit
• **`np.float64`** : nombre à virgule flottante 64-bit
• **`np.int32`** : nombre entier 32-bit
• **`np.int64`** : nombre entier 64-bit
• **`np.bool`** : booléen

Copperhead interprète également les types en utilisant les conventions **numpy** :

• **`bool`** est interprété comme un booléen
• **`float`** est interprété comme un nombre à virgule flottante 64-bit
• **`int`** est interprété comme un entier. Point important, Copperhead n'implémente pas de type entier infini comme peut l'être le **`int`** de Python.

En plus des cinq types scalaires de base, les programmes Copperhead peuvent utiliser les tuples de ces types, et ces tuples peuvent être imbriqués. Si Copperhead ne supporte pas les classes, les tuples peuvent servir de structures de données de nature différente. L'utilisation de tuples est même assez facile. Le **listing 7** compare par exemple l'élément **`k`** de deux paires key-value et retourne la paire ayant la plus petite clé. Notez que nous avons exprimé les tuples de deux façons différentes, pour montrer que ces deux notations sont supportées.

On accède aux éléments d'un tuple en le décompressant en plusieurs éléments, comme en Python. Par exemple,

```
k0, v0 = kv0
```

décompresse `kv0` en deux éléments. Cette décompression peut être effectuée via une instruction dédiée, comme le montre le listing 6, ou directement dans les paramètres d'une fonction.

Les tuples se créent en assignant plusieurs éléments à une variable ou en retournant des éléments multiples. Ainsi,

```
kv0 = k0, v0
```

compresse deux éléments en un tuple, et

```
return x, y
```

retourne un tuple de deux éléments.

Contrairement à Python standard, Copperhead ne supporte pas l'accès dynamique aléatoire au tuples via l'opérateur `[]`. Les tuples étant typés de façon hétérogène, l'accès aléatoire nécessiterait des types dynamiques, non supportés par Copperhead.

En complément des scalaires et les tuples, les programmes Copperhead utilisent aussi des séquences. Elles sont identiques aux listes Python, à ceci près que, comme les tableaux **numpy**, elles doivent contenir des types homogènes. Ces séquences peuvent être indexées en utilisant l'opérateur `[]`, mais uniquement en lecture. Les éléments ne peuvent être assignés via `[]` dans la mesure où Copperhead ne supporte pas les effets de bord, comme nous l'avons déjà mentionné.

Lors des appels de fonction, les valeurs en entrée doivent êtres de types que le runtime Copperhead puisse comprendre. Ces types sont les suivants :

• les 5 types scalaires de **numpy**, ainsi que les trois types scalaires de Python que nous avons décrits précédemment.

• les listes Python de types homogènes, dans lesquelles chaque élément est soit un des types de scalaires décrits précédemment, soit un tuple de ces types.

• des tableaux **numpy** unidimensionnels dont chaque élément est un type scalaire supporté.

• les tuples des cinq types scalaires et des types de séquence.

• `copperhead.cuarray`. Ce type séquence, renvoyé par les programmes Copperhead, gère la mémoire automatiquement dans les espaces mémoires hétérogènes. En prime, sachant qu'il transfère les données en mode "lazy" (c'est-à-dire en différé), il préserve la bande passante.

C'est le runtime qui convertit tous les types de séquences en types `copperhead.cuarray` avant que soit appelées les fonctions Copperhead. Pourquoi ? Pour préparer les données à l'exécution GPU, si nécessaire. Le runtime peut utiliser les listes Python et les tableaux **numpy**, mais les conversions sont coûteuses si elles sont répétitives. Par conséquent, dès lors que les fonctions Copperhead sont appelées à partir d'une boucle depuis l'interpréteur Python, il est recommandé de construire explicitement les types `copperhead.cuarray`, pour éviter les conversions inutiles.

Sachez enfin que les littéraux Copperhead sont faiblement typés : un entier littéral sera de type `np.int64` et un réel de type `np.float64`, à moins qu'ils ne soient utilisés dans un contexte de typage alternatif.

On peut utiliser des fonctions en tant que valeurs dans Copperhead, mais les lambdas ne peuvent actuellement pas sortir de la portée dans laquelle on les a définies. En pratique, cela revient à dire qu'un programme Copperhead ne peut accepter une fonction en tant qu'argument lorsqu'il est appelé à partir de Python, ni produire une fonction et la renvoyer à Python. Il se pourrait toutefois que cette restriction soit bientôt supprimée…

Un runtime intelligent

C'est au programmeur de déterminer si l'exécution aura lieu sur le CPU ou sur le GPU. Comment ? En utilisant les instructions `with`, comme au listing 7. A charge pour le runtime Copperhead de migrer les données en différé entre les espaces mémoires si nécessaire. Par exemple, lorsqu'on appelle une fonction Copperhead sur le GPU, le runtime vérifie que les données sont sur le GPU avant d'exécuter la fonction. En fin de course, le résultat se trouvera également en mémoire GPU. Si une fonction utilise ce résultat en entrée, et si elle est invoquée sur le même GPU, les

données ne seront pas déplacées. Si elle est invoquée à un autre endroit, par exemple sur le CPU, les données seront déplacées vers le CPU avant que la fonction soit exécutée.

Les valeurs renvoyées par les procédures Copperhead sont des `futures`. Dans certains cas, l'interpréteur Python reprend immédiatement le contrôle du programme cependant que le calcul est effectué en dehors de lui. Une synchronisation peut également avoir lieu lors du déréférencement d'un objet `copperhead.cuarray` depuis l'interpréteur Python. Les appels successifs aux fonctions Copperhead attendent toujours que les appels précédents soient terminés avant de démarrer, pour préserver l'ordre logique du code.

Le runtime Copperhead invoque automatiquement les compilateurs pour créer des binaires des fonctions Copperhead. Ces binaires sont persistés en cache dans un répertoire `__pycache__` qui contient les extensions Python pour toutes les fonctions Copperhead appelées à partir du module, ainsi que le code source C++/CUDA généré pour la fonction. Lorsqu'une fonction a été compilée et mise en cache, la latence de l'appel depuis l'interpréteur Python est généralement comprise entre 10 et 100 microsecondes, selon le nombre d'arguments.

Notez pour finir que le compilateur Copperhead fusionne de façon intensive les opérations sur les primitives, afin d'optimiser autant que possible l'utilisation de la bande passante. Il génère du code Thrust pour implémenter les calculs, la sortie du compilateur restant disponible dans le répertoire `__pycache__` après que le programme ait été exécuté.

De ce qui précède, il ressort que Copperhead allie élégamment la productivité de Python et l'efficacité de la programmation parallèle multiplateformes. Conçu à partie de Thrust, il libère le développeur des contraintes liées aux itérateurs complexes de C++. Le sujet vous intéresse ? Copperhead vous paraît convaincant ? Faites-le nous savoir, et si tel est le cas, nous lui consacrerons un article avec des exemples applicatifs complets.

Bons développements ! ■

Hervé Di Rosa - Good (2005)

Abonnez-vous*

[gratuitement]

Recevez chaque mois le seul magazine exclusivement dédié aux environnements HPC et Big Data...

...et profitez des contenus Web qui vous sont réservés

* sur www.hpcmagazine.fr