

Impetus-Afea, pionnier du calcul GPU

Afin d'exploiter une nouvelle génération d'algorithmes, le Bureau d'Ingénierie a choisi de porter sur processeur graphique une grande partie de son code de calculs

L'éditeur Impetus-Afea est constitué de trois petites sociétés indépendantes situées en Suède, en Norvège et en France, qui développent en commun un solveur par éléments finis avec des accords de type consortium. Leur objectif depuis leur création en 2007 est d'écrire un nouveau code plus robuste et plus précis. Ce solveur permet par exemple de modéliser les impacts sur des réservoirs d'avions, qui sont des pièces composites très critiques mais dont les



Jérôme Limido, ingénieur chez Impetus-Afea France, a présenté à la dernière conférence de l'association Nafems, « la modélisation du coup de bélier hydrodynamique sur structure composite », une simulation accélérée par GPU.

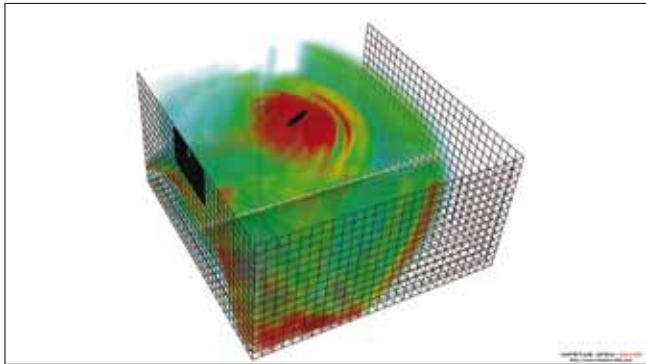
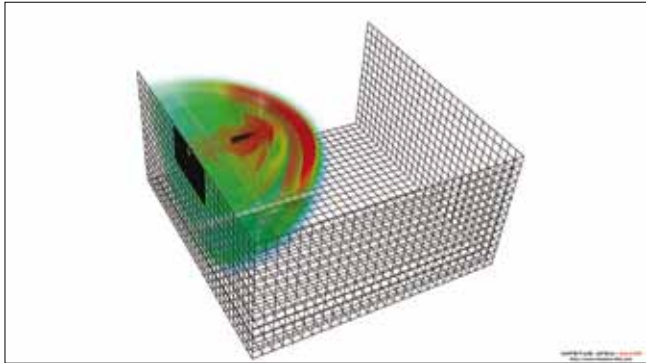
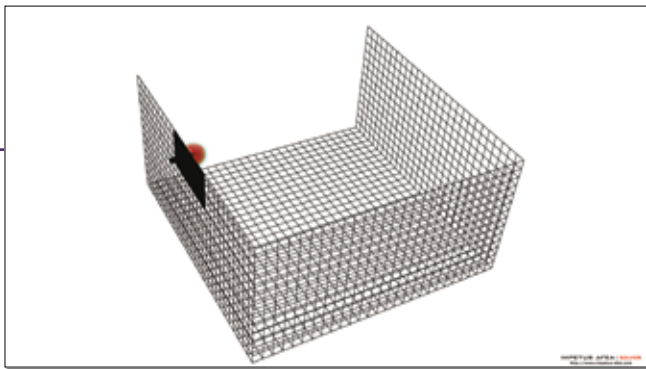
tests de solidité en réel sont coûteux. Positionné sur les marchés civils et militaires et solidement implanté dans le milieu universitaire, Impetus Afea est aussi en train de nouer des liens privilégiés avec de grosses entités industrielles dans le cadre de projets collaboratifs.

Une simulation gourmande en calculs

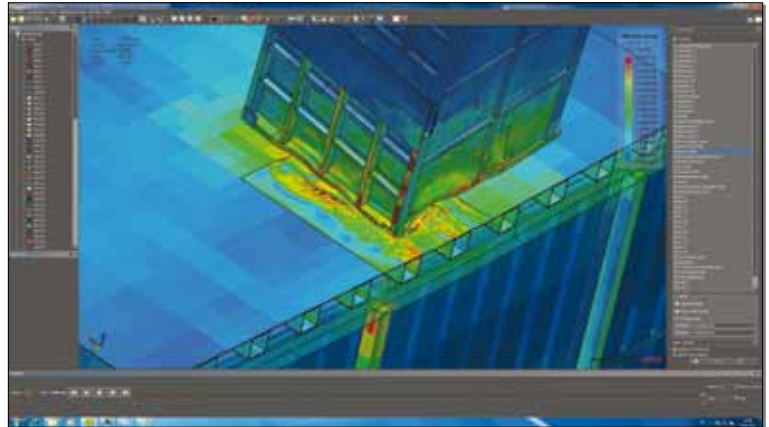
La modélisation numérique classique de l'interaction entre un fluide et une structure déformable exploite un maillage pour décrire la surface du fluide, avec une qualité de résultats fonction des dimensions des mailles. En raffinant le maillage de manière adaptative, la précision du résultat peut être améliorée mais des fuites peuvent se produire dans le modèle, l'énergie n'est plus conservée et les temps de calculs explosent. Ce type de solveurs éléments finis est usuellement développé en Fortran, et parallélisé avec MPI/OpenMP. « Nous développons une méthode particulière de représentation du fluide 3D plus récente dite sans maillage ou SPH pour smooth particle hydrodynamics. », avance

Jérôme Limido, ingénieur chez Impetus-Afea France. Si les éléments finis sont développés au sein de la structure suédoise d'Impetus et le post-processeur en Norvège, c'est la structure française qui développe la méthode SPH, le couplage et des modèles de comportement.

Les travers classiques de la méthode SPH sont des problèmes de précision et des temps de calcul élevés, d'où le besoin de paralléliser les algorithmes. Avec l'émergence en 2007 du calcul sur processeurs graphiques (GPU), la jeune structure a choisi de paralléliser son code avec le langage Cuda de nVidia. Quasiment tout est porté à l'heure actuelle sur GPU et les résultats sur des cas industriels sont concluants. « Le GPU est très adapté aux calculs SPH qui se parallélisent naturellement, et qui gagne en rapidité un facteur allant de 10 à 100. Nous arrivons même à extraire les champs de pression précis du SPH », se félicite Jérôme Limido. Pour les éléments finis en revanche, le gain est d'environ 10 et la parallélisation des algorithmes de contact est encore plus ardue. Leur priorité reste la robustesse et la précision du résultat, néanmoins le calcul doit être suffisamment rapide



Evolution de l'onde choc au cours du temps lors de la perforation d'une balle dans un réservoir déformable partiellement rempli (vue en coupe) à 0,01 ms, 0,2 ms et 0,4 ms.



Postprocesseur en haute résolution de la méthode des éléments finis sur GPU, sur une simulation d'explosion dans un container de transport d'une plateforme pétrolière.

pour être utilisable industriellement. Typiquement les calculs sont lancés de nuit sur la station de travail et doivent durer douze heures maximum.

L'évolution rapide d'une jeune technologie

Les langages de développement sur GPU ne sont pas légion et Impetus s'est porté sur le plus efficace, Cuda de nVidia, face à son concurrent OpenCL jugé moins mature. La technologie qui était encore un peu brute de forme à exploiter il y a peu, a fortement gagné en souplesse, avec moins de verrous et un debuggage facilité et plus convivial. Côté puissance, il y a deux ans, il était possible de gérer un million de particules en SPH. Désormais, avec le GPU sur cartes Tesla, on

peut aller jusqu'à vingt millions de particules dans le même temps de traitement !

La majeure partie du code est écrite en double, en Fortran/C++ et en Cuda, sans vraiment d'automatisme pour passer d'une version à l'autre. Par ailleurs, les ingénieurs d'Impetus ont démarré sur Linux, mais migrent de plus en plus sur Windows, en pratique Win server 64 bits. Des calculs effectués sur de gros modèles sur une semaine ont heureusement confirmé la stabilité de la configuration matérielle GPU exploitée dans un environnement Windows.

Le développement sur GPU demande un investissement, il faut être très précis pour optimiser le code et savoir gérer les transferts CPU et GPU ainsi que les mémoires partagées. Il

peut s'avérer rentable de porter sur GPU une partie du code qui se parallélise pourtant mal, simplement afin d'éviter les transferts de données entre le CPU et le GPU qui constituent un goulot d'étranglement. Impetus fait d'ailleurs partie des rares industriels à exploiter le GPU. « Certains ne sont pas convaincus suite à des tests incomplets, car par exemple la latence des transferts des protocoles masquent le gain en calculs. Le développement sur GPU demande une approche globale des algorithmes, ce qui est plus difficile avec un gros code déjà écrit. », soutient Jérôme Limido.

L'avenir est aux machines hybrides

Comme leur base utilisateurs est en construction, Impetus propose à ses clients de fournir le matériel avec le logiciel, ce qui garantit aussi une robustesse de la solution. Typiquement, ils vendent une station de travail Xeon 6 cœurs, avec

deux cartes graphiques nVidia Tesla type C2070 ou C2075 pour environ 10 000 euros. « Ce pack logiciel et matériel bénéficie donc d'un prix du matériel relativement peu cher, avec deux cartes Tesla dont l'une est dédiée au calcul des éléments finis et l'autre à la méthode sans maillage, complète Jean-Luc Lacome, Président d'Impetus-Afea France. Nos clients industriels poussent au multi-GPU, mais nous en sommes aux prémices, pas encore en phase industrielle. »

Le multi-GPU doit aussi permettre de lever la limite actuelle de 6 Go de mémoire RAM des cartes graphiques. Par ailleurs ajoute Jean-Luc Lacome, « par exemple pour l'algorithme de contact nous ne sommes pas encore fixés si le CPU ou le GPU est plus rapide. L'avenir pour nous réside plutôt dans les machines hybrides avec un CPU multi-cœurs associé à plusieurs GPU ». Une couche amont MPI permettrait de distribuer le calcul sur les CPU et sur les cartes GPU associées, chaque cœur pilotant une carte GPU. ■