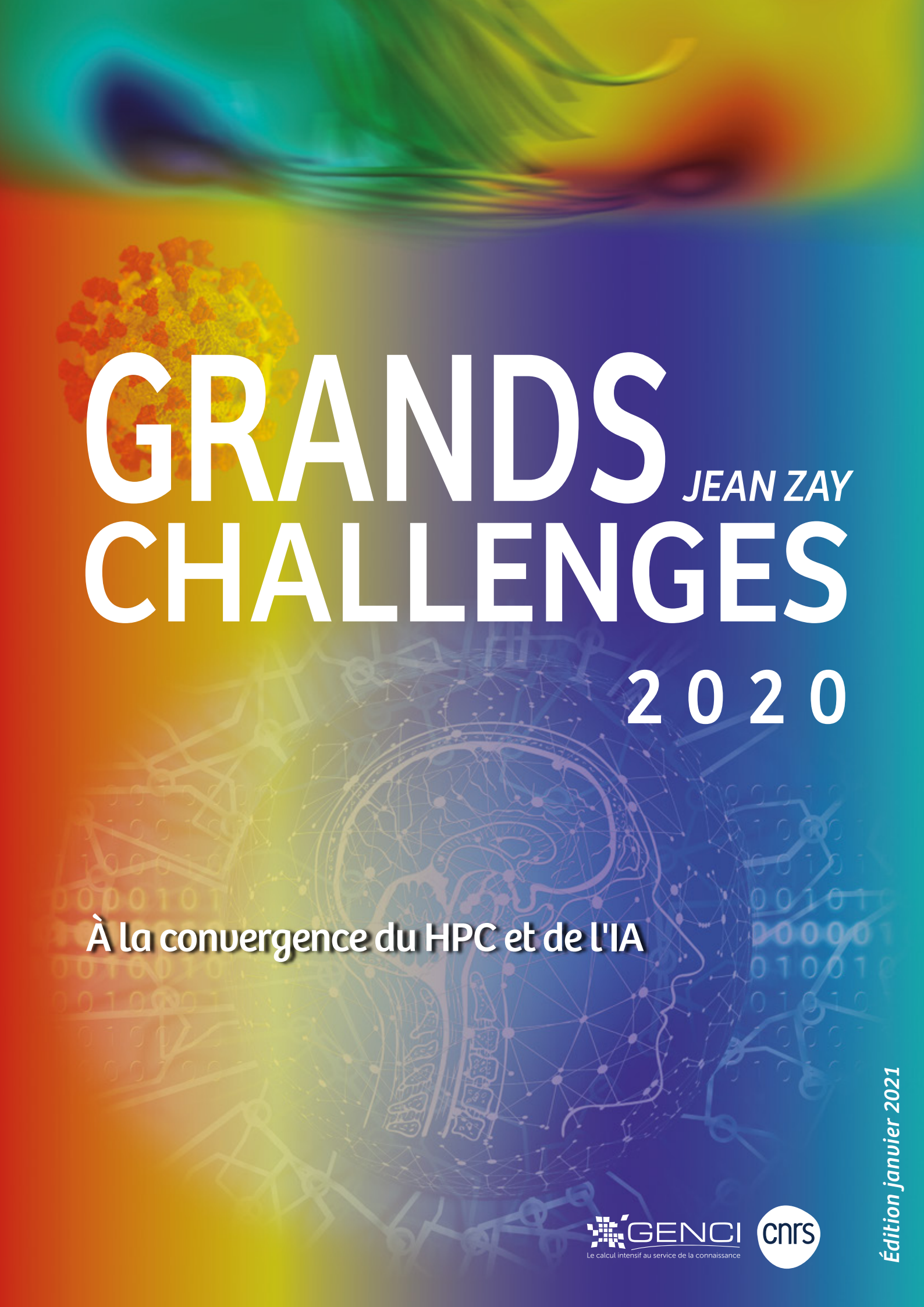


GRANDS CHALLENGES

JEAN ZAY

2020

À la convergence du HPC et de l'IA

GENCI
Le calcul intensif au service de la connaissance

cnrs

Édition janvier 2021

Sommaire

ÉDITORIAL

Directeur-adjoint de la Direction des données ouvertes de la recherche du CNRS, en charge des infrastructures numériques

Denis Veynante

Philippe Lavocat

Président-Directeur Général de GENCI

Éditorial

L'arrivée d'un nouveau supercalculateur dans un des trois centres nationaux est toujours un événement. Grâce aux progrès constants des technologies informatiques, la dernière machine acquise augmente considérablement la puissance de calcul que la Très Grande Infrastructure de Recherche GENCI peut offrir à ses utilisateurs, leur permettant des simulations numériques toujours plus ambitieuses dans tous les domaines de la science tels que climat, astrophysique, physique, biologie, chimie, matériaux, ingénierie… pour n'en citer que quelques-uns. Toutes ces disciplines contribuent à assurer un socle scientifique et technique considérable à la démarche nationale et européenne visant à répondre aux grands enjeux sociétaux tels que la santé, l'énergie, les transports, la sécurité ou l'aide à la décision.

Au-delà de cette augmentation de puissance, l'installation du supercalculateur « Jean Zay » à l'IDRIS, le centre de calcul intensif du CNRS situé sur le plateau de Saclay, est particulière à plusieurs titres. Pour le CNRS, elle coïncide simultanément avec les 80 ans de la création de l'organisme par Jean Zay et Jean Perrin, motivant sa dédicace au ministre d'alors de l'Éducation Nationale et des Beaux-Arts, et avec les 50 ans du centre de calcul à Orsay, d'abord le CIRCE (Centre inter-régional de calcul électronique), puis l'IDRIS (Institut du développement et des ressources en informatique scientifique), à partir de 1993. D'un point de vue scientifique, elle marque l'arrivée de nouveaux usagers en provenance de la communauté de la recherche en intelligence artificielle, jusque-là non-utilisatrice des moyens de GENCI, dans le cadre du plan *AI for humanity* voulu par le Président de la République, la faisant ainsi bénéficier de moyens de simulations toujours plus ambitieuses pour traiter des volumes de données sans cesse croissants. L'accueil de cette communauté, aux pratiques différentes de celles des utilisateurs traditionnels du calcul intensif et peu habituée aux contraintes d'accès d'un centre national, a conduit GENCI à faire évoluer ses modes d'accès aux ressources, dans le respect de règles de sécurité indispensables, vu la valeur et le caractère stratégique des matériels mis à disposition, vers plus de flexibilité et de souplesse dont profiteront bientôt tous les autres utilisateurs. Parallèlement, l'IDRIS a mis en place un environnement logiciel adapté et un support dédié de très haut niveau.

Le supercalculateur « Jean Zay » remplace à l'IDRIS deux machines IBM acquises par GENCI en 2012, l'une généraliste « Ada » (dotée de 332 nœuds larges pour un total de 10 624 cœurs de calcul et une puissance de 230 TFlop/s), l'autre « Turing », plus spécifique (dotée de 6 144 nœuds pour un total de 98 304 cœurs de calcul et une puissance de 1258 TFlop/s). Cette dernière, une BlueGene/Q, aux processeurs relativement lents mais en grand nombre et reliés par un excellent réseau d'interconnexion, était particulièrement adaptée aux calculs massivement parallèles nécessitant peu de mémoire par cœur, illustration de la volonté de diversité des architectures que GENCI propose à ses utilisateurs. A l'issue d'une procédure de dialogue compétitif menée par une équipe intégrée GENCI-CNRS/IDRIS incluant pour la première fois des spécialistes de l'intelligence artificielle, le choix s'est porté sur une machine de la gamme SGI de Hewlett Packard Enterprise (HPE). D'une puissance crête initiale de 16 PFlop/s et d'une architecture équilibrée (en termes de capacité de calcul, capacité mémoire, débits réseau et entrées-sorties), elle dispose de deux partitions. La première partition, généraliste, est composée de 61 120 cœurs de calcul Intel CascadeLake interconnectés par un réseau Intel Omni-Path (OPA). La seconde, accélérée, est constituée à ce jour de 2 696 GPU Nvidia V100 dont une partie est dédiée aux recherches en intelligence artificielle dans le cadre d'un financement spécifique. Cette partition, dite « convergée » est homogène, permettant d'adapter la « frontière » calcul intensif / intelligence artificielle au gré des besoins de l'exploitation. Cette machine est aussi la première du parc GENCI à offrir un premier niveau de stockage haut débit de 2,5 Po à plus de 0,5 To/s basé sur des mémoires flash et non des disques rotatifs. Deux opportunités, une donation de Facebook et l'anticipation de l'extension de la machine initialement prévue en 2021 ou 2022 et destinée à préserver son homogénéité, ont porté maintenant ses caractéristiques à 28 PFlop/s.

Soucieux de réduire l'impact environnemental des machines de calcul intensif, le CNRS et GENCI ont choisi une solution de refroidissement à l'eau chaude. L'eau, qui circule au plus près des processeurs, entre dans la machine à 32°C, à comparer aux 12°C d'une solution usuelle à l'eau froide, pour en ressortir à 42°C, limitant l'énergie nécessaire à la climatisation. Outre une réutilisation pour le chauffage du bâtiment, une partie des calories sera récupérée par l'Établissement public d'aménagement Paris-Saclay (EPAPS) pour contribuer au chauffage du campus.

L'installation d'une machine, c'est aussi associer un environnement technique : l'arrivée de ce supercalculateur a été accompagnée du renouvellement de l'environnement de stockage global et du remplacement des groupes électrogènes du centre. Ces trois opérations ont été financées par GENCI. Par ailleurs, l'utilisation d'accélérateurs de type GPU étant encore limitée dans le monde du calcul intensif, GENCI a accompagné l'installation de la machine de la mise en place d'un « contrat de progrès » dans le cadre duquel, les équipes du centre de calcul et de HPE ont travaillé au portage de 6 applications « phares », représentatives de la diversité des usages, avec l'aide de leurs développeurs. Couronnée de succès et unanimement saluée, cette idée novatrice sera reprise et étendue pour l'acquisition du calculateur qui remplacera prochainement la machine « Occigen » au CINES, Centre Informatique National de l'Enseignement Supérieur associé à la TGIR GENCI.

L'installation et la mise en service du supercalculateur « Jean Zay » ont été réalisées avec le soutien de GENCI par les équipes de l'IDRIS et de HPE démontrant leur savoir-faire dans la mise en œuvre dans les délais prévus d'une grande infrastructure de calcul. Une période particulière de trois mois, dédiée au démarrage, aux tests et à la mise au point d'une architecture complexe, a été réservée à quelques utilisateurs, acceptant par principe les aléas de conditions opérationnelles pas totalement stabilisées, en contrepartie d'une occasion unique d'accéder à des ressources de calculs pouvant aller jusqu'à l'intégralité du supercalculateur, pour la réalisation de simulations ou de traitements exceptionnels, appelés « Grands Challenges ». La présence pendant cette période de rodage, des équipes du centre de calcul (spécialistes systèmes et applicatifs), de GENCI et des experts du constructeur informatique, mobilisés pour résoudre les difficultés éventuelles de démarrage de la machine, est alors un atout supplémentaire : la collaboration étroite qui s'établit entre les équipes de chercheurs et ces spécialistes permet bien souvent d'optimiser la mise au point des logiciels de simulation et régler les problèmes de mise en production du supercalculateur. Ainsi sont réalisées des simulations numériques qui poussent aux limites non seulement les capacités de la machine mais aussi les logiciels de simulation eux-mêmes et l'ensemble de l'environnement informatique pour en exploiter les résultats. Elles sont ainsi essentielles pour franchir les changements d'échelle, aussi bien applicatifs que scientifiques, rendus possibles par les avancées technologiques dans le domaine du calcul intensif et de l'intelligence artificielle.

Après un appel à candidatures lancé par GENCI auprès des différents Comités Thématiques représentant les disciplines utilisatrices, 30 Grands Challenges (18 en simulation numérique et 12 en intelligence artificielle) ont donc été retenus et ont partagé le supercalculateur principalement entre juillet et octobre 2019. Côté recherche académique, la majorité des domaines utilisateurs était représentée : climat et environnement, écoulements non-réactifs, réactifs ou multiphasiques, astrophysique, imagerie médicale appliquée au traitement du cancer ou la simulation du fonctionnement du cerveau, physique des plasmas, informatique et algorithmique, biologie humaine, dynamique moléculaire et propriétés des nouveaux matériaux. Côté recherche ouverte industrielle, 3 grands challenges étaient portés par le CERFACS (couplant simulation numérique et intelligence artificielle), l'IMFT (associée à EDF) et l'IRMA (associé à la PME AxesSim). Enfin, coté intelligence artificielle, ils ont permis notamment le passage à l'échelle de l'apprentissage automatique de modèles dans le domaine de la vision appliqué à la médecine, du traitement automatique des langages ou la classification d'attaques malicieuses (dites adversiales) contre des modèles en IA, par exemple appliqués à la vision pour la reconnaissance de panneaux de signalisation.

Les résultats exceptionnels décrits dans ce numéro, parfois mêmes des « premières » mondiales, illustrent la diversité et l'intérêt du calcul numérique intensif pour soutenir la modélisation par simulation numérique, comme le développement de l'intelligence artificielle. A l'issue de cette phase, la machine « Jean Zay » a été ouverte à tous les utilisateurs le 1er novembre 2019. Son extension, opérationnelle depuis l'automne 2020, constitue elle-même l'opportunité de nouveaux « Grands Challenges ».

Cet apport de la simulation numérique et de l'intelligence artificielle à la recherche ouverte, activité stratégique au service de la compétitivité scientifique et économique, combiné à l'utilisation, pour la première fois dans le parc de machines de GENCI, d'une machine massivement accélérée, est une nouvelle étape vers l'utilisation des supercalculateurs européens de classe pré-exaflopique qui seront mis en service courant 2021, puis exaflopique vers les années 2023-2024 et qui permettront des percées scientifiques d'encore plus grande ampleur. Parallèlement, stimulés par cette réussite dans l'ouverture et l'élargissement des moyens de calcul, le CNRS et GENCI travaillent au déploiement d'une offre combinée calcul intensif / traitement de données massives appuyée sur les deux centres de calcul d'envergure nationale du CNRS, IDRIS et CC-IN2P3, à destination des grandes infrastructures de recherche, une communauté encore absente des utilisateurs des ressources offertes par GENCI. Nul doute que cette action donnera lieu à des nouveaux projets toujours plus passionnants.

Pour finir, au-delà du projet et de la prouesse technique, il faudra retenir de cette aventure, le rapprochement fertile, toujours en cours, entre deux communautés qui ne se connaissaient pas ou peu mais qui travaillent désormais ensemble : 350 projets relatifs à l'intelligence artificielle ont vu le jour en une année dont certains mixent déjà calcul intensif et IA pour de la co-simulation ou du post-traitement des données.

Le CNRS et GENCI sont très heureux et très fiers de s'associer pour féliciter et remercier chaleureusement tous les acteurs de cette très belle opération au service de la recherche, génératrice de nouvelles connaissances extraordinaires.

Très bonne lecture !!!

L'IDRIS

L'IDRIS (Institut du développement et des ressources en informatique scientifique), fondé en novembre 1993, est le centre national du CNRS pour le calcul numérique intensif de très haute performance (HPC) et l'intelligence artificielle (IA), au service des communautés scientifiques de la recherche publique ou privée (sous condition de recherche ouverte avec publication des résultats), tributaires de l'informatique extrême.

À la fois centre de ressources informatiques et pôle de compétences en HPC et IA, l'IDRIS (www.idris.fr) est une unité d'appui à la recherche du CNRS rattachée administrativement à l'Institut des sciences de l'information et de leurs interactions (INS2I), mais dont la vocation à l'intérieur du CNRS est pluridisciplinaire. Les modalités de fonctionnement de l'IDRIS sont proches de celles des très grands équipements scientifiques, tout en conservant une flexibilité et une adaptabilité remarquable, comme cela fut le cas durant la crise sanitaire de la Covid-19.

L'objectif principal assigné à l'IDRIS est de contribuer aussi efficacement que possible à l'excellence de la recherche scientifique dans le domaine de la modélisation, du HPC et de l'IA.

Pour ce faire, l'IDRIS intervient à deux niveaux :

- Comme structure de services, par la mise en place et l'exploitation d'un environnement de calcul intensif d'avant-garde, diversifié, polyvalent et évolutif, adapté aux très grands défis scientifiques dans les domaines de la simulation numérique et de l'IA. Cet environnement englobe une interface performante de support aux utilisateurs, qui offre des services à très forte valeur ajoutée. Ainsi, l'IDRIS ne se limite pas seulement au conseil et à la formation mais s'implique également dans le développement et l'optimisation des codes scientifiques.
- Comme agent de transfert de technologies, de la recherche et du développement en informatique vers les infrastructures nationales de calcul de haute performance. Situé à l'intersection de la science (la simulation numérique) et de la technologie (l'informatique scientifique) et très proche des utilisateurs scientifiques, l'IDRIS se trouve dans une situation privilégiée pour l'intégration progressive des nouvelles technologies dans le système national de la recherche scientifique. Cette activité s'est traduite, dans les années 90, par une contribution importante à la diffusion du calcul parallèle. L'IDRIS poursuit aujourd'hui ce type d'actions pour favoriser le passage au parallélisme massif associé à l'utilisation des accélérateurs matériels, notamment de type GPU. Ces évolutions représentent des enjeux majeurs pour les années à venir dans tout ce qui relève des domaines du HPC et de l'IA.

L'IDRIS, structuré en sept équipes (Système/Exploitation/Infrastructure, Réseau, Sécurité, Support aux Utilisateurs HPC, Support aux Utilisateurs IA, Administration et Communication) pour un effectif de 29 permanents, a reçu en 2020 un cristal collectif CNRS pour le projet Jean Zay.

GENCI

Créé en 2007 par les pouvoirs publics, GENCI a pour mission de démocratiser l'usage de la simulation numérique par le calcul haute performance associé à l'usage de l'intelligence artificielle, pour soutenir la compétitivité scientifique et industrielle française.
GENCI est une Très Grande Infrastructure de Recherche (TGIR), organisée en société civile détenue à

- 49 % par l'État représenté par le Ministère en charge de l'Enseignement supérieur, de la Recherche et de l'Innovation,
- 20 % par le CEA,
- 20 % par le CNRS,
- 10 % par les Universités représentées par la Conférence des Présidents d'Université,
- 1 % par Inria.

GENCI met à disposition des communautés de recherche académiques et industrielles, des moyens de calcul et de traitement de données massives performants et répond à trois missions :

- mettre en œuvre la stratégie nationale d'équipement en moyens de calcul à haute performance et de traitement/stockage de données massives au bénéfice de la recherche scientifique française en lien avec les trois centres nationaux de calcul de ses Associés,
- participer et soutenir pro-activement la réalisation d'un écosystème intégré du calcul à haute performance associé à l'intelligence artificielle à l'échelle européenne,
- promouvoir la simulation numérique par le calcul intensif auprès de la recherche ouverte académique et industrielle.

GENCI dispose d'un budget annuel de 39 millions d'euros.

Le supercalculateur Jean Zay

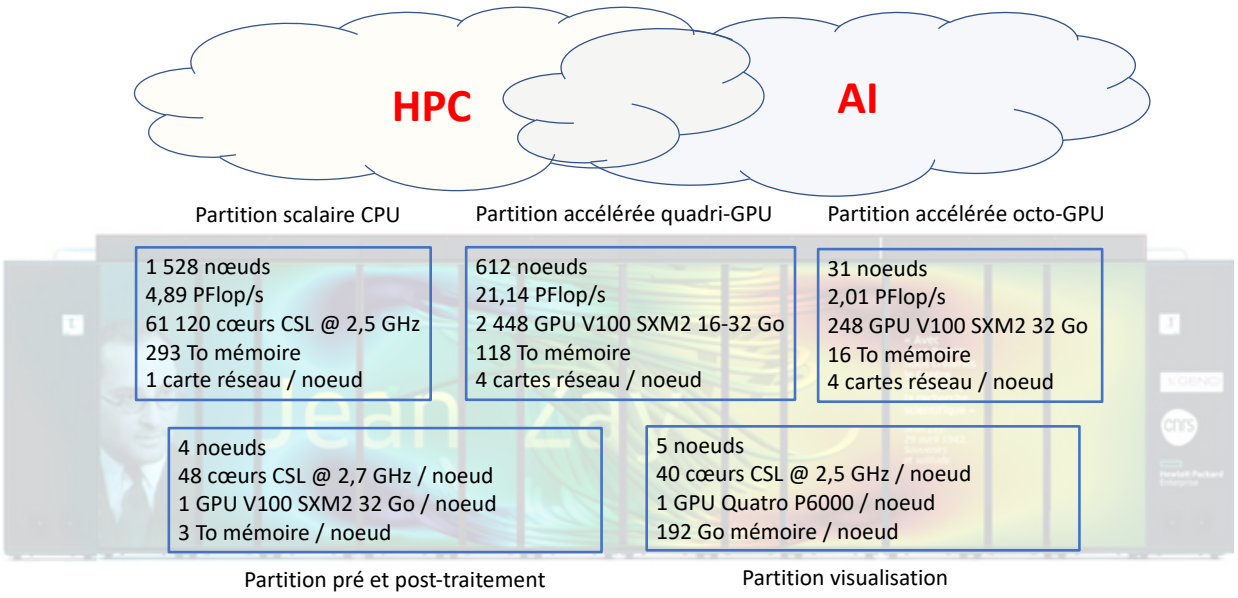
La collaboration entre l'IDRIS et GENCI pour le remplacement des anciennes configurations de calcul Ada et Turing a été initiée dès le début de l'année 2017 et a abouti début 2019, par l'acquisition auprès de la compagnie Hewlett-Packard Entreprise (HPE), du premier supercalculateur hybride accéléré Tier1 Français.

Ce supercalculateur de nouvelle génération, d'une puissance crête initiale de 16 PFlop/s (16 millions de milliards d'opérations en virgule flottante par seconde), a été installé à l'IDRIS au cours du premier semestre 2019. Cette configuration a depuis été étendue à 28 PFlop/s au printemps 2020.

Les caractéristiques techniques détaillées ci-dessous en font une architecture particulièrement bien équilibrée et aussi bien adaptée aux problématiques du HPC qu'à celles de l'IA :

- 86 344 cœurs de calcul Intel Cascade Lake cadencés à 2,5 GHz,
- 1 292 GPU NVIDIA V100 avec 32 Go de mémoire HBM2 (phase 1),
- 1 404 GPU NVIDIA V100 avec 16 Go de mémoire HBM2 (extension),
- 427 To de mémoire totale,
- Un réseau d'interconnexion Intel Omni-Path à 100 Gb/s avec une topologie de type *Enhanced-Hypercube 8D,*
- Un système de stockage de premier niveau en technologie Full Flash de 2,2 Po, accessible avec une bande passante supérieure à 500 Go/s,
- Un système de stockage capacitif de second niveau en technologie disques rotatifs de 35 Po de volumétrie accessible avec une bande passante supérieure à 150 Go/s.

Jean Zay est organisé en trois partitions de calcul, une partition dédiée au pré et post-traitement et une dernière partition pour la visualisation.



ENERGETICS : *ocEaN atmosphERe couplinG from Eddy To basIn sCaleS*

Jean-Marc Molines, IR CNRS à l'IGE,
Sébastien Masson, CNAP au LOCEAN


J.M. Molines


Sébastien Masson

Liens
IMMERSE :
http://immerse-ocean.eu
CONTACTS:
http://meom-group.github.io/projects/contacts/
MICRA :
https://ru.ambafrance.org/Projets-PHC-Kolmogorov-en-cours#MICRA

La circulation océanique globale joue un rôle essentiel dans la redistribution d'énergie au sein du système climatique : les interactions océan-atmosphère sont au cœur de la variabilité du climat. Les deux milieux échangent continuellement de l'énergie thermique et mécanique, et de la matière (eau douce, gaz, éléments chimiques, etc.) sur une vaste gamme d'échelles spatio-temporelles. Certains couplages se font à grande échelle, comme par exemple le phénomène El Nino dans le bassin océanique du Pacifique qui a un impact planétaire.
Cependant les courants océaniques sont contrôlés par des échelles spatiales bien plus fines que celle des bassins (Fig. 1). Aux échelles de l'ordre de la centaine de kilomètres, les tourbillons de mésoéchelle et les courants s'écoulant sur les talus ou les plateaux continentaux contribuent au transport de chaleur de l'équateur vers les pôles. A des échelles encore plus fines (quelques km) les processus de sous-mésoéchelle (ondes internes, filament de vorticité, ….) contrôlent d'importants échanges verticaux d'énergie et de matière dans l'océan superficiel.
Des études récentes montrent l'existence d'intenses échanges océan-atmosphère à méso-échelle et sous-méso-échelle dans toutes les régions du globe avec des phénomènes complexes de rétroaction positive (frontogénèse atmosphérique, amplification potentielle de tempêtes). Les rétroactions entre courants océaniques de surface et vents dans la couche limite atmosphérique par exemple affectent sensiblement la quantité d'énergie mécanique échangée entre les deux fluides.
Ces échanges mettent en jeu une gamme de processus de couplage dynamique et thermodynamique à fine échelle entre l'océan et l'atmosphère dont l'observation est très partielle. Leur étude est souvent abordée par des simulations couplées sur de petits domaines, ce qui ne permet d'appréhender ni leur variabilité spatiale ni la manifestation de leurs effets à plus grande échelle.
Notre projet ENERGETICS vise à la fois à mieux comprendre les processus de couplage air-mer à fine échelle et à guider la construction de systèmes de modélisation océanique de nouvelle génération, tant pour la recherche que pour la prévision opérationnelle. Ce projet met en œuvre une simulation couplée océan-atmosphère à très haute résolution à la fois dans l'océan et l'atmosphère afin de prendre en compte les interactions entre les deux milieux à ces fines échelles. Cette simulation couvre en outre un large secteur de l'océan Atlantique, afin d'étudier d'une part les caractéristiques régionales de ces processus et leurs impacts à l'échelle d'un grand bassin océanique.

Simulations Numérique ENERGETICS

Nous avons réalisé une simulation couplée océan-atmosphère-banquise de 3 ans (2004-2006) sur l'océan Atlantique et ses mers connexes (10°S-70°N, Golfe du Mexique, Mer des Caraïbes, Méditerranée, Mer Noire), avec une représentation explicite des fines échelles océaniques et atmosphériques.

Le modèle océan/banquise est une configuration du code NEMO avec une résolution de 1/36° (environ 3 km) et 150 niveaux verticaux (~2,8 milliards de points de grille). Le modèle atmosphère/surface continentale est une configuration du code WRF à une résolution de 1/12° (~9 km). Le couplage est réalisé par OASIS-MCT. Les modèles sont initialisés et contraints à leurs frontières par des réanalyses, océanique (GLORYS12) pour NEMO, et atmosphérique (ERA5) pour WRF. Le serveur XIOS (dont NEMO et WRF sont clients) est utilisé pour sauvegarder à une fréquence horaire les champs bidimensionnels de surface et journalière les champs tridimensionnels.
Les trois exécutables (NEMO, WRF et XIOS) sont lancés simultanément et communiquent via MPI2. Les études de scalabilité de NEMO+XIOS et WRF+XIOS pris séparément montrent un excellent passage à l'échelle sur la machine Jean-Zay (efficacité voisine de 1 jusqu'au-delà de 20 000 cœurs). Il faut cependant équilibrer les charges entre océan et atmosphère qui se donnent rendez-vous au moment du couplage (toutes les heures), exercice délicat qui dépend aussi du flux des I/O. Compte tenu de la charge de la machine (et pour limiter les attentes) nous avons utilisé 238 nœuds de calcul (5460 cœurs pour NEMO, 3942 pour WRF et 101 pour XIOS). Il faut alors ~5h30 pour réaliser un mois de simulation. Les étapes de réglage et de correction (en particulier de l'interface de couplage) ont été assez lourdes, en particulier à cause des temps d'attente.

Résultats

Les modèles ne sont contraints vers les réanalyses qu'à leurs frontières et leur évolution est totalement libre à l'intérieur du domaine. Les réanalyses océaniques et atmosphériques utilisées pour l'initialisation ont été produites de façon indépendante (par Mercator Ocean et ECMWF, respectivement) et la mise à l'équilibre de ce système couplé régional est donc un processus complexe que les trois ans de simulation ne permettent pas de finaliser. L'évolution du système couplé au cours de la simulation indique un refroidissement des couches de surfaces. Dans l'océan, ceci se remarque par une température de surface (SST) plus froide que celle de la réanalyse GLORYS12 sur la même période aux latitudes tropicales (équateur et subtropiques). Dans l'atmosphère, la température de l'air à 2 m est notablement plus froide que celle des réanalyses ECMWF sur l'ensemble du domaine à l'exception de l'Amérique du Sud ou l'on remarque un réchauffement des basses couches atmosphériques. Le refroidissement est plus marqué sur les continents (Afrique, Europe, Amérique du Nord) que sur les océans. On remarque également des vents plus forts que dans les réanalyses en mer des Caraïbes. Bien que le climat simulé ne représente pas fidèlement le climat actuel, nos simulations permettront d'étudier les interactions O/A à fine échelle. La Figure 1 montre l'omniprésence de structures de mésoéchelle (~100 km) dans l'océan, plus

Equipe Scientifique
Jean-Marc Molines, IR CNRS à l'IGE, et Sébastien Masson, CNAP au LOCEAN ont conjointement préparé le système couplé et réalisé les simulations, avec le soutien éclairant de Eric Maisonnave du CERFACS pour la préparation du coupleur.
Toute l'équipe impliquée dans la conception du projet ENERGETICS et sa valorisation scientifique remercie l'IDRIS et tout spécialement Pascal Voury et Rémi Lacroix dont l'aide a été précieuse et généreuse.

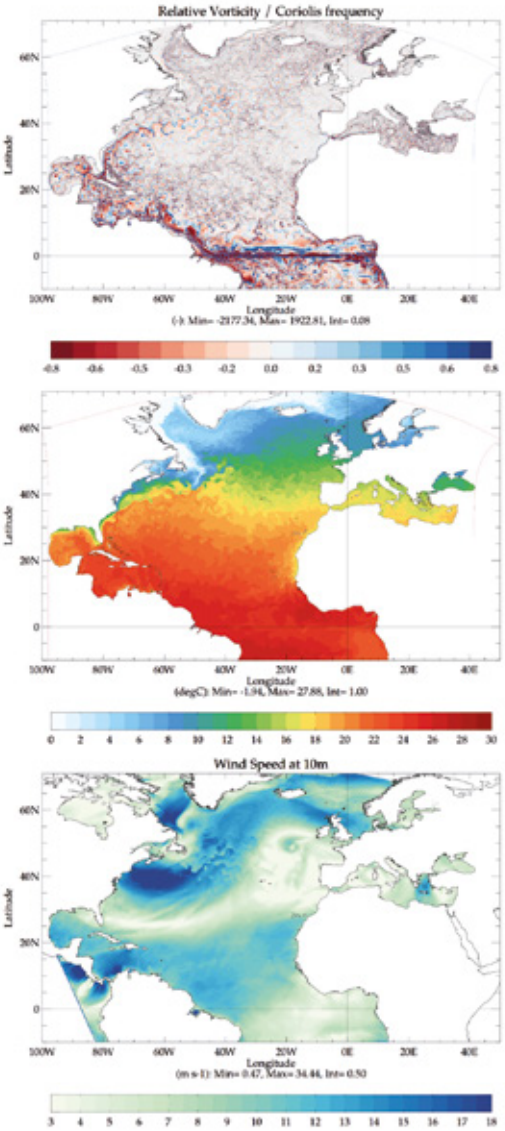


Figure 1 : *Moyennes journalières (19 décembre de la 3ème année d'intégration) de la vorticité relative (normalisée par la vorticité planétaire) à la surface de l'océan, illustrant l'omniprésence des fines échelles dans la dynamique océanique, de la température de surface de l'océan (°C) et de la vitesse du vent à 10 m (m.s-1). On distingue au large de Terre Neuve des échelles caractéristiques de la mésoéchelle océanique dans la SST comme dans le vent de surface.*

marquées en termes de gradients dans la région du Gulf Stream et de son extension au large de Terre Neuve. Côté atmosphère, le vent de surface montre également des structures comparables à celles de la mésoéchelle océanique dans cette région, indiquant des interactions entre les deux fluides à ces échelles, suffisamment fortes pour affecter la circulation à grande échelle.
Si l'on se focalise sur le Gulf-Stream, les champs de vent et de flux turbulents de chaleur montrent une forte cohérence avec les structures de courant et de température de surface aux échelles inférieures à 250 km (Figure 2). On observe même dans ces champs atmosphériques des structures filamentaires caractéristiques de la sub-mésoéchelle océanique. Notre analyse montre que ces structures se retrouvent également dans la profondeur de la couche limite atmosphérique.

Perspectives

La simulation ENERGETICS a produit une base de données unique par son emprise géographique, sa résolution et sa durée, décrivant à fines échelles spatio-temporelles l'évolution couplée des couches limites océanique et atmosphérique sur l'Atlantique Nord et l'ensemble des mers européennes (Mer Méditerranée, Mer Noire). Du fait de difficultés liées autant à la nouveauté du calculateur qu'à la mise au point de ce système couplé très haute résolution, nous n'avons réalisé qu'une partie du programme proposé. Néanmoins, la base de données va permettre aux différentes équipes engagées dans le projet d'étudier l'énergétique des interactions océan-atmosphère dans différentes situations, sur une gamme inédite d'échelles spatio-temporelles.

La simulation ENERGETICS a produit une base de données unique par son emprise géographique, sa résolution et sa durée

Ces données seront en particulier valorisées dans plusieurs projets: le projet H2020 IMMERSE dédié à l'amélioration de NEMO et notamment son couplage avec la CLA Albatros ; les projets MOPGA CONTaCTS et JPIOC MEDLEY dédiés notamment à l'étude de l'énergétique de surface océanique à fine échelle et son hétérogénéité ; le projet PHC Kolmogorov MICRA dédié à la dynamique fine échelle des masses d'eau subpolaires. Cette simulation répond à un objectif majeur de l'International Research Network DRAKKAR (France, Allemagne, UK) de favoriser le développement de simulations couplées-guidées dans cette communauté. Les conclusions de ce projet et ses applications en termes de forçage seront directement utilisables pour la préparation du prochain système global de prévision du Copernicus Marine Environment and Monitoring Services (résolution cible identique : 1/36°).

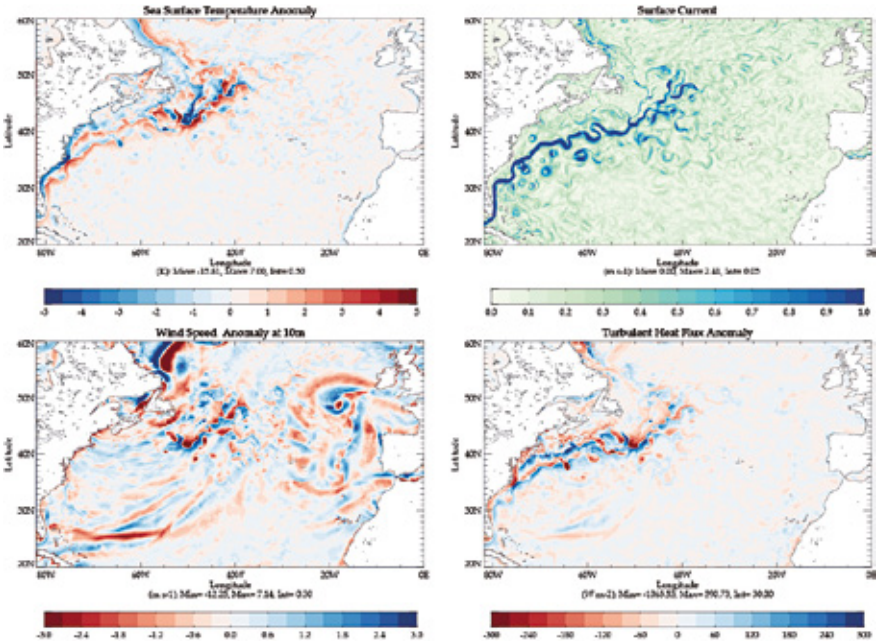
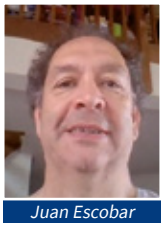


Figure 2 : *Champs océaniques et atmosphériques produits par le système couplé le 19 décembre 2006, filtrés spatialement par un filtre gaussien retirant les plus grandes échelles (> 250 km). Les flux turbulents d'énergie thermodynamique (positifs lorsque gain de chaleur pour l'atmosphère) et le vent de surface montrent une cohérence marquée avec les courants océanique et la SST à ces échelles.*

Giga-LES d'une supercellule par Méso-NH


J.P. Chaboureau


Juan Escobar


Philippe Wautelet

Jean-Pierre Chaboureau, Juan Escobar, Philippe Wautelet
Laboratoire d'Aérologie (Université de Toulouse et CNRS)

Les supercellules font partie des orages les plus violents et les plus destructeurs qui provoquent des pluies torrentielles, des rafales descendantes, de la grêle, des éclairs et des tornades. Elles ont un aspect circulaire dû au changement de la direction du vent avec l'altitude. La forte instabilité convective combinée au mouvement rotatif du vent entraîne un courant ascendant jusqu'à 15 km d'altitude renforcé par sa synchronisation avec le front de rafales descendantes. Mieux comprendre et prévoir ce type d'orage violent pour prévenir les dommages aux personnes et aux biens est un enjeu sociétal majeur.

Le cas simulé ici est le premier cas en conditions réalistes, c'est-à-dire issues d'analyses météorologiques.

Une supercellule présente une voûte sans écho dans les observations radar. Le pied de la voûte est composé d'un rideau d'embryons de grêle dans un courant ascendant tandis que son toit est la signature d'un mésocyclone, une circulation horizontale du vent. La production de gros grêlons s'explique ainsi par un concept de recyclage microphysique. Les embryons de glace circulent dans le mésocyclone et pénètrent plusieurs fois à l'intérieur du rideau d'embryons de grêle jusqu'à ce que la grêle soit suffisamment grosse pour se précipiter dans la région dite de la cascade de grêle. Les rafales descendantes s'expliquent par la fonte du rideau d'embryons de grêle. Cependant, ces explications soulèvent des questions. La théorie du rideau d'embryons avec recirculation est-elle suffisante pour expliquer ces grêlons gros et lourds ? L'évaporation de l'embryon de glace est-elle une raison raisonnable pour une explosion responsable de l'effondrement d'un nuage en quelques minutes avec des vents violents à la surface ?

Notre compréhension des phénomènes violents associés aux supercellules est en partie limitée par la capacité à résoudre explicitement à la fois l'orage dans son ensemble et les détails de ses composantes. La giga-LES (giga pour milliard et LES pour large-eddy simulation ou simulation des grands tourbillons) est un outil révolutionnaire car elle résout un large éventail d'échelles. Elle permet ainsi de représenter la supercellule avec une représentation explicite des tourbillons les plus énergétiques tout au long de sa propagation. Elle surpasse ainsi la simulation à maille kilométrique en représentant explicitement les échanges turbulents entre les cellules convectives individuelles et leur environnement. Jusqu'à présent, les giga-LES de supercellule sont rares et réalisées en conditions idéalisées. Le cas simulé ici est le premier cas en conditions réalistes, c'est-à-dire issues d'analyses météorologiques.

Le modèle météorologique utilisé est le code communautaire Méso-NH (http://mesonh.aero.obs-mip.fr/) développé conjointement par le Laboratoire d'Aérologie (Université de Toulouse et CNRS) et le CNRM (Météo-France et CNRS). Porté sur l'ensemble des supercalculateurs du GENCI, ce code est écrit en Fortran 90 et utilise MPI. La scalabilité du code a été montrée jusqu'à 2 000 000 cœurs virtuels (threads, Lac et al. 2018). Pour Jean Zay, le temps de communication pour les entrées sorties a été amélioré par l'emploi de la librairie Global Array. La configuration choisie pour la simulation d'un maillage cubique de 100 m est une grille de 5120 x 4800 points et 240 niveaux verticaux (5,9 milliard de points).

Cette simulation de 9 h temps réel a été réalisée avec 24 000 cœurs, a consommé 5 millions d'heures de calcul et a généré un volume de 300 To de données.

Le cas étudié est un cas observé à Sao Borja (sud du Brésil) pendant la campagne internationale de terrain RELAMPAGO en novembre 2018. La supercellule a produit des rafales de vent destructrices typiques d'une tornade de niveau 2 sur l'échelle de Fujita, des pluies intenses et un cumul de grêle de 4 cm. Elle a été bien échantillonnée par radar. Le radiosondage effectué dans la région montre une atmosphère avec une instabilité convective forte (2000 J kg^{-1} d'énergie potentielle de convection disponible) et une forte hélicité (126 m^2 s^{-2}). La figure 1 montre une vue tridimensionnelle des nuages à 15h20 sur le domaine entier de la simulation. Plusieurs systèmes orageux sont développés générant averses et rafales de vent. Celui situé au nord-ouest du domaine montre une couverture nuageuse circulaire typique des supercellules.

La figure 2 est une vue tridimensionnelle de la supercellule à pleine maturité à 17h00. Elle montre une structure en voûte avec des précipitations de la surface à 15 km d'altitude à son pied. La vitesse verticale est très élevée, jusqu'à plus de 200 km h^{-1}. Ce courant ascendant permet la formation et la suspension de ces précipitations au sein de l'orage. À environ 6 km d'altitude, des précipitations sont présentes en dehors du courant ascendant. Cette région fait partie d'une boucle du mésocyclone, la circulation horizontale du vent dans la supercellule. Un très fort contenu en précipitation peutdonc être simulé dans des conditions réalistes sans se trouver dans un courant ascendant. Il reste à déterminer le rôle et l'importance respective des différents processus d'échelle fine (dégagement de chaleur latente, mélange turbulent avec l'air ambiant, cisaillement de vent, refroidissement des rafales descendantes par évaporation de la pluie…).

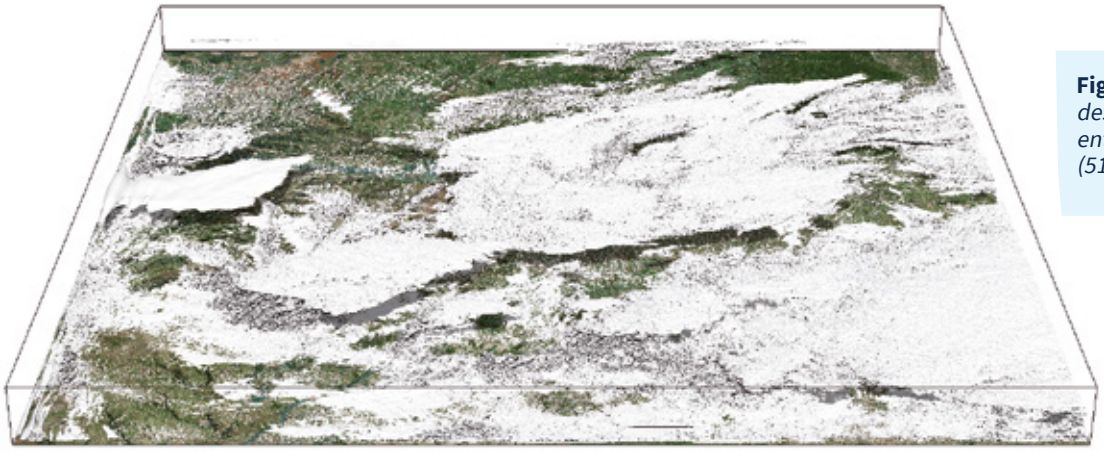


Figure 1 : *vue tridimensionnelle des nuages à 15h20 sur le domaine entier de la simulation (512 x 480 km²).*

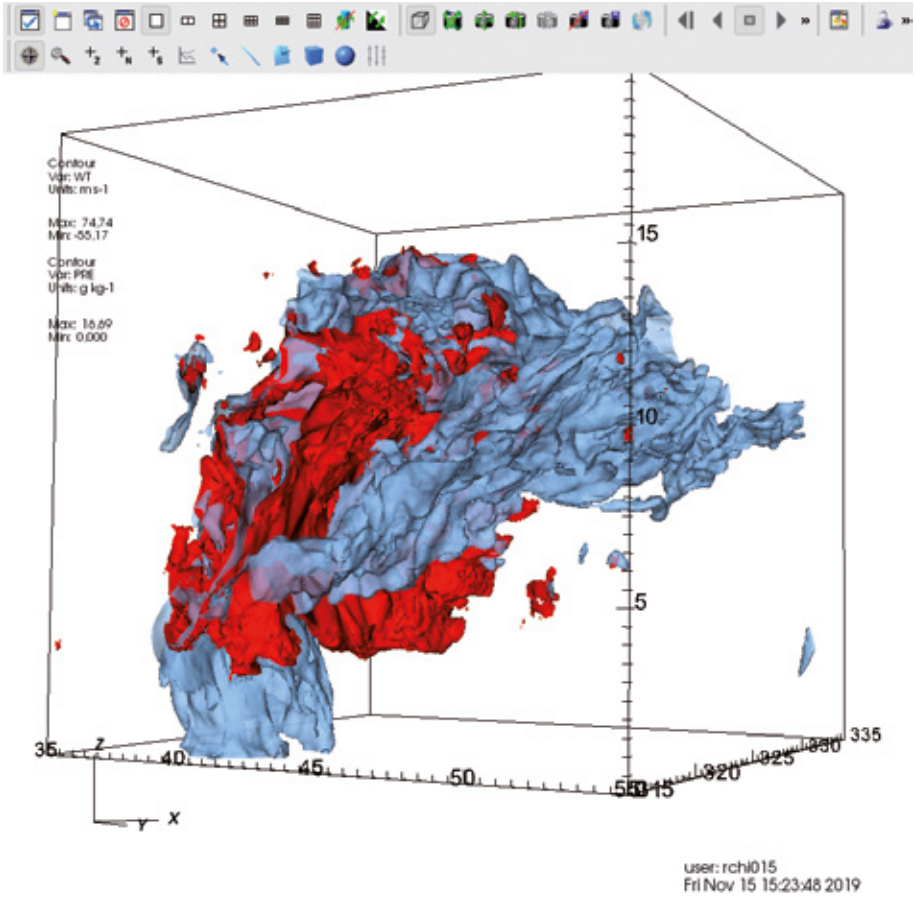


Figure 2 : *vue tridimensionnelle à travers la supercellule à 17h00. Le champ coloré en rouge représente la vitesse verticale supérieure à 20 m s $^{-1}$ (72 km h $^{-1}$) et celui en bleu les précipitations supérieures à 5 g kg $^{-1}$.*

Référence

Lac, C., J.-P. Chaboureau, V. Masson, J.-P. Pinty, P. Tulet, J. Escobar, M. Leriche, C. Barthe, B. Aouizerats, C. Augros, P. Aumond, F. Auguste, P. Bechtold, S. Berthet, S. Bieilli, F. Bosseur, O. Caumont, J.-M. Cohard, J. Colin, F. Couvreux, J. Cuxart, G. Delautier, T. Dauhut, V. Ducrocq, J.-B. Filippi, D. Gazen, O. Geoffroy, F. Gheusi, R. Honnert, J.-P. Lafore, C. Lebeaupin Brossier, Q. Libois, T. Lunet, C. Mari, T. Maric, P. Mascart, M. Mogé, G. Molinié, O. Nuissier, F. Pantillon, P. Peyrillé, J. Pergaud, E. Perraud, J. Pianezze, J.-L. Redelsperger, D. Ricard, E. Richard, S. Riette, Q. Rodier, R. Schoetter, L. Seyfried, J. Stein, K. Suhre, O. Thouron, S. Turner, A. Verrelle, B. Vié, F. Visentin, V. Vionnet, and P. Wautelet, Overview of the Meso-NH model version 5.4 and its applications, Geosci. Model Dev., 11, 1929-1969, 2018. https://doi.org/10.5194/gmd-11-1929-2018

Modélisation haute-résolution de la distribution des espèces végétales par apprentissage profond

Alexis Joly, Inria, LIRMM, univ. de Montpellier, CNRS
Maximilien Servajean, LIRMM, univ. de Montpellier, CNRS, AMIS, UPVM, Montpellier.
Benjamin Deneu, Inria, LIRMM, univ. de Montpellier, CNRS
Pierre Bonnet, CIRAD, AMAP, univ. de Montpellier
François Munoz, LiPhy, université de Grenoble-Rhône Alpes

On estime qu'il existe environ 391 000 espèces de plantes actuellement connues de la science et de nouvelles espèces de plantes sont encore découvertes et décrites chaque année. Cette diversité végétale a été l'un des éléments majeurs du développement de la civilisation humaine (alimentation, médecine, matériaux de construction, loisirs, gènes, etc.) et on sait qu'elle joue un rôle crucial dans le fonctionnement et la stabilité des écosystèmes. Cependant, notre connaissance des plantes au niveau des espèces n'en est encore qu'à ses débuts. Pour la grande majorité des espèces, nous n'avons aucune idée de leur rôle exact dans les éco-systèmes ou de leur utilisation potentielle par l'homme.

L'apprentissage de tels modèles sur des données haute-résolution et à l'échelle territoriale n'aurait pas pu s'envisager sans l'utilisation du supercalculateur Jean Zay

Même nos connaissances sur la répartition géographique et l'abondance des populations restent très limitées pour la plupart des espèces. La communauté internationale a déployé des efforts importants au cours des deux dernières décennies pour développer des initiatives mondiales, des plateformes numériques et des outils pour aider les biologistes à organiser, partager, visualiser et analyser les données sur la biodiversité. En parallèle, des projets de sciences participatives de grande ampleur ont vu le jour et permettent de collecter des observations à des échelles sans précédent. La plateforme Pl@ntNet, en particulier, a été la première au niveau mondial à mettre en œuvre une application mobile d'identification automatique des plantes. Celle-ci permet de collecter chaque année des dizaines de milliers d'observations de plantes un peu partout dans le monde. Maintenant que ces données massives ont été collectées, il est nécessaire de les analyser et c'est là le principal objectif des expérimentations de grande ampleur mises en œuvre grâce aux supercalculateurs du GENCI. Nous nous sommes en particulier intéressés à la modélisation de la distribution des espèces à l'échelle de la flore française et nord-américaine.
L'objectif est d'inférer la distribution spatiale d'un grand nombre d'espèces (typiquement plusieurs milliers) à partir d'un ensemble d'occurrences géo-localisées de ces espèces.

Comme illustré par la figure 1, ceci est généralement atteint par des approches de modélisation de niche environnementale, c'est à dire en prédisant la distribution dans l'espace géographique sur la base d'une représentation de leur répartition dans l'espace environnemental (= niche écologique). L'environnement est dans la plupart des cas représenté par les données climatiques (telles que la température et les précipitations), les données caractérisant le type de sol et la couverture terrestre, ou encore la proximité à l'eau.
Les modèles de distribution d'espèces ("Species Distribution Models" en anglais) sont devenus de plus en plus importants au cours des dernières décennies pour l'étude de la biodiversité, la macroécologie, l'écologie des communautés et l'écologie de la conservation. Une connaissance précise de la répartition spatiale des espèces et de leurs préférences écologiques est en effet d'une importance cruciale pour de nombreux domaines d'application tels que la planification de la conservation ou la gestion des paysages et des territoires. Cependant, les modèles actuels (cf. figure 1) sont créés à grande échelle avec une résolution spatiale assez grossière, de l'ordre de la centaine ou dizaine de kilomètres. Ce manque de précision spatiale est critique pour de nombreuses applications telles que la conservation et l'aménagement du territoire qui exigent souvent une résolution bien plus élevée. Dans ce travail, nous nous sommes intéressés à créer des modèles de distribution d'espèces à des échelles spatiales beaucoup plus fines (de l'ordre de la centaine ou dizaine de mètres). Nous avons pour cela dû mobiliser des données environnementales, satellitaires et aériennes de très haute-résolution (1 m par pixel) ainsi que des volumes d'observations de plantes sans précédent (plusieurs millions). Pour analyser ces données, nous avons également dû faire mettre en œuvre des modèles différents de ceux utilisés habituellement en écologie, à savoir des réseaux de neurones convolutionnels (cf. figure 2).
Ces derniers sont habituellement utilisés pour les applications de reconnaissance d'images et ont le grand avantage de pouvoir modéliser la structure spatiale de l'environnement en plus des facteurs classiques tels que température, altitude, type de sol, etc.

L'apprentissage de tels modèles sur des données haute résolution et à l'échelle territoriale nécessite cependant des ressources de calcul très importantes et cette étude n'aurait pas pu s'envisager sans l'utilisation du supercalculateur Jean Zay du Genci. Les paramètres du réseau de neurones convolutionnel nécessitent en particulier d'être optimisés sur des processeurs de type GPU (Graphical Processing Unit) qui sont présents en très grand nombre sur la machine Jean Zay (plus de 1000 GPU).
L'entrainement d'un seul modèle nécessite ainsi environ 2500 heures de calcul GPU et nous avons dû en entraîner près d'une dizaine pour sélectionner les meilleurs hyper-paramètres tels que l'architecture du réseau de neurones, le choix de la fonction d'optimisation ou bien encore la manière d'encoder les données d'entrée.
Grâce à ces calculs, nous disposons désormais d'un premier modèle haute-résolution couvrant près de 10,000 espèces en France et aux Etats-Unis. La prochaine étape sera maintenant d'interpréter les prédictions de ce modèle et d'estimer dans quelle mesure il pourrait être utilisé pour des applications concrètes de conservation ou d'aménagement du territoire.

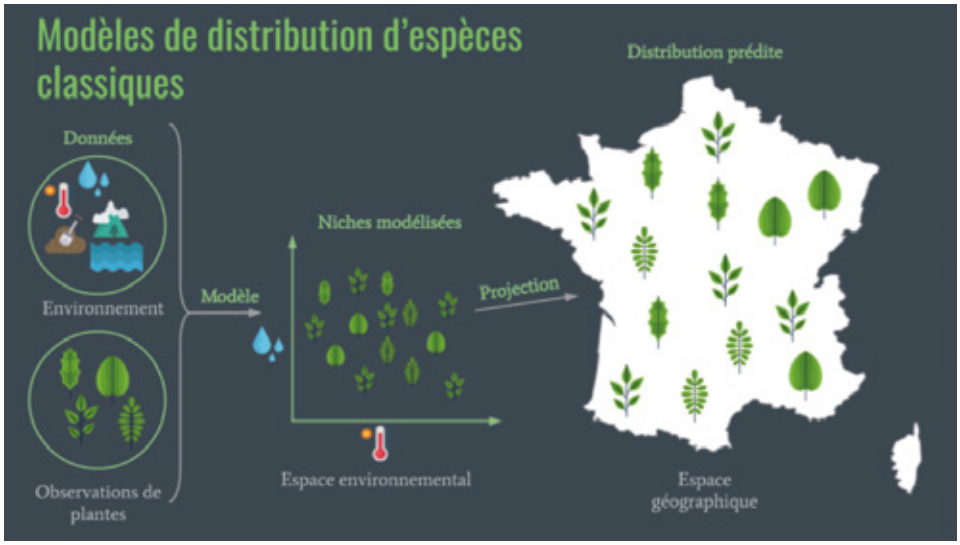


Figure 1 : *Illustration du processus de modélisation de la distribution des espèces par une approche classique*

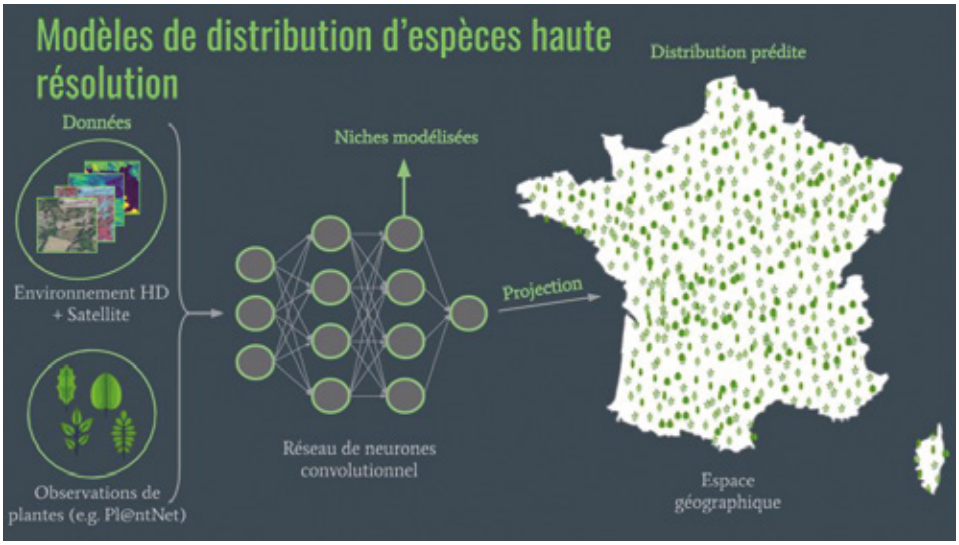


Figure 2 : *Illustration du processus de modélisation de la distribution des espèces par l'approche haute-résolution mise en œuvre grâce aux supercalculteurs du GENCI*

Turbulence induite par l'ascension d'un nuage de bulles

Anne Boulin, Jean-François Haquet, Florian Le Roy De Bonneville, Frédéric Risso, Rémi Zamansky

Florian Le Roy De Bonneville

Rémi Zamansky

Turbulence induite par l'ascension d'un nuage de bulles

Les écoulements à bulles font partie de la famille des écoulements polyphasiques dans lesquels des particules, solides, liquides ou gazeuses, sont dispersées dans un liquide. On retrouve ce type d'écoulement dans nombreux procédés industriels (lits fluidisés, colonnes à bulles ou d'extraction) ou naturels (vagues déferlantes, panaches volcaniques). On les rencontre aussi dans des situations accidentelles, comme la fusion du coeur d'une centrale nucléaire. L'injection de bulles dans un liquide favorise grandement les transferts et le mélange entre les deux phases.

Ces simulations en configuration homogène ont permis de valider notre modélisation et révéler les mécanismes physiques de production et de transfert d'énergie de la turbulence à bulles

Plusieurs facteurs vont influencer ces mécanismes : le nombre de bulles, leur taille, leur vitesse, la viscosité du fluide dans lequel elles évoluent, etc. C'est un système complexe dans lequel les bulles ont une influence sur le fluide mais le fluide a également une influence sur les bulles.

Dans notre étude nous nous intéressons à l'écoulement induit par l'ascension d'un essaim de bulles de grande taille (dont le nombre de Reynolds est de plusieurs centaines : typiquement des bulles l'air de 2.5 mm de diamètre dans de l'eau). Ce type d'écoulement est fortement marqué par les interactions entre les sillages des bulles et présente de nombreuses caractéristiques originales [1, 4, 2, 5], l'une des plus frappantes étant l'existence d'un comportement spectral singulier mettant en jeu un régime évoluant en puissance -3 du nombre d'onde. Nous visons une compréhension fine des mécanismes de transfert turbulent inter-échelle et de mélange dans ce type d'écoulement. Pour cela nous proposons de simuler l'écoulement avec une méthode Lagrangienne pour le suivi des bulles. Dans notre approche numérique, chaque bulle est modélisée comme une source volumique de quantité de mouvement répartie sur quelques éléments de maillage et donc les plus petites échelles de l'écoulement (c'est-à-dire des échelles beaucoup plus petites que le diamètre des bulles) ne sont pas finement résolues. Cela nous permet de simuler une fraction volumique élevée avec un grand nombre de bulles. Dans les travaux de [3], dans lesquels les bulles étaient considérées comme des sources fixes de quantité de mouvement, il a été montré que l'on pouvait reproduire l'interaction entre les sillages des bulles et obtenir les statistiques de l'écoulement en relativement bon accord avec les expériences malgré un maillage grossier.

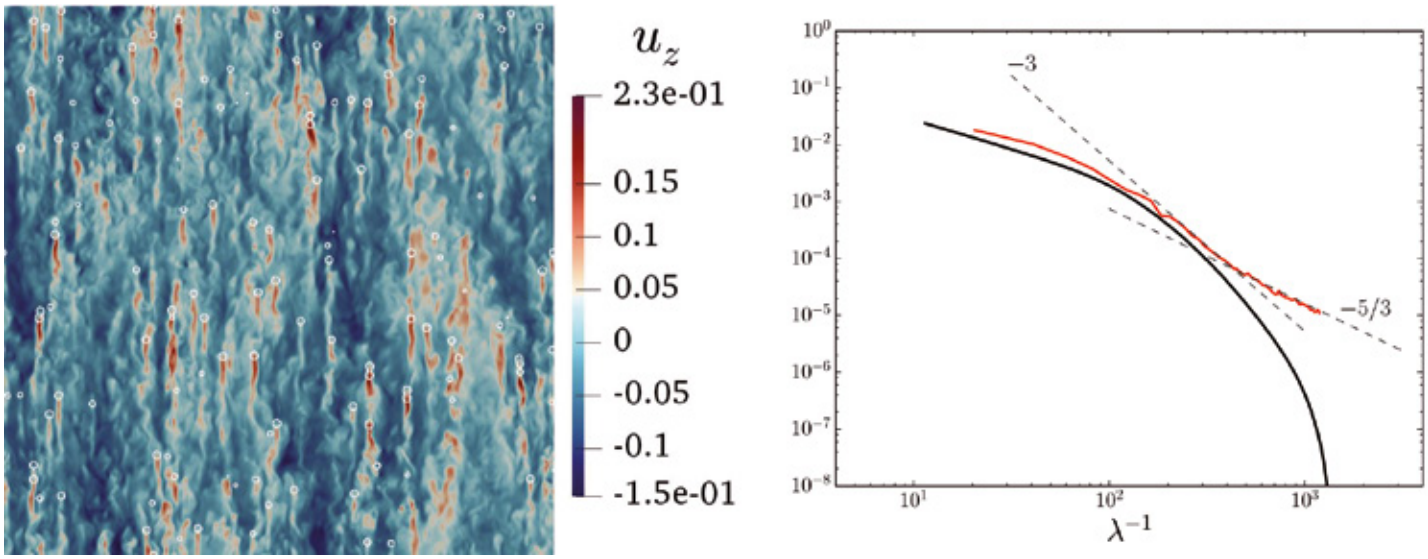
Dans notre travail, nous nous intéressons aux bulles qui peuvent avoir un mouvement relatif entre elles afin de pouvoir simuler l'évolution de la fraction volumique. Pour calculer la trajectoire de chaque bulle, il est nécessaire de calculer les forces hydrodynamiques exercées sur elle. Le nombre de Reynolds des bulles étant grand, il est indispensable de prendre en compte la perturbation de l'écoulement générée par la bulle elle-même afin d'annuler son effet dans le bilan des forces.

Nous avons établi un modèle pour déterminer cette perturbation de l'écoulement permettant ainsi de calculer de façon rigoureuse la force de trainée, de masse ajoutée ou de portance. Le modèle nécessite la connaissance de l'historique de la quantité de mouvement échangée entre une bulle et l'écoulement ainsi que sa position. Nous montrons qu'il est ainsi possible d'obtenir des simulations numériques de l'agitation induite par la montée d'un essaim de bulles en accord avec l'expérience. Ceci est illustré par la figure 1 ci-contre.

Sur la gauche, nous présentons un instantané de la vitesse verticale montrant l'interaction entre les sillages. A droite, nous comparons les spectres de vitesse verticale obtenus à partir des expériences de [4] (bulle d'air dans l'eau avec $d_b = 2.5$mm et $= 2.5\%$), avec les spectres obtenus à partir de notre simulation numérique pour des conditions similaires. Ces simulations en configuration homogène ont permis de valider notre modélisation et révéler les mécanismes physiques de production et de transfert d'énergie de la turbulence à bulles. Le modèle numérique va maintenant être appliqué à la simulation du bain de corium produit lors d'un accident de fusion du coeur d'une centrale nucléaire dans un but de prévention des risques.

Notre allocation sur le supercalculateur Jean Zay était de 1.5 Millions d'heures. La plupart de nos simulations étaient menées sur des domaines de calcul de ayant $2\,048^3$ mailles et nous avons utilisé jusqu'à 4 096 coeurs pour ces calculs. L'intérêt des Grands Challenges et des supercalculateurs est la possibilité de simuler de grands domaines avec un nombre de mailles conséquent. Dans notre cas, la caractérisation précise de l'agitation induite par les bulles nécessite un domaine de calcul suffisamment grand pour s'affranchir des effets de confinement et une résolution suffisamment fine pour capturer la dynamique aux petites échelles et nécessite donc une puissance de calcul considérable.

Figure 1 : *(Gauche) Champ instantané de la vitesse verticale du liquide (en m/s). Les cercles matérialisent les bulles. (Droite) Densité spectrale de puissance pour la vitesse du liquide (composante verticale dans la direction verticale). La courbe noire représente nos simulations numériques et la courbe rouge les expériences de [4].*



Bibliographie

[1] M. Lance and J. Bataille. Turbulence in the liquid phase of a uniform bubbly air–water flow. Journal of Fluid Mechanics, 222 :95–118, 1991.

[2] V. N. Prakash, J. Martínez Mercado, L. van Wijngaarden, E. Mancilla, Y. Tagawa, D. Lohse, and C. Sun. Energy spectra in turbulent bubbly flows. Journal of Fluid Mechanics, 791 : 174–190, 002 2016.

[3] G. Riboux, D. Legendre, and F. Risso. A model of bubble-induced turbulence based on large-scale wake interactions. Journal of Fluid Mechanics, 719 :362–387, 2013.

[4] G. Riboux, F. Risso, and D. Legendre. Experimental characterization of the agitation generated by bubbles rising at high reynolds number. J. Fluid Mech., 643 :509–539, 2010.

[5] Frédéric Risso. Agitation, mixing, and transfers induced by bubbles. Annual Review of Fluid Mechanics, 50 :25–48, 2018.

CO-simulatiOn Learning

V. Xing, C. Lapeyre, A. Misdariis, L. Labarrère, G. Staffelbach (CERFACS)


Victor Xing


Gabriel Staffelbach


Corentin Lapeyre

La simulation numérique est un outil indispensable dans le domaine de la combustion, qui représente la grande majorité de l'énergie produite dans le monde. Pour représenter la dynamique instationnaire d'une flamme dans un écoulement turbulent, on emploie généralement des simulations des grandes échelles (LES) utilisant des modèles de sous-maille pour tenir compte des plus petites échelles spatiales. En particulier, il est important de bien représenter les phénomènes physico-chimiques à l'échelle de l'épaisseur du front de flamme, environ un dixième de millimètre, souvent inférieure à la résolution du maillage.

L'implémentation et la validation de ce couplage massivement parallèle ont constitué un enjeu majeur de ce projet. Nous avons ainsi réussi à coupler jusqu'à 2560 instances AVBP (CPU) avec 64 CNN (GPU)

Une solution est donnée par le formalisme de flamme épaissie qui produit un front de flamme artificiellement épaissi sans affecter la vitesse de la flamme. Cette opération altère le plissement du front de flamme qui résulte de l'interaction flamme-turbulence. Or, le plissement modifie la surface du front, donc le taux de réaction total et la vitesse de propagation de la flamme.
Pour corriger ce problème, on introduit dans les équations une fonction d'efficacité qui contient l'information du plissement non résolu. Récemment, une nouvelle fonction d'efficacité basée sur un réseau de neurones convolutionnel (CNN) a été proposée [1]. Par rapport aux modèles existants qui s'appuient sur des arguments physiques, le CNN a l'avantage de pouvoir agréger de l'information contextuelle à plusieurs échelles.
Ce projet "Grands Challenges" visait à intégrer le CNN dans une simulation de grande ampleur en mettant en œuvre un couplage CPU-GPU entre le solveur numérique AVBP et le CNN. Les résultats de la simulation permettraient d'évaluer la performance à la volée du CNN sur une configuration complexe.
Nous avons réalisé la LES du brûleur à fente R3 étudié en simulation numérique directe (DNS) par Luca et al. [2]. Le CNN a été entraîné sur la configuration R2 qui a la même résolution spatiale mais des dimensions 8 fois plus petites. L'objectif était de montrer que le CNN peut fournir une fonction d'efficacité très précise pour une LES coûteuse, moyennant le coût plus raisonnable d'une LES à échelle réduite qui est utilisée pour l'entraînement. La base d'apprentissage du réseau est construite en appliquant un filtre spatial aux champs de variable de progrès et d'efficacité des solutions DNS instantanées puis en les interpolant sur un maillage 4 fois moins résolu. Le CNN apprend alors à transformer le champ de variable de progrès, qui est une grandeur résolue en LES, en un champ d'efficacité qui représente l'information non résolue du plissement de sous-maille (voir figure 1).

Dans cet exercice, dit a priori, la performance du CNN sur un jeu de test est excellente. Sa performance a posteriori, c'est-à-dire en étant intégré à la volée à un solveur numérique, est l'objet d'étude de ce projet.
L'écoulement est simulé à l'aide du code parallèle AVBP résolvant les équations de Navier-Stokes 3D instationnaires réactives en régime compressible. Le domaine, de dimensions 1,0 cm x 3,8 cm x 5,8 cm, est initialement rempli de gaz brûlés de la combustion du méthane à la température adiabatique de combustion 2 229 K. Un mélange méthane-air de richesse 0.7 à la température de 800 K est injecté dans la fente du brûleur à une vitesse de 100 m/s et une intensité turbulente de 10%. Le front de flamme est épaissi uniquement dans les zones réactives par un senseur de flamme dynamique. Le facteur d'épaississement est de 4, comme dans la base d'entraînement du CNN.
Le domaine est muni d'un maillage non structuré de 95 millions d'éléments. La zone proche des lèvres du brûleur est maillée finement car l'interaction entre le front de flamme initialement laminaire et la couche de mélange turbulente gouverne le comportement du reste de la flamme. Plus loin, on déraffine progressivement le maillage jusqu'à la zone centrale qui englobe la portion de flamme plissée en sous-maille. Comme la notion d'efficacité n'a pas d'intérêt en dehors des zones réactives, le CNN n'est utilisé que dans cette zone centrale.
Pour cette simulation couplée, AVBP utilise un maillage non structuré couvrant tout le domaine de calcul et partitionné entre chaque processeur CPU. Le calcul de l'efficacité dans la zone centrale est partitionné entre plusieurs instances du CNN, chacune localisée sur son propre GPU et possédant un maillage structuré correspondant à sa région de la zone. La bibliothèque de couplage CWIPI, développée par l'ONERA, assure les échanges des champs interpolés à travers une interface géométrique non conforme. Elle dispose d'interfaces Fortran et Python permettant de distribuer le couplage dans chaque code. Les communications sont réalisées par le protocole MPI toutes les 100 itérations, ce qui permet de réduire le surcoût de calcul dû au couplage sans perdre en précision de calcul. L'implémentation et la validation de ce couplage massivement parallèle ont constitué un enjeu majeur de ce projet. Nous avons ainsi réussi à coupler jusqu'à 2560 instances AVBP (CPU) avec 64 CNN (GPU).

Ce projet a permis d'effectuer la première co-simulation de ce type. Elle retrouve bien le comportement dynamique qualitatif de la flamme. On retrouve un front de flamme plan laminaire près de la sortie des lèvres de l'injecteur. Plus en aval, du plissement apparaît sous l'effet de la turbulence (voir figure 2).

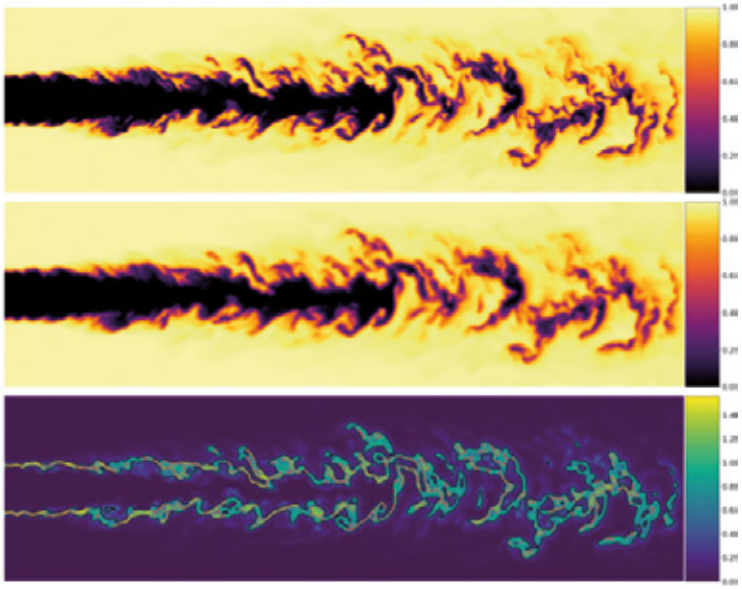


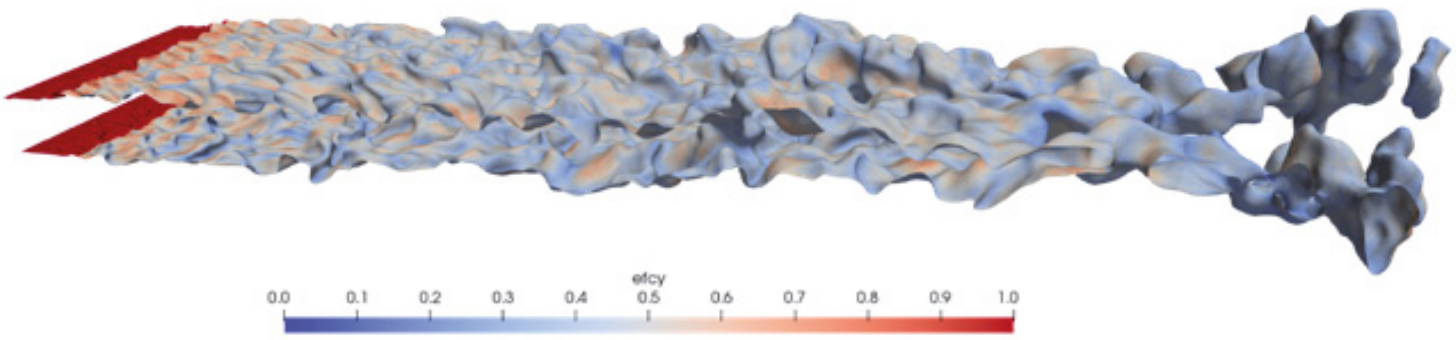
Figure 1 : *Coupes longitudinales d'un instantané de la flamme R3.*
Haut : variable de progrès de la DNS.
Milieu : variable de progrès filtrée et déraffinée sur maillage LES = champ d'entrée du CNN.
Bas : efficacité avec iso-contours à 0.5 en noir et à 1 en rouge = champ de sortie du CNN.

Nous avons cependant observé que le CNN prédisait un champ d'efficacité plus faible que prévu. Ceci nous a permis d'identifier une faiblesse de notre approche : sur ce cas complexe, les champs de la flamme épaissie sont trop différents des champs filtrés de la base d'apprentissage, en particulier au niveau des profils de variable de progrès à travers le front de flamme. En effet, l'opération de filtrage explicite réalisée sur la base d'apprentissage n'est pas strictement équivalente à la combinaison d'un filtrage implicite et d'un épaississement dans la LES d'une flamme épaissie. Le CNN n'arrive donc pas à généraliser sur les données de la simulation. Nous n'avions pas observé ce phénomène sur des configurations plus simples, ce qui souligne l'intérêt de cette simulation de grande ampleur rendue possible grâce à ce projet. C'est donc un résultat très intéressant qui nous mènera à améliorer notre modèle.
Grâce à ce projet "Grand Challenges", nous avons implémenté et validé un couplage hybride massivement parallèle entre le solveur numérique AVBP et un modèle d'efficacité prédit par des CNN. Ceci a permis de réaliser la première co-simulation de ce type avec AVBP. Les résultats obtenus nous ont fait progresser dans notre exploration de l'utilisation de techniques de deep learning en simulation de combustion turbulente.

Nous remercions Dr. Antonio Attili de l'Institute for Technical Combustion, RWTH Aachen University qui nous a fourni les données DNS R2 et R3. Un grand merci également à l'IDRIS et l'équipe support Jean-Zay pour leur aide précieuse durant ce projet.

[1] Lapeyre, Corentin J., et al. «Training convolutional neural networks to estimate turbulent sub-grid scale reaction rates.» Combustion and Flame 203 (2019): 255-264.
[2] Luca, Stefano, et al. «On the statistics of flame stretch in turbulent premixed jet flames in the thin reaction zone regime at varying Reynolds number.» Proceedings of the Combustion Institute 37.2 (2019): 2451-2459.

Figure 2 : *Visualisation de la flamme : iso-surface de variable de progrès c = 0.5 colorée par l'efficacité prédite par les CNN*



Simulation numérique massivement parallèle de l'hydrodynamique et des transferts thermiques d'un réacteur gaz-particule (lit fluidisé) réactif polydisperse à l'échelle industrielle (maillages de 1 et 8 milliards de mailles)

Hervé Neau[a], - Maxime Pigou[a], - Pascal Fede[a], - Renaud Ansart[b], - Cyril Baudry[d], Nicolas Mérigoux[e], - Jérome Laviéville[e], - Yvan Fournier[e], - Nicolas Renon[c], - Olivier Simonin[a].

Nous avons bénéficié du supercalculateur Jean Zay de l'IDRIS dans le cadre des Grands Challenges HPC 2020. Notre Grand Challenge sur la simulation numérique massivement parallèle de l'hydrodynamique et des transferts thermiques d'un réacteur gaz-particule (lit fluidisé) réactif polydisperse à l'échelle industrielle est un projet de mécanique des fluides numérique du comité thématique « Ecoulements réactifs et multiphasiques (CT2b) ».

Ce travail a démarré par un méso-challenge CALMIP (mésocentre de calcul régional Midi-Pyrénées Olympe – Tiers-2) réalisé en 2018 et un grand challenge EDF R&D 2019 (supercalculateur Gaïa). Ce travail préliminaire était le fruit d'une étroite collaboration entre les équipes de deux laboratoires de recherche, d'un mésocentre de calcul régional et d'un industriel français : l'IMFT/LGC (UMRs CNRS/Toulouse INP/UPS), CALMIP et EDF R&D. Ces deux challenges ont permis de réaliser une simulation de 25s de temps physique et de démontrer les excellentes performances parallèles du code NEPTUNE_CFD sur une configuration industrielle avec un maillage non structuré de plus d'un milliard de mailles en utilisant de 30 à 1 000 nœuds de calcul, soit de 1 080 à 36 000 cœurs. Une telle simulation était déjà une première mondiale. Elle a fait l'objet d'une *keynote lecture* au congrès international FLUIDIZATION XVI à Guilin (Chine) en mai 2019 (Neau et al., 2019) et d'un article scientifique (Neau et al., 2020, https://doi.org//10.1016/j.powtec.2020.03.010).

Seul un supercalculateur de dernière génération tel que Jean Zay pouvait permettre de mener à bien une telle étude qui est une première mondiale.

Le grand challenge sur Jean Zay à IDRIS (Tiers-1) a été porté par la même équipe mixte : IMFT/LGC, CALMIP et EDF R&D avec le support de l'IDRIS. L'IMFT développe et intègre des modèles gaz/particules réactifs dans le code de calcul NEPTUNE_CFD dont le cœur est développé par le consortium EDF/CEA/IRSN/Framatome. Ce challenge s'appuyait sur la complémentarité des centres de calcul nationaux et régionaux dans le cadre d'une thématique scientifique forte (développement de modèles et contraintes applicatives industrielles) avec un code co-développé par un industriel et des universitaires.

Le procédé étudié de lit fluidisé est basé sur la mise en suspension d'une poudre par un gaz ascendant. Cela permet d'obtenir des propriétés hydrodynamiques très intéressantes pour des procédés industriels, notamment énergétiques. Ce procédé présente des qualités exceptionnelles de mélange, de mise en contact fluide-particules et d'inertie thermique. Il est ainsi possible de réaliser des réactions chimiques exothermiques à une température quasiment constante et uniforme, en évitant la formation de points chauds et l'emballement du réacteur chimique. Ces lits fluidisés gaz-particule sont utilisés dans de très nombreux procédés industriels :
- les technologies éprouvées avec des réactions à catalyses hétérogènes : colonnes de craquages du pétrole, réacteurs de polymérisation des oléfines (matières plastiques)
- la diminution de la production de gaz à effet de serre en permettant la capture du CO_2 : chambres de combustion ou de gazéification de combustibles solides, chaudières à combustible fossile avec chemical looping, production du ciment
- la transition énergétique : gazéification de la biomasse, systèmes de séchage, récepteurs solaires à lit fluidisé.

D'un point de vue théorique, ces systèmes sont très complexes. Les écoulements mis en jeu sont tridimensionnels, turbulents, instationnaires, multiphasiques, anisothermes et réactifs. Ils sont le siège de transferts de chaleur et de masse entre phases et de réactions chimiques en phase gazeuse ou aux interfaces gaz/particules. Ces écoulements sont fortement multi-échelles induisant des problèmes spécifiques de modélisation physique et de raffinement de maillage.

D'un point de vue applicatif, le besoin d'outils de simulation 3D des écoulements de lits fluidisés réactifs à l'échelle des installations industrielles est très fort (compréhension du fonctionnement, optimisation, nouveaux concepts) d'autant plus que les parois métalliques de ces réacteurs et les conditions opératoires interdisent quasiment toute mesure expérimentale. Cependant leur mise en œuvre pratique pose toujours de nombreux problèmes tant du point de vue de la validité des modèles physiques que de la précision des schémas numériques ou des performances des codes. L'obligation de prendre en compte la géométrie complète à l'échelle industrielle, de plusieurs dizaines de mètres de haut, incluant des obstacles internes ou des injecteurs de quelques millimètres, couplée à la nécessité de résoudre l'écoulement à très petite échelle, impose des simulations instationnaires très lourdes qui ne peuvent être menées à bien que sur des supercalculateurs de dernière génération, avec des codes massivement parallèles. La puissance du supercalculateur Jean Zay était indispensable pour réaliser une telle simulation avec plus de 8 milliards de mailles.

Le code massivement parallèle NEPTUNE_CFD résout les équations de type Navier-Stokes moyennées par une méthode Volumes Finis sur maillage non structuré. Il résout en 3D et pour chacune des phases considérées des équations instationnaires Eulériennes de masse, quantité de mouvement et enthalpie, fortement couplées par l'intermédiaire des termes de transferts entre les phases.

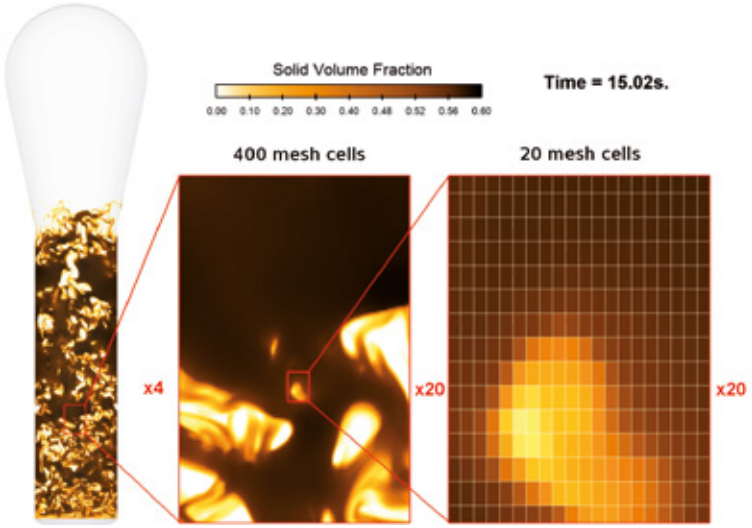


Figure 1 : *Champ instantané de taux de présence des particules dans un lit fluidisé 3D réactif à l'échelle industrielle (zooms successifs faisant apparaitre le maillage)*

La simulation effectuée correspond à l'hydrodynamique et aux transferts thermiques avec une réaction exothermique dégageant 20 MW d'un réacteur industriel à lit fluidisé d'une hauteur totale de 30m pour un diamètre de 5 m contenant 100 tonnes de particules. Le maillage utilisé compte plus de 8 milliards de cellules hexaédriques de 2,5 millimètres de côté. Ce maillage est 8 fois plus raffiné que le plus fin jamais utilisé.

Ce grand challenge a permis :

- de démontrer la faisabilité de la simulation massivement parallèle d'un lit fluidisé à l'échelle industrielle avec un maillage 8 fois plus fin que le plus gros jamais utilisé soit 8 milliards de mailles. Cela a permis aussi de démontrer que NEPTUNE_CFD permet d'utiliser efficacement la totalité du supercalculateur Jean Zay soit 60 000 cœurs CPU
- de simuler 1.5 secondes de temps physique avec une précision inédite sur ce maillage extrêmement raffiné au-delà des 25 s obtenues avec le maillage à 1 milliard de cellules. Les petites structures, clusters et bulles, sont capturées extrêmement finement et elles jouent un rôle crucial sur le comportement macroscopique de l'écoulement.

L'utilisation d'un maillage aussi raffiné a permis la prise en compte réaliste des injections latérales de catalyseur dans le lit fluidisé dont le rôle au niveau thermique est fondamental et de tester l'indépendance des résultats au maillage
- d'évaluer les performances du code et sa capacité de passage aux grandes échelles, de 240 à 1 528 nœuds sur le maillage à 8 milliards de cellules
- de constituer une base de données de référence de champs 3D instationnaires complets et de s'en servir pour développer des modèles de sous-maille par filtrage spatial permettant de reproduire correctement les résultats physiques macroscopiques malgré des maillages moins raffinés

La démonstration des excellentes performances HPC du code et de la machine en pleine charge viennent compléter les challenges précédents et permettent de comparer les technologies. Seul un supercalculateur de dernière génération tel que Jean Zay pouvait permettre de mener à bien une telle étude qui est une première mondiale.

Cette étude se poursuit actuellement sur le supercalculateur Irene AMD du TGCC (Tiers-0) sur le même cas avec un maillage de 64 milliards de mailles qui requiert la totalité de la machine. Ce projet montre bien la complémentarité des différentes échelles des centres de calcul français et européens.

[*a*] Institut de Mécanique des Fluides de Toulouse (IMFT), Université de Toulouse, CNRS, INPT, UPS Toulouse, France

[*b*] Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France

[*c*] UMS CALMIP 3667 Université de Toulouse, CNRS, INPT, INSA, ISAE, UPS, Toulouse, France

[*d*] Délégation technologies et systèmes d'information, EDF R&D, Palaiseau, France

[*e*] *Fluid Mechanics, Energy and Environment Dpt.*, EDF R&D, Chatou, France

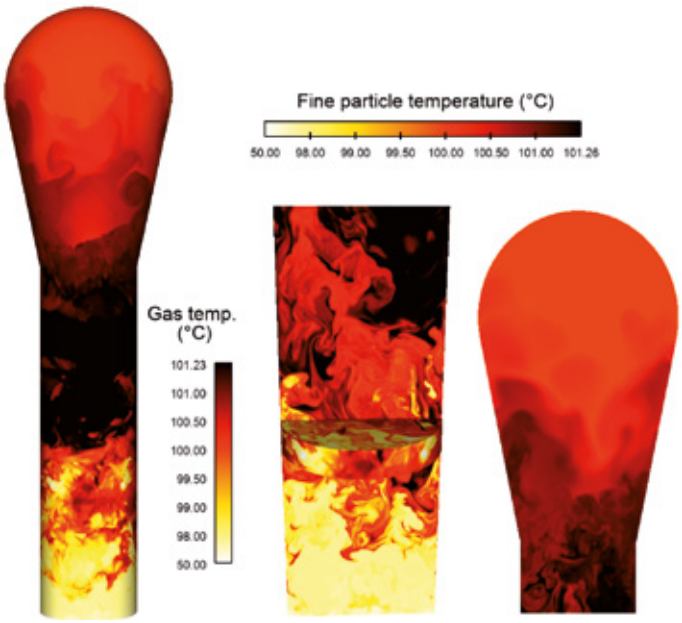
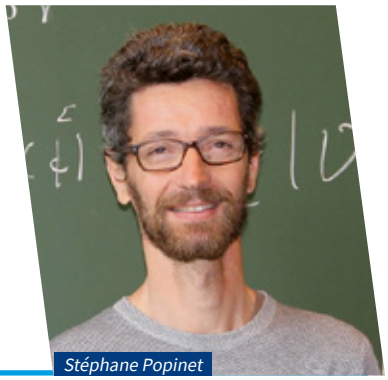


Figure 2 : *Champs instantanés de températures du gaz et des particules fines (vue de peau, plans de coupe)*

Étude d'écoulements multiphasiques complexes

Stéphane Popinet (DR CNRS)
& Stéphane Zaleski (Pr Sorbonne Université et IUF)
Institut Jean Le Rond d'Alembert, Sorbonne Université, CNRS et IUF, Paris.


Stéphane Popinet

L'étude d'écoulements multiphasiques complexes possède un intérêt considérable dans les sciences de l'ingénieur et les sciences de la nature. Un exemple est le déferlement des vagues de mer, qui contribue fortement aux échanges thermiques et de masse entre l'océan et l'atmosphère et par conséquent constitue un élément important des modèles climatiques.

Jean Zay performe également mieux que les autres supercalculateurs lorsque le nombre de mailles par cœur est augmenté

Un autre exemple est le transfert de masse dans les procédés industriels qui fait appel à la détermination du calcul de l'arrachage des gouttes à partir des nappes liquides cisaillées.

Le calcul fait appel à la résolution des équations de la mécanique des fluides, notamment les équations de Navier-Stokes avec termes capillaires et fortes variations de densité. Ce calcul se heurte à de nombreuses difficultés dont l'une des plus manifestes est la vaste gamme d'échelles de longueur en cause, depuis les plus petites bulles ou gouttes de l'ordre du micron jusqu'aux tailles de vagues ou des procédés industriels, de l'ordre du mètre. Un calcul complet n'est pas possible et plusieurs stratégies sont utilisées. Celle de ce projet est d'utiliser une forme de *Simulation Numérique Directe* ce qui permet d'éviter les incertitudes sur les équations. Cependant le projet est alors encore plus coûteux en temps de calcul que si des simulations basées sur des modèles simplifiés étaient utilisées. Ce temps de calcul important est cependant réduit d'une part par la réduction de la taille du problème physique (des vagues plus courtes ou des modèles réduits des activités industrielles sont utilisés), et d'autre part par l'utilisation de méthodes de maillage hiérarchique adaptatif en « arbre *octobranche* » *(octtree)* dans laquelle le maillage s'adapte et se raffine automatiquement autour des régions contenant les petites structures de la solution comme les tourbillons et les gouttes. Ce type de calcul est cependant très difficile à paralléliser et à optimiser, et les résultats peuvent dépendre du type de calculateur utilisé, et de l'environnement logiciel.

Les codes ont fonctionné avec des performances excellentes sur Jean Zay (comme le montrent les tests de *weak scaling* de la figure 1) . Cinq millions d'heures scalaires ont été utilisées sur un nombre de cœurs moyen de deux mille. Les calculs ont permis de contribuer à une étude poussée du phénomène de déferlement de vagues. La plupart des experts considèrent que le déferlement est principalement un phénomène fluide bidimensionnel ou plan. Le travail de Moustert, Popinet et Deike (JFM, soumis) montre que ce n'est pas le cas et que au contraire des structures transversales à la direction de propagation de la vague se forment (figures 2 et 3). Un régime asymptotique indépendant du nombre de Reynolds $Re = \rho U L / \mu$ est obtenu quand $Re > 10^5$.

Une perspective fascinante de ce travail et d'obtenir une prédiction numérique du transfert de masse et de chaleur dans ce type d'écoulement. Pour l'instant les calculs sont effectués avec uniquement les termes de quantité de mouvement, sans l'équation de la chaleur ou de la diffusion nécessaires pour l'estimation de ces transferts. La difficulté est d'avoir des mailles de calcul suffisamment fines pour résoudre les zones très minces, appelées couches limites, où se produit la diffusion. Les résultats déjà obtenus donnent bon espoir que ce nouveau challenge pourra être surmonté.

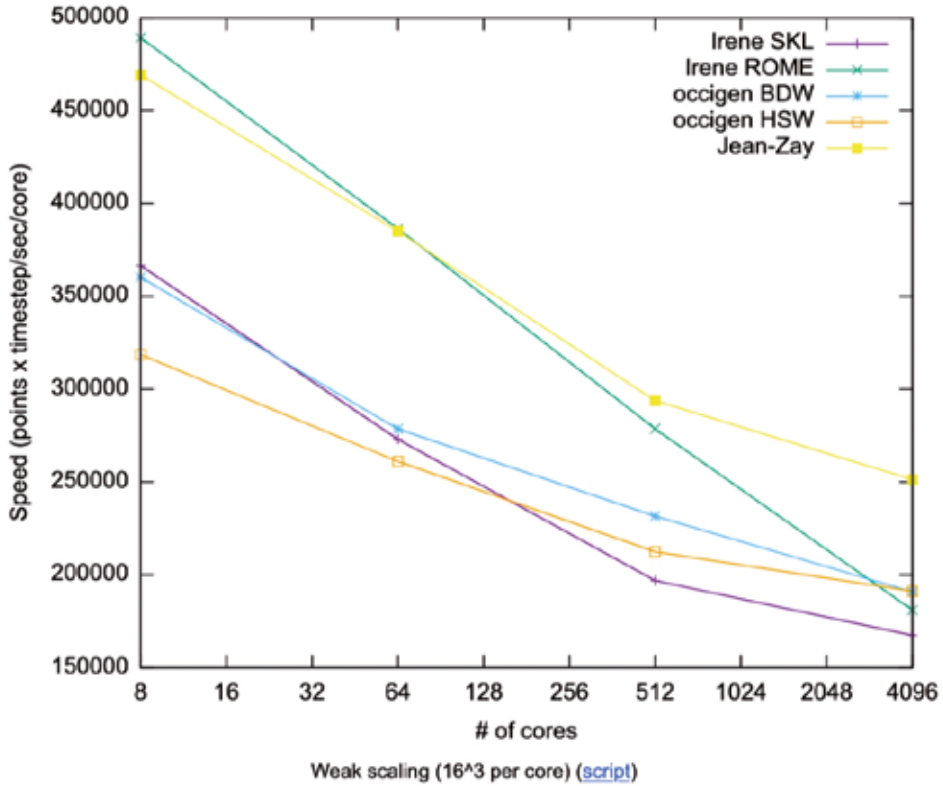


Figure 1 : *Test en échelonnement faible de la parallélisation du code sur divers supercalculateurs. Le cas de seize mailles au cube est montré, mais Jean Zay performe également mieux que les autres supercalculateurs lorsque le nombre de mailles par cœur est augmenté comme montré dans http://basilisk.dalembert.upmc.fr/sandbox/joubert/benchmark/weak_scaling.c*

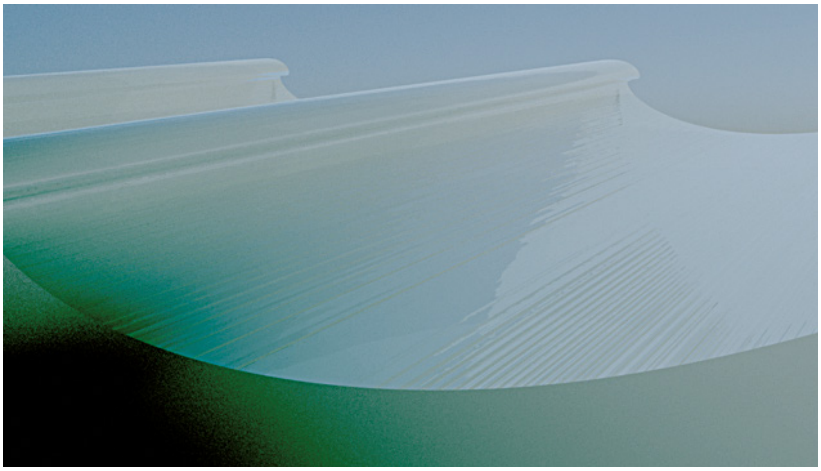


Figure 2 : *Rendu par tracé de rayon d'une vague déferlante eau air dans le cas Reynolds = 10^5, avec une résolution de 2 048 mailles dans chaque direction. Cambrement non-linéaire à t/T=0.37.*

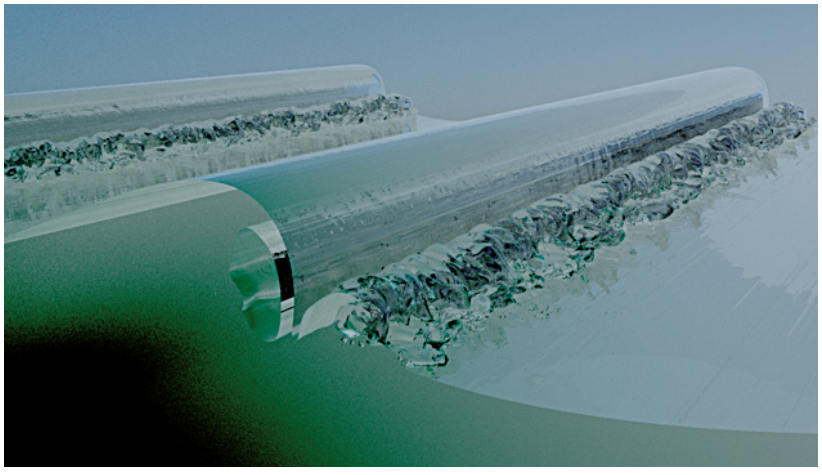


Figure 3 : *Rebond de la vague principale (splash-up).*

Simulation d'une image nucléaire SPECT 4D d'un traitement du cancer

David Sarrut [1,3], Thomas Baudier [1,3], Antoine Robert [1,2], Sébastien Jan [4]

[1] *Univ.Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69373, Lyon, France.*

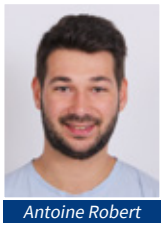
[2] *Kitware SAS, 6 Cours André Philip, 69100 Villeurbanne.*

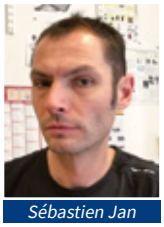
[3] *Centre Léon Bérard, 28, rue Laennec, 69373 Lyon Cedex 08, France.*

[4] *Université Paris-Saclay, CEA, CNRS, Inserm, Biomaps, Service Hospitalier Frédéric Joliot, 4 place du général Leclerc, 91401 ORSAY France*


David Sarrut


Thomas Baudier


Antoine Robert


Sébastien Jan

Le but de ce projet Grand Challenge était de réaliser une grande simulation sur des clusters de calcul haute performance partagés. La tâche est une simulation Monte Carlo réaliste d'une image SPECT complète pendant le traitement du cancer au Lutécium 177 (^{177}Lu), en tenant compte du mouvement respiratoire du patient. Cette simulation, supposant environ 17 ans de temps CPU, a été exécutée, ici, en environ 30 jours. Il illustre la faisabilité d'une simulation Monte Carlo réaliste à grande échelle d'un protocole de cancer théranostique. Tous les développements et résultats sont ouverts et disponibles pour la communauté au sein de la plateforme GATE.

1. Collaboration et centres de calcul

Cette simulation, supposant environ 17 ans de temps CPU, a été exécutée ici en environ 30 jours.

Ce travail a été réalisé dans le cadre d'une collaboration entre des chercheurs du laboratoire CREATIS (CNRS, Lyon), du centre de lutte contre le cancer Léon Bérard (Lyon) et du Service Hospitalier Frédéric Joliot (CEA, Orsay). Deux centres de calcul ont été utilisés pour effectuer le calcul. Le premier est le supercalculateur Jean Zay géré par l'IDRIS (CNRS-INS2I), à Orsay. Il a été installé en 2019 et contient 1528 nœuds de calcul. Le second est le centre de calcul de l'IN2P3 (CNRS-IN2P3) à Lyon. Ce centre contient environ 32700 CPU. Les deux centres ont été utilisés dans un mode « non-dédié », c'est-à-dire que les demandes de calculs sont placées en concurrence dans une même liste d'attente. Le temps de calcul global dépend donc de la charge du cluster.

2. Contexte médical

L'objectif était de réaliser une simulation Monte Carlo d'une image tomographique SPECT *(Single Photon Emission Computed Tomography)* acquise pendant un traitement du cancer par radiothérapie interne utilisant du ^{177}Lu, et tenant compte des mouvements respiratoires du patient. Le schéma de désintégration radioactive du ^{177}Lu permet d'irradier les tumeurs neuroendocrines par émission bêta tandis que les émissions gamma permettent en même temps de contrôler la distribution de dose par imagerie SPECT.

Cependant, les images souffrent de plusieurs incertitudes, comme celle due à la respiration du patient qui diminue la qualité de l'image et nuit à la quantification [1].

Dans ce contexte, les simulations Monte Carlo sont un élément clé pour comprendre et corriger ses phénomènes. Malgré l'existence de certaines techniques de réduction de variance, les simulations Monte Carlo restent très couteuses en temps de calcul et ne sont, en pratique, pas utilisées pour ce type d'applications cliniques.

Nous pensons ainsi qu'une simulation de référence sera utile à la communauté.

3.Simulations

Tous les principaux éléments impliqués dans une acquisition SPECT d'un patient traité par du ^{177}Lu ont été simulés (Figure1) incluant les éléments suivants :

- Une image CT d'un patient traité avec du ^{177}Lu au centre Léon Bérard (incluant le mouvement respiratoire) ;
- Le système d'imagerie SPECT General Electric NM670 ;
- La distribution d'activité a été obtenue à partir d'une acquisition SPECT, obtenue 24 heures après injection. Les valeurs des voxels de l'image ont été converties en concentration d'activité radioactive (MBq/mL).

Les mouvements respiratoires de cette source ont été simulés grâce aux champs de vecteurs obtenus par recalage déformable des images morphologiques CT 4D. L'activité du ^{177}Lu a été simulée par une émission gamma isotrope correspondant au spectre énergétique de l'isotope.

Une fois la simulation terminée, un signal respiratoire a été extrait des données en utilisant la méthode décrite dans [2]. Les évènements ont été regroupés temporellement en trames de 200 ms et spatialement en projections de 256 × 256 pixels. Puis, le signal obtenu a été divisé en 10 phases respiratoires calculées en détectant les positions d'inspiration finale. Les projections ont été triées en fonction de ces phases en dix sous-ensembles. Chaque sous-ensemble a été reconstruit à l'aide de RTK (Reconstruction Toolkit) [3]. Les reconstructions ont été faites avec 15 itérations et 4 sous-ensembles en utilisant l'algorithme OSEM combiné à un débruitage basé sur la variation totale (TV). La correction du diffusé a été prise en compte à l'aide de la méthode Double Energy Window [4]. Les corrections d'atténuation (AC) et de la PSF ont été faites au cours du processus itératif avec la méthode décrite dans [5]. Les deux lits ont été reconstruits indépendamment l'un de l'autre. Ensuite, les deux résultats ont été fusionnés pour obtenir une reconstruction SPECT 4D corrélée à la respiration, de 10 images.

[1]Institut du développement et des ressources en informatique scientifique

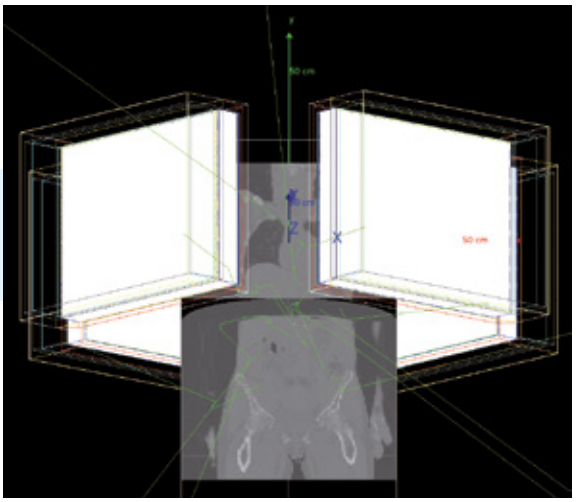


Figure 1 : *Simulation d'un imageur SPECT composé ici de 4 têtes de détection.*

4. Résultats et discussion

4.1. Temps de calcul

Avec les paramètres utilisés, le temps de calcul était d'environ 2100 PPS (particules par seconde) sur les ordinateurs du CC-IN2P3 (XA730i Intel Cascade Lake 6248), et de 3500 PPS sur le cluster Jean Zay (Intel Xeon CPU E5-2650). Compte tenu du nombre total de particules primaires simulées, autour de 9.9E11, cela correspond à un total d'environ 17.4 ans de temps processeur. Au total, 8000 jobs ont été lancés. Si les jobs avaient pu être tous exécutés en parallèle, le temps de calcul aurait été de 19 heures. Cependant, les deux centres de calcul utilisés n'étaient pas dédiés exclusivement à cette tâche et les jobs étaient en concurrence avec ceux des autres utilisateurs (mode dit « non-dédié »). De plus, le nombre de job simultanés en file d'attente était limité. Nous avons également dû faire face à un arrêt dû à une maintenance inattendue.
Globalement, la durée totale a été d'environ 30 jours, conduisant à une accélération moyenne d'environ 200. Bien entendu, l'utilisation d'un cluster dédié aurait permis de réaliser un calcul très rapide (moins d'un jour), mais cette expérience en « conditions réelles » montre le gain réel qui peut être attendu grâce à ce type de centre.

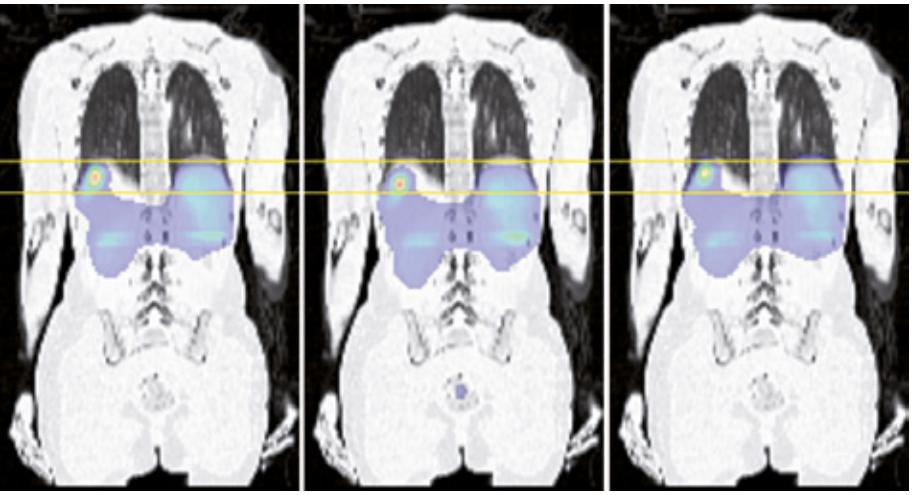


Figure 2 : *Reconstruction des images issues des simulations. A gauche : image reconstruite sans correction de mouvement. Au centre et à droite : image corrélée à la respiration, fin d'inspiration et fin d'expiration.*

4.2 Simulation

Au total, les 10 phases respiratoires et les 2 pas de lits ont représenté environ 81 Go de données, soit un total d'environ 255 millions d'événements détectés.

Notre algorithme d'analyse nous a permis d'extraire un signal respiratoire de ces données et de reconstruire une image SPECT par phase respiratoire (Figure2). Nous avons vérifié que le signal extrait automatiquement était proche de celui de référence : une période moyenne de 3.97 secondes a effectivement été trouvé pour les 4 secondes de référence.

5. Conclusion

Une simulation numérique d'une image nucléaire d'un traitement du cancer, durant environ 17 ans de calcul a été réalisée en 30 jours sur un centre de calcul CNRS, en environnement non dédié, conduisant à une accélération moyenne d'un facteur 200. Toutes les données de simulation sont disponibles sur demande.

[1] Remco Bastiaannet, Max A Viergever, and Hugo W A M de Jong. Impact of respiratory motion and acquisition settings on SPECT liver dosimetry for radioembolization. Medical physics, 44(10) :5270–5279, October 2017.

[2] James C. Sanders, Philipp Ritt, Torsten Kuwert, A. Hans Vija, and Andreas K. Maier. Fully Automated Data-Driven Respiratory Signal Extraction From SPECT Images Using Laplacian Eigenmaps. IEEE transactions on medical imaging, 35(11) :2425– 2435, November 2016.

[3] S Rit, M Vila Oliva, S Brousmiche, R Labarbe, D Sarrut, and G C Sharp. The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK). Journal of Physics : Conference Series, 489 :012079, March 2014.

[4] Ronald J. Jaszczak, Kim L. Greer, Carey E. Floyd, C. Craig Harris, and R. Edward Coleman. Improved SPECT Quantification Using Compensation for Scattered Photons. Journal of Nuclear Medicine, 25(8) :893–900, January 1984.

[5] Gengsheng Zeng, Chuanyong Bai, and Grant T. Gullberg. A projector/backprojector with slice-to-slice blurring for efficient three-dimensional scatter modeling.
IEEE transactions on medical imaging, 18 :722–32, September 1999.

SALT: *Shining A Light Through the dark ages*

Pierre Ocvirk[1],
Joe Lewis[1],
Yohan Dubois[2],
Dominique Aubert[1]
[1]*Observatoire Astronomique de Strasbourg*
[2]*Institut d'Astrophysique de Paris*


Dominique Aubert


Pierre Ocvirk


Joe Lewis


Yohan Dubois

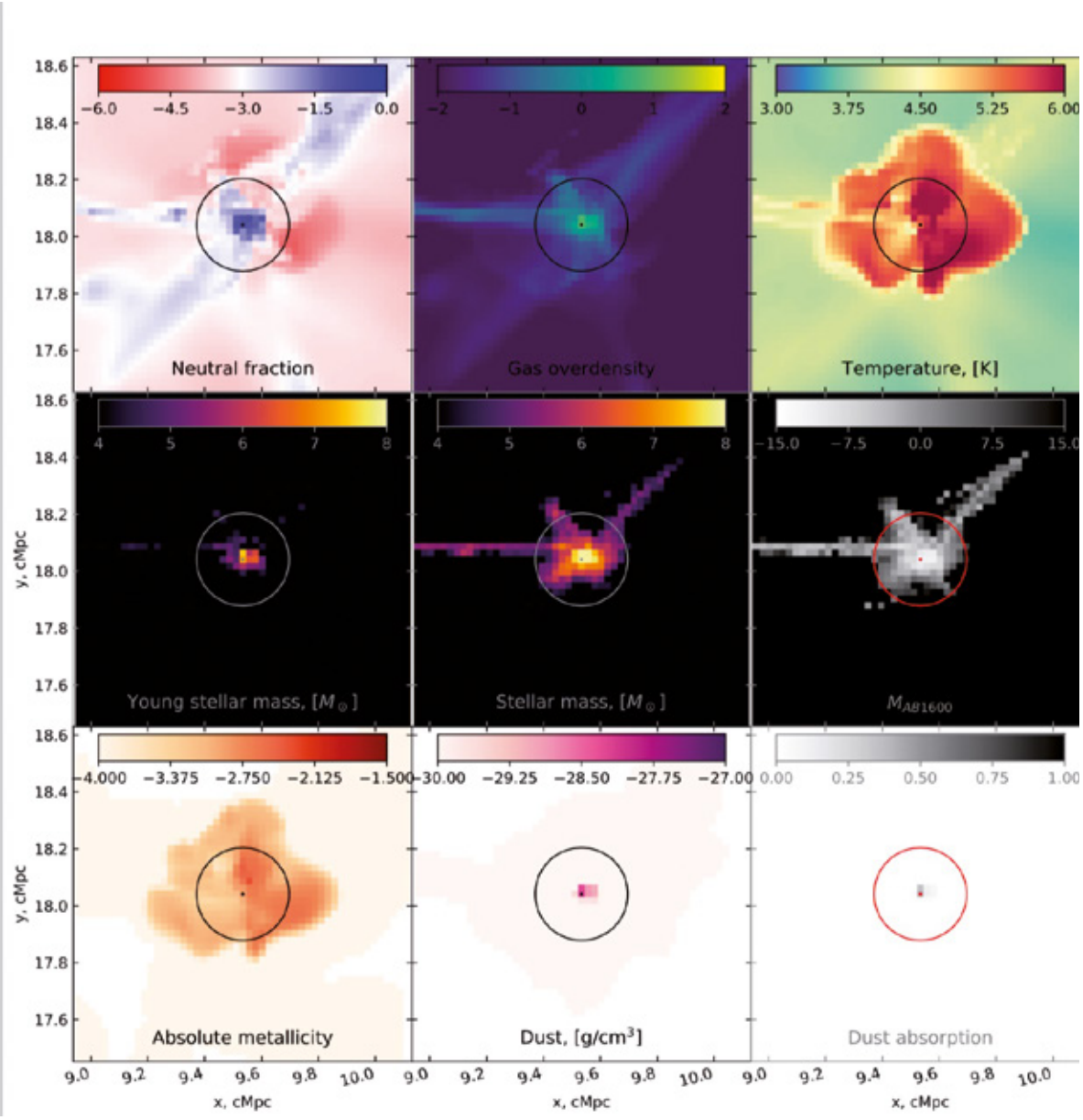
Le projet grand challenge SALT: *Shining A Light through the dark ages* vise à modéliser la formation des galaxies pendant le premier milliard d'années de l'Univers, période qui comprend la naissance des toutes premières étoiles, marquant ainsi la fin de la période des "âges sombres". Le contexte du projet est celui des nouvelles données apportées par les observatoires dédiés à l'étude de cette époque reculée, tels que NenuFAR (Station de Radioastronomie de Nançay), le James Webb Space Telescope, et les futurs télescopes géants SKA et ELT.

Nous avons pu constater le très bon comportement de Jean Zay sous la charge de ces simulations, pour lesquelles 128 nœuds, 512 GPUs et 4096 cœurs sont utilisés en parallèle.

Peu après le Big Bang, la matière composant l'Univers est sous la forme d'un plasma très chaud. C'est seulement 300 000 ans plus tard, lorsque l'Univers s'est suffisamment refroidi, que les premiers atomes d'Hydrogène peuvent se former. Le plasma originel devient alors un gaz neutre et relativement froid. C'est à partir de ce gaz que les toutes premières étoiles se forment, environ 150 millions d'années après le Big Bang. En se propageant, leur intense lumière casse les atomes d'Hydrogène, ramenant ainsi le milieu intergalactique à l'état de plasma qui prévalait juste après le Big Bang. C'est ce bégaiement dans l'histoire d'ionisation de l'Univers, qui amène les astrophysiciens à parler de "ré-ionisation", puisque l'Univers redevient ionisé avec l'apparition des premières étoiles. Cette réionisation s'accompagne d'un échauffement parfois lourd de conséquences : le gaz devient suffisamment chaud pour échapper à la faible gravité des galaxies les moins massives, les privant ainsi du matériau qui leur permettait de former des étoiles. Ainsi, les étoiles jeunes produisent un rayonnement ionisant qui tend, dans certaines conditions, à freiner ou empêcher la formation des générations stellaires suivantes. Ce type de rétroaction rend la modélisation de la formation des galaxies particulièrement délicate, des lors que l'on prend en compte le rayonnement, essentiel pour bien décrire l'époque de la réionisation. On parle de couplage d'échelles : le cœur des petites galaxies de l'Univers jeune mesure quelques centaines d'années lumière tout au plus, et pourtant leur capacité à collecter (par gravité) et garder du gaz capable de former des étoiles peut être affectée par le rayonnement collectif des galaxies voisines et jusqu'à des distances de plusieurs dizaines de millions d'années lumière. Nos simulations doivent ainsi pouvoir décrire le très petit comme le très grand. Notre projet Grand challenge devait initialement réaliser des simulations idéalisées de régions à très grande résolution spatiale, afin d'étudier l'impact de l'irradiation UV par les premières étoiles et galaxies sur les structures gazeuses du milieu inter-galactique et proto-galactique. Ces simulations ont été réalisées et nous avons pu constater le très bon comportement de Jean Zay sous la charge de ces simulations, pour lesquelles 128 nœuds, 512 GPUs et 4096 cœurs sont utilisés en parallèle. Nous avons aussi choisi d'utiliser une partie du temps grand challenge pour développer, tester et calibrer de nouveaux modules de physique pour notre code RAMSES-CUDATON.

Le doctorant strasbourgeois Joe Lewis (soutenance prévue en septembre 2020) a pour projet de thèse d'implémenter les processus d'enrichissement chimique et surtout de formation des poussières et leur impact sur le *transfert radiatif* ionisant dans notre code de simulation RAMSES-CUDATON. Ce travail été mené en collaboration avec Y. Dubois (IAP), et a tiré parti du temps grand challenge alloué, qui nous a permis de progresser dans la compréhension et la prise en compte de ces processus pendant l'époque de la réionisation. C'est une physique complexe, impliquant de nombreuses boucles de rétro-action, positive et négative. Par exemple, les métaux (les astronomes appellent métal tout élément autre que hydrogène et hélium), dans le milieu interstellaire, forment des grains de poussière. Ces grains absorbent le rayonnement ionisant émis par les étoiles jeunes, et diminuent donc l'intensité du rayonnement disponible pour ré-ioniser l'Univers à grande échelle. De plus, ces grains de poussière augmentent la capacité du gaz d'hydrogène à refroidir, et donc à recombiner et redevenir neutre, augmentant ainsi encore l'opacité du milieu interstellaire des galaxies. Cependant, ce refroidissement va aussi conduire à augmenter le taux de formation d'étoiles de ces galaxies, entrainant une augmentation de la production intrinsèque de photons ionisants. Le but des simulations entreprises sur le temps grand challenge est de déterminer l'impact final de ce réseau de processus physiques sur la capacité des galaxies a ré-ioniser l'Univers. Leur analyse est encore en cours.
La figure ci-contre montre l'environnement complexe d'une galaxie pesant 10^{11} fois la masse du soleil, tirée d'une de nos simulations grand challenge, lorsque l'Univers avait 1 milliard d'années. Chaque panneau correspond a un champ physique différent. Toutes les échelles de couleur sont logarithmiques sauf l'absorption des poussières, qui est linéaire. En haut, de gauche à droite : fraction d'hydrogène neutre, sur-densité du gaz, température.

Au milieu, la masse stellaire jeune (moins de 10 millions d'années), au centre, la masse stellaire totale, la magnitude UV (1 600 Angström) par cellule. En bas, la métallicité, la densité de poussières, et l'absorption due aux poussières. Il apparait que l'absorption à 1 600 Angström par les poussières est circonscrite aux régions de formation d'étoiles de la galaxie, et peut atteindre 60%, c'est a dire une magnitude. Dans le continu ionisant, l'opacité est 3 à 4 fois plus importante, et la transmission du rayonnement ionisant peut donc être réduite d'un facteur 10 par rapport à une galaxie sans poussières. Ces résultats sont exposés dans un article en préparation ainsi que dans le manuscrit de thèse de Joe Lewis. La calibration du modèle obtenue lors du Grand challenge sera raffinée grâce a une autre allocation DARI (en cours), puis déployée sur Summit (Oak Ridge National Laboratory) pour réaliser Cosmic Dawn III, une simulation géante de formation des galaxies pendant l'époque de la réionisation, pour un coût total de 3 600 000 heures GPU. Cette nouvelle simulation, qui utilisera simultanément jusqu'à 24 576 GPUs pour produire environ 20 Péta-octets de données en une semaine de calculs, permettra de mieux capturer les propriétés et l'évolution de la population d'absorbants du milieu intergalactique et du mélange complexe des différentes phases de gaz chaud et froid entourant les galaxies.



ExoConv : Simulation à haute résolution de la convection dans les atmosphères d'exoplanètes de type terrestre

S. Daley-Yates, T. Padioleau, P. Tremblin, P. Kestener, M. Mancip
Université Paris-Saclay, UVSQ, CNRS, CEA, Maison de la Simulation, 91191, Gif-sur-Yvette, France


Simon Daley-Yates

L'objectif principal du grand challenge ExoConv est d'étudier à haute résolution le comportement convectif des atmosphères d'exoplanètes de type terrestre fortement irradiées ou jeunes (donc à haute température) et sujette à une transition chimique comme la transition entre CO et CO_2. Cette transition a aussi potentiellement eu lieu dans l'atmosphère de la Terre au moment de sa formation dans le système solaire. Une meilleure compréhension de ce mécanisme pourrait donc apporter un nouvel éclairage sur la paléoclimatologie de la Terre primitive.

La simulation, réalisée sur 1 000 GPUs, a permis de confirmer par une étude de convergence la réduction du gradient de température dans l'atmosphère.

Nous avons réalisé une simulation hydrodynamique haute résolution en milieu stratifié, en présence d'un gradient de poids moléculaire moyen dans l'atmosphère entre le haut de la boîte (i.e. dominé par CO_2) et le bas de la boîte (dominé par CO). L'atmosphère est en présence d'un gradient instable au critère de Ledoux pour la convection. Le gaz CO_2-"lourd" va vouloir descendre alors que le gaz CO-"léger" aura tendance à remonter. La chimie va alors s'activer et convertir le CO monté en haut de la boite en CO_2 et inversement, le CO_2 tombé au fond de la boite en CO. Le gradient instable est alors restauré et on maintient un cycle convectif turbulent associé au gradient de poids moléculaire et à la chimie (voir figure 1). La conversion CO/CO_2 va alors induire un chauffage/refroidissement dans l'atmosphère induit par le transfert radiatif et surtout par le changement de poids moléculaire. De manière contre-intuitive ce chauffage/refroidissement amène à une réduction du gradient de température à une valeur « sous-adiabatique » alors qu'une convection adiabatique de type Rayleigh-Bénard est censée maintenir le gradient de température à un gradient adiabatique (voir figure 2).

Cette simulation a été réalisée avec le code ARK[1] (Padioleau et al. 2019) qui s'appuie sur la bibliothèque Kokkos[2] pour obtenir la portabilité de performance entre les différentes architectures de calcul émergentes à l'ère de l'exascale. Nous avons ainsi pu réaliser avec le même code (et une seule implémentation des noyaux de calcul) une étude paramétrique dans le cadre de l'appel DARI A6 de GENCI sur la partition KNL du supercalculateur Joliot-Curie @CEA/TGCC et ce grand challenge sur la partition GPU du supercalculateur Jean Zay @CNRS/IDRIS. La simulation grand challenge a une résolution de $5\,000^3$ et a été réalisée sur 1 000 GPUs pour une consommation totale de 125 000 heures GPU. Elle a permis de confirmer par une étude de convergence la réduction du gradient de température dans l'atmosphère associée à la convection radiative CO/CO_2 et aussi d'étudier la dynamique des petites échelles et la cascade turbulente dans ce type de convection.

Ce mécanisme est nouveau et a été récemment proposé par Tremblin et al. (2019) dans le cadre d'une théorie générale pour la convection thermo-compositionnelle diabatique (i.e. avec termes sources, transfert radiatif et chimie pour ce cas d'étude). Cette théorie permet de décrire aussi bien la convection thermohaline dans les océans terrestres, la convection humide dans l'atmosphère terrestre et potentiellement la convection radiative CO/CH4 dans les atmosphères d'exoplanètes géantes et de naines brunes.

La convection radiative CO/CO_2 est une extension directe de ce formalisme pour les exoplanètes de type terrestre (avec une atmosphère pauvre en hydrogène) et serait une toute nouvelle prédiction pour de futures observations en astrophysique, avec les futurs télescopes destinés à la caractérisation des atmosphères d'exoplanètes (par exemple le télescope spatial James Webb).

[1] https://gitlab.erc-atmo.eu/erc-atmo/ark
[2] https://github.com/kokkos

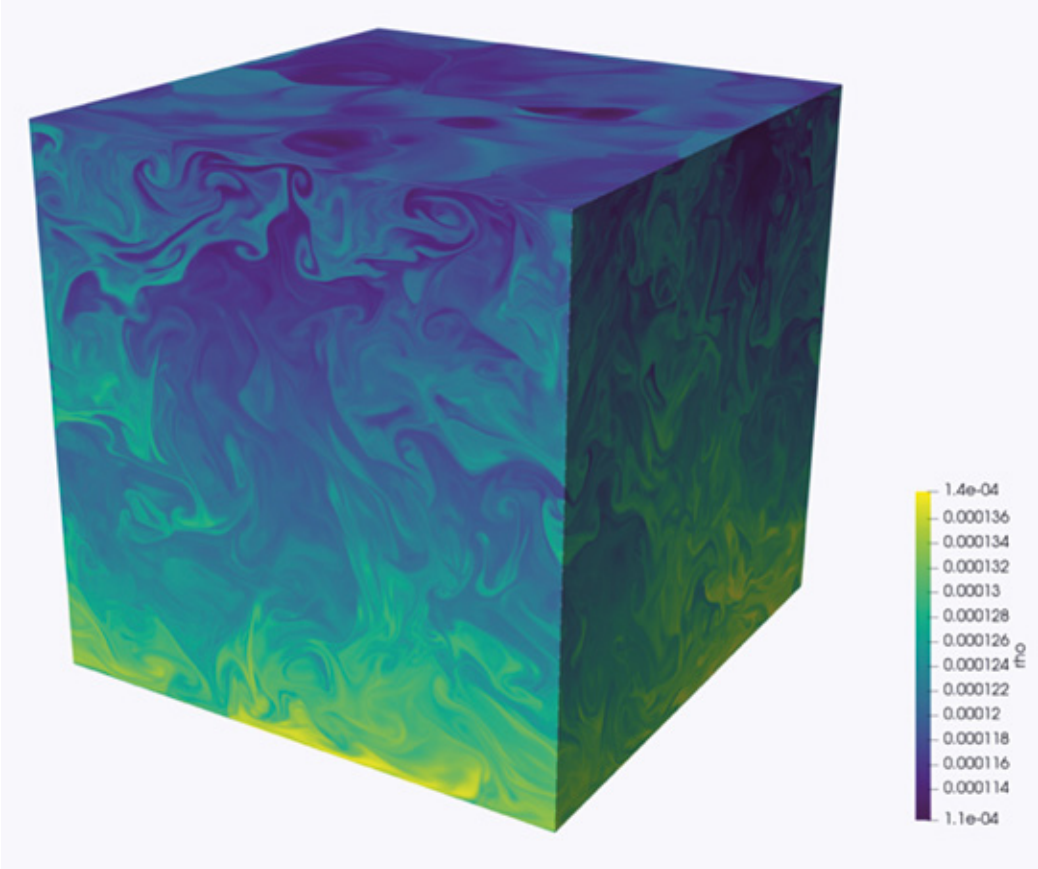


Figure 1 : *Visualisation de la densité de la simulation grand challenge à une résolution de 5000^3, pour un temps auquel le régime turbulent stationnaire de la convection radiative CO/CO_2 est atteint.*

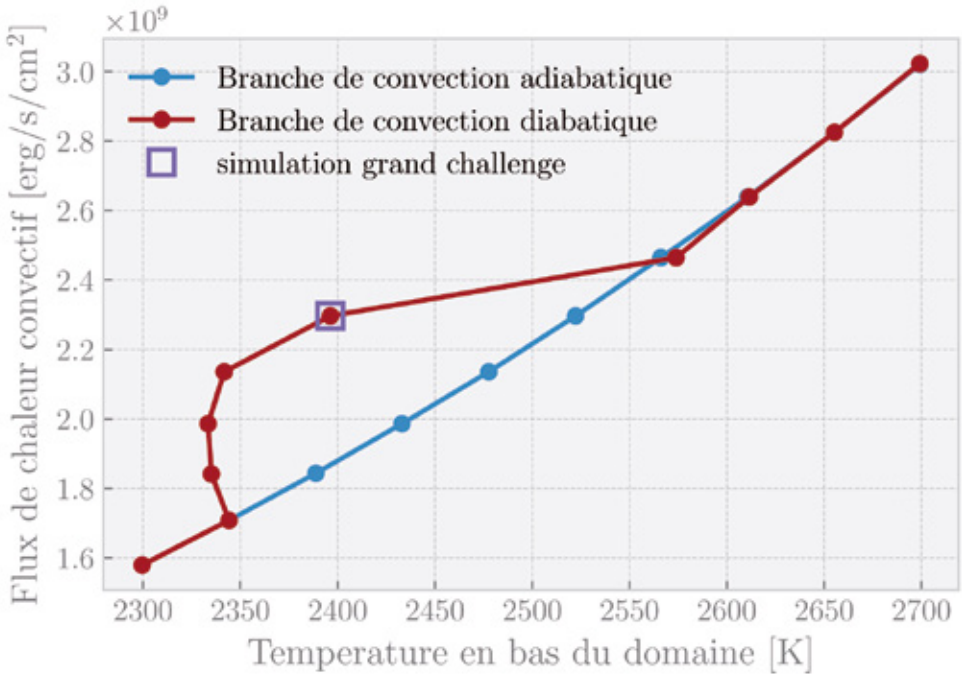


Figure 2 : *Étude paramétrique du flux convectif en fonction de la température au fond de l'atmosphère pour la convection radiative CO/CO_2, pour un temps auquel la réduction du gradient de température est la plus importante. La simulation grand challenge effectuée à haute résolution est indiquée par le carré magenta.*

Émulation de simulations de Réionisation par apprentissage profond


Emilie Thélie


Dominique Aubert


Jonathan Chardin


Pierre Ocvirk

Collaborateurs :
Dominique Aubert,
Jonathan Chardin,
Pierre Ocvirk & Emilie Thélie
Observatoire Astronomique de Strasbourg (ObAS) UMR 7550 Universite de Strasbourg CNRS

Contexte

La *Réionisation* désigne une époque reculée, durant laquelle l'Univers voit apparaître les premières étoiles et les premières galaxies. Ces tout premiers objets vont produire et remplir le cosmos de rayonnement ultraviolet qui va réchauffer et ioniser tout le gaz d'hydrogène. Ce faisant, un réseau de régions ionisées va s'établir autour de tous les sites de formation d'étoiles, épousant les contours des grandes structures de l'Univers. Durant le premier milliard d'années après le Big Bang, ces régions vont croître et multiplier pour aboutir à un cosmos complètement ionisé, tel qu'il est toujours aujourd'hui.

Jean Zay aura permis une convergence entre la production de données d'apprentissage et l'apprentissage lui-même sur un seul et même site de façon très efficace

Ce processus est fondamental car il permet en principe de tracer les stades initiaux de la formation des galaxies et des étoiles, toujours aujourd'hui mal connus. En particulier, il relie les grandes échelles cosmologiques, qui se font ioniser, à celles plus « petites » et qui correspondent aux sites de production de ce rayonnement de première génération. Toutefois, il ne pourra être étudié que si les astronomes parviennent à construire les instruments, satellites et télescopes, qui permettront d'observer l'Univers à très grande distance et donc l'observer tel qu'il était il y a très longtemps dans le passé, durant les époques où procède la Réionisation. Pour ces raisons la communauté effectue actuellement un très gros effort pour mettre en place des instruments tels que JWST (télescope spatial infrarouge, 2021), SKA (interféromètre radio au sol, 2025+) ou bien ATHENA (télescope spatial à rayons X, 2028+) durant la prochaine décennie : ces instruments seront capables de voir ces époques distantes et reculées, pour améliorer notre connaissance des processus à l'œuvre lors de la formation des toutes premières étoiles.

Côté théorie, un des outils de prédilection est la simulation numérique cosmologique. Grâce à des codes dédiés, capables de résoudre de façon couplée les processus de gravitation, de mécanique des fluides, de formation d'étoiles et de propagation du rayonnement, les groupes tels que celui de l'Observatoire de Strasbourg produisent des Univers « synthétiques » qui cherchent à reproduire la physique à l'œuvre dans toute sa complexité. L'objectif de ces études sur simulations est d'essayer, d'une part, d'anticiper quels paramètres et processus physiques dans les premières galaxies et étoiles vont influencer le cosmos dans son entier pour le réioniser. D'autre part, ces productions permettent de prédire dans quelle mesure les observations de la réionisation sont des contraintes efficaces sur la physique en place aux petites échelles.

Notre projet et ses résultats

Dans le cadre du Grand Challenge, notre groupe a contribué à cet effort théorique en produisant ce type de simulations sur le super-calculateur Jean-Zay. Pour cela nous avons utilisé le code de simulation cosmologique EMMA, développé à l'Observatoire Astronomique de Strasbourg. Ce code, dédié à la simulation de la réionisation, est capable de résoudre les physiques à l'œuvre durant ces époques (expansion de l'Univers, gravitation, mécanique des fluides, formation stellaire, chimie et refroidissement de l'hydrogène) et notamment la physique de la propagation du rayonnement et de son interaction avec la matière (appelée aussi transfert radiatif). Cette dernière physique est particulièrement exigeante à cause de ses temps caractéristiques très courts et elle domine le bilan final des calculs réalisés lors de telles simulations. A priori le coût associé en heures de calculs devrait être prohibitif, rendant impossible le type de simulation que nous visons, à moins d'approximations parfois grossières. Grâce aux GPUs, qui accélèrent les calculs de cette physique du rayonnement d'environ un facteur 10, ce coût reste sous contrôle dans le code de simulation EMMA. Cela nous permet de réaliser des simulations uniques dans la communauté et sans compromis, à très grands nombre d'éléments de résolution. EMMA est un code massivement parallèle, capable d'être déployé sur un grand nombre de coeurs CPUs et de cartes graphiques GPUs : l'accès à Jean-Zay, avec ces multiples coeurs et surtout de multiples cartes graphiques de dernières génération, nous ont ainsi permis la production d'un jeu de simulation de grande ampleur.

Nous avons ainsi produit 5 simulations de l'époque de Réionisation : chacune d'entre elle modélise un cube d'Univers différent de 550 millions d'années lumières de côté, résolu initialement avec un milliard d'éléments de résolution, nombre qui double en fin de simulation grâce au raffinement adaptatif de maille (voir fig. 1 pour une sortie de simulation). Lors de notre grand Challenge, chacune de ces simulations a été déployée sur 4096 coeurs CPU et 512 GPUs de la partition accélérée de Jean-Zay.

Ces simulations reproduisent la Réionisation dans toute sa complexité et cet ensemble constitue d'ores et déjà un jeu de donnée dont les exploitations peuvent être multiples, voire même non-anticipées. Par exemple, ces simulations ont déjà été exploitées par une équipe de l'Institut d'Astrophysique Spatiale d'Orsay pour prédire l'impact de la Réionisation sur le signal du Fond Diffus Cosmologique. L'émergence d'électrons libres durant cette époque est susceptible de diffuser les photons du rayonnement fossile et l'analyse des modifications imprimées sur ce dernier permet de remonter à l'histoire de la Réionisation et la géométrie de son réseau de bulles ionisées. Cette étude, soumise au journal *Astronomy & Astrophysics*, a utilisé le jeu de simulations produites lors de ce grand challenge pour mettre de nouvelle contraintes statistiques sur l'observation de ce processus.

Cet ensemble de simulations est également mis à profit actuellement pour étudier la topologie du processus de Réionisation. Grâce à des outils de théorie de Morse, qui permettent de caractériser les points critiques des cartes de réionisation (maxima, minima, points selles), de construire le « squelette » du réseau de régions ionisées ou bien de segmenter les simulations en zones d'influence des sources de lumières, nous sommes en train d'étudier la géométrie de la propagation du rayonnement et sa relation avec la distribution de matière sous-jacente.

Enfin, et c'est le coeur de la proposition de notre grand challenge, nous avons utilisé ce jeu de simulations pour pousser plus loin les performances de notre émulateur de simulation numérique de la Réionisation. Cet émulateur utilise des réseaux profonds, de type réseaux de convolution, pour prédire la distribution et l'évolution temporelle des bulles ionisées de la Réionisation, *à partir de la seule donnée de la distribution des étoiles dans une simulation et sans simuler la physique extrêmement coûteuse du transfert radiatif*. Nous avons démontré l'an dernier que ce type de réseaux parvient à émuler cette physique et à reproduire de façon satisfaisante l'histoire de la distribution des bulles ionisées en quelques minutes là où une simulation complète met plusieurs jours sur des configurations matérielles de type super-calculateur. Le point crucial cependant est la disponibilité de données d'apprentissage pour entraîner ces réseaux profonds. Le grand challenge nous aura permis de multiplier par 3 le nombre de jeu de données. Par ailleurs, ces processus d'apprentissage de nos émulateurs se font sur GPUs et Jean-Zay aura ainsi permis une convergence entre la production de données d'apprentissage et l'apprentissage lui-même sur un seul et même site, de façon très efficace. Nous avons d'ores et déjà pu constater que ces émulateurs sont désormais capables de reproduire un plus grand nombre de configurations et d'histoires de réionisation de façon satisfaisante **(voir fig. 2)**. Nous sommes actuellement en train de les exploiter pour réaliser des prédictions rapides, et en grand nombre, d'observations radio de l'époque de Réionisation.

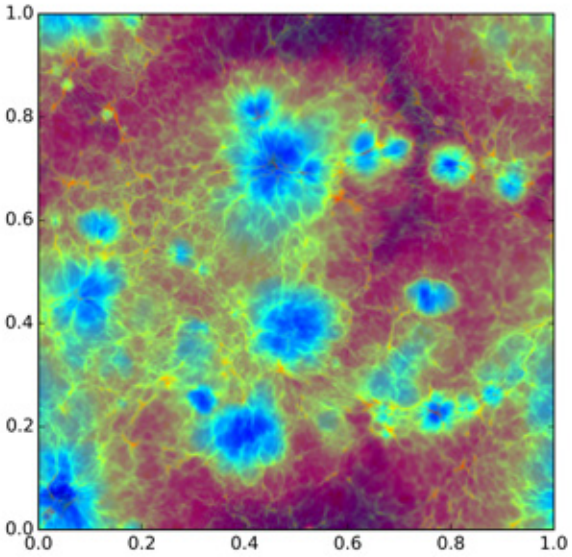


Figure 1 : *une des simulations réalisée lors du GC. Cela représente la distribution de gaz filamentaire dans un Univers agé de 1 milliard d'année, avec en surimpression la distribution des bulles de rayonnement ionisant, créées par les premières étoiles. L'échelle transverse représente environ 500 millions d'années-lumières de côté.*

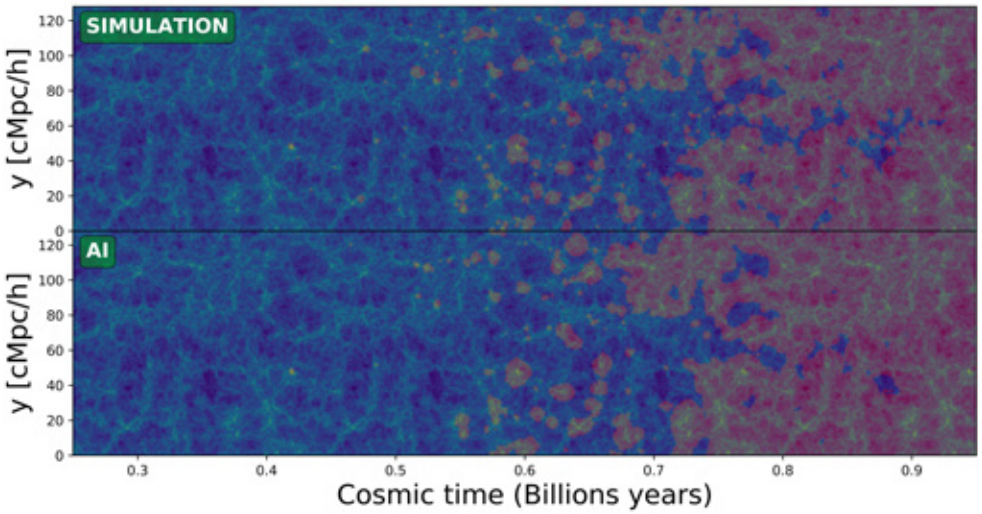


Figure 2 : *une frise temporelle qui montre l'évolution de la distribution et de la taille des bulles de réionisation au cours du premier milliard d'années de l'Univers. En rouge les régions ionisées, superposées à l'évolution de la distribution de matière en bleu. La première frise a été obtenue sur l'une des simulations du grand challenge, la seconde a été obtenue en quelques minutes par un réseau profond à partir de la distribution de matière seulement, entraîné sur les données produites durant le GC.*

Simulation haute fidélité d'accélération d'électrons par sillage laser.

Arnaud Beck[2], Imene Zemzemi[2], Julien Derouillat[1], Francesco Massimo[2], Mathieu Lobet[1], Arnd Specka[2]

[1] *Maison de la Simulation*
[2] *Laboratoire Leprince-Ringuet*

Léquipe (de gauche à droite) :
Mathieu Lobet, Arnd Specka, Imene Zemzemi, Arnaud Beck, Julien Derouillat, Francesco Massimo.

Contexte scientifique et objectifs

L'accélération par sillage laser [1, 2] est une technique d'accélération de particules imaginée dans les années 70 et mis en œuvre à partir du début des années 2000.
Elle permet des accélérations foudroyantes près de 1000 fois supérieures aux accélérations obtenus dans les accélérateurs de particules « standards » d'aujourd'hui qui utilisent des cavités radio-fréquence. L'enjeu sociétal est donc une réduction significative de la taille de ces accélérateurs. Des applications peuvent alors être envisagées en imagerie ou en médecine.

Le cœur de cette étude est donc de produire des simulations en 3D sans hypothèse simplificatrice et à très haute résolution.

Cette technique consiste à créer une onde plasma dans le sillage d'un laser ultra-intense se propageant dans un gaz sous-critique (transparent à la longueur d'onde du laser) et à piéger des électrons dans cette onde. Les électrons piégés dans l'onde plasma vont, à la manière de surfeurs accélérés par une vague, pouvoir récupérer une partie de l'énergie sous forme d'énergie cinétique. Ce faisant, les électrons atteignent quasi-instantanément des énergies relativistes et le paquet d'électrons injectés devient un faisceau d'électrons relativistes. Les électrons et l'onde plasma se propagent tous deux à une vitesse proche de celle de la lumière dans le vide et l'accélération continue donc tant que dure l'interaction entre eux.
Pour atteindre de très hautes énergies, l'un des grands challenges d'aujourd'hui est de prolonger cette interaction le plus longtemps possible. Pour cela, l'onde plasma doit être maintenue mais aussi être stable et symétrique pour ne pas dégrader la qualité du faisceau d'électrons. Pour les applications visées, il est en effet crucial de produire des faisceaux à très faible émittance et le plus mono-énergétique possible. Enfin, il est impératif d'être capable de reproduire le même faisceau d'électron à chaque tir laser.

Mais, le laser utilisé pour générer cette onde n'est lui même pas parfaitement symétrique et présente des variations tir à tir. L'objectif de ce grand challenge est d'étudier dans quelle mesure les imperfections d'un laser peuvent impacter le faisceau d'électrons résultants.

Résultats

Cette étude numérique a pour particularité de se faire en collaboration avec l'équipe expérimentale du laser Apollon. Celle-ci nous fournit les éléments de description du profil laser les plus précis possibles afin de le restituer au mieux dans des simulations haute-fidélité. Ainsi, il est question ici, non de simulations standards de lasers parfaitement gaussiens, mais de simulations de lasers réalistes dont le profil est exactement celui mesuré sur site.

Afin de pouvoir discerner quels sont les paramètres du laser qui importent et quels sont ceux qui n'ont que peu d'influence, la même simulation est reproduite trois fois en introduisant progressivement des asymétries et des imperfections. 14 millions d'heures ont été attribués pour ce projet. Les simulations sont faites à résolution élevée pour prendre en compte les imperfections les plus fines des mesures et utilisent chacune 12 000 cœurs de calcul (600 nœuds de Jean-Zay). Pour illustration, la figure 1 compare le laser parfait (super-gaussien), utilisé comme référence, au laser « réaliste » tel que reproduit dans la simulation à partir des données expérimentales.

Comme on le voit sur cette figure, les imperfections du laser sont complexes et largement visibles même après une longue propagation. Les approches simplificatrices généralement faites pour accélérer le calcul sous la forme d'hypothèse sur la géométrie ou la lente variation du laser dans le temps ne sont peut-être pas capables de rendre fidèlement les effets de ces défauts. Le cœur de cette étude est donc de produire des simulations en 3D sans hypothèse simplificatrice et à très haute résolution.

La comparaison des 3 simulations nous permet alors de déterminer quels paramètres jouent un rôle critique. Ainsi par exemple, il apparaît désormais évident que la phase joue un rôle primordial. La figure 2 illustre une comparaison de la densité électronique dans le sillage du laser pour les 3 configurations. Dans le cas d'une enveloppe laser symétrique mais d'une phase réaliste, on note déjà de très fortes perturbations qui ne sont guère accentuées lorsqu'une enveloppe asymétrique est utilisée. C'est donc bien l'introduction d'imperfection dans la phase qui influence le résultat final au premier ordre.

Perspectives

Au delà des résultats de physique déjà très intéressants, les données récoltées lors de cette campagne vont pouvoir dorénavant servir d'étalon à toute une gamme d'autres méthodes numériques.

En effet, les techniques utilisées dans ces simulations sont reconnues comme étant très fidèles mais d'un coût très élevé et possible uniquement dans le cas d'un grand challenge.
Posséder ces données va dorénavant nous permettre de discriminer parmi la multitude de modèles réduits dont nous disposons lesquels reproduisent au mieux la simulation de référence. Cela nous indiquera aussi sous quelles conditions ils sont capables de le faire. En effet, à ce jour personne ne connaît vraiment la validité de la plupart des modèles numériques simplifiés dans le cas d'un laser réaliste.
A ce jour, une doctorante travaille déjà à l'établissement du nombre nécessaire de modes azimutaux pour reproduire ces résultats dans une simulation quasi-cylindrique beaucoup moins coûteuse.
Ce travail est fondamental car si nous parvenons à établir un modèle réduit reproduisant fidèlement ces résultats, des études paramétriques beaucoup plus poussées pourront être effectuées.

[1] T. Tajima, J.M. Dawson, Phys. Rev. Lett. 43 (4), 1979, p. 267.

[2] E. Esarey, C. B. Schroeder, W. P. Leemans, Rev. Mod. Phys. 81, 2009, p. 1229

Figure 1 : *Comparaison entre un laser parfait (demi plan droit) et le laser mesuré (demi plan gauche), après 1 mm de propagation dans le plasma*

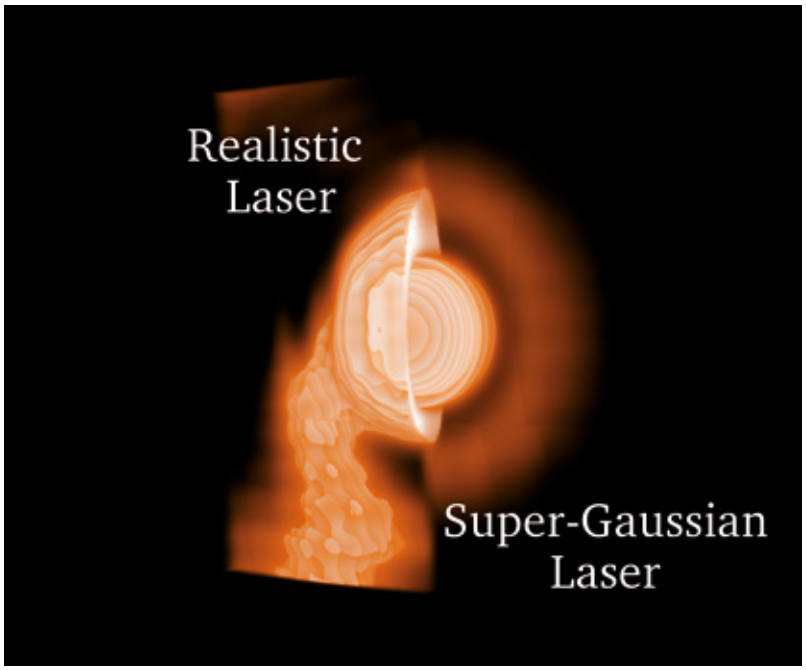
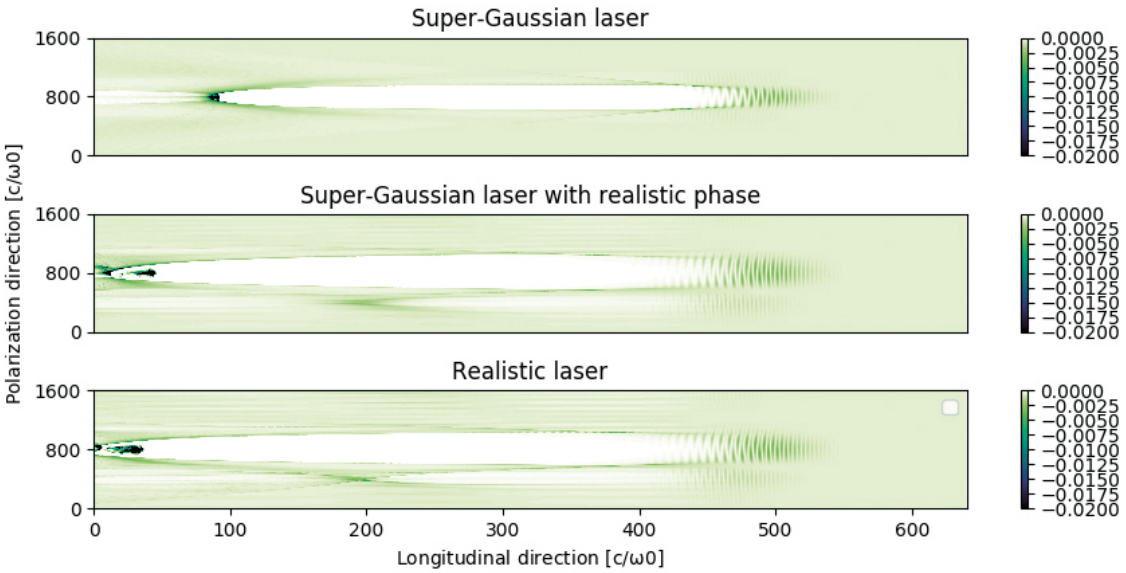


Figure 2 : *Densité électronique dans le sillage du laser pour les trois configurations différentes.
Panel du haut: enveloppe super gaussienne et phase plane.
Panel du milieu: enveloppe super gaussienne avec phase mesurée.
Panel du bas :
enveloppe et phase mesurées.
La densité est exprimée
en unité de densité critique*



Une nouvelle physique fondamentale est-elle nécessaire pour expliquer la mesure du moment magnétique du muon ?

Laurent Lellouch, directeur de recherche CNRS,
Centre de physique théorique, CNRS, Aix-Marseille U., U. de Toulon

Le modèle standard de la physique des particules, qui fournit une description quantique de toutes les particules subatomiques connues, est l'une des grandes réalisations scientifiques du 20ème siècle. Le boson de Higgs, dont le modèle prédit l'existence depuis 1964, a finalement été observé au *Large Hadron Collider* (LHC) du CERN en 2012. Avec cette observation, les particules du modèle standard sont au complet. Malgré cela, des mesures observationnelles et expérimentales, ainsi que certains aspects théoriques du modèle, suggèrent qu'il est incomplet. Des scientifiques du monde entier recherchent donc des indices susceptibles de nous indiquer ce qui pourrait se trouver au-delà. Une de leurs approches consiste à mesurer des propriétés de particules élémentaires avec une précision extrême et à comparer ces mesures avec des prédictions du modèle standard, tout aussi précises : tout désaccord significatif signalerait la découverte d'une nouvelle physique fondamentale.

Pour la première fois, un calcul ab initio de cette contribution rivalise en précision avec l'approche usuelle, basée sur l'exploitation théorique de données expérimentales.

Une propriété particulièrement prometteuse concerne le muon, un cousin éphémère de l'électron. Comme toute particule élémentaire ayant un spin et une charge électrique, le muon se comporte comme un minuscule aimant caractérisé par un moment magnétique. Aujourd'hui, ce moment magnétique est mesuré et prédit avec une précision relative d'environ $6 \times 10^{(-10)}$!
À ce niveau de précision, pratiquement tous les aspects du modèle standard sont impliqués dans la prédiction théorique, au travers de fluctuations quantiques. Selon les méthodes de calcul usuelles, les résultats expérimental et théorique sont en désaccord d'environ 3 à 4 écarts-types, un désaccord suggestif, mais encore trop peu important pour sceller le sort du modèle standard. Heureusement, une nouvelle expérience est en cours au *Fermi National Laboratory* de Chicago, qui vise à réduire l'erreur de la mesure d'un facteur 4 au cours des 2 prochaines années. Une autre expérience en cours d'élaboration au J-PARC d'Ibaraki, au Japon, a des objectifs similaires. Si les prédictions du modèle standard peuvent aussi être améliorées et rendent le désaccord encore plus important, une nouvelle physique fondamentale aura été découverte.

C'est dans ce contexte que notre équipe a complété un calcul ab initio de la contribution qui limite le plus la précision de la prédiction du moment magnétique du muon dans le modèle standard. Cette contribution, connue sous le nom de polarisation hadronique du vide (HVP), est due à des effets hautement non linéaires de la force nucléaire forte, qui ne peuvent être calculés « à la main ». Son calcul nécessite de résoudre les équations de la chromodynamique quantique (QCD), avec de l'ordre de 10^9 variables. Grâce aux développements théoriques, algorithmiques et numériques, qu'ont connu le domaine de la QCD sur réseau ces dernières années, des superordinateurs massivement parallèles, tels Turing et Jean-Zay à l'IDRIS, ont pu être exploités très efficacement pour relever cet énorme défi [1,2]. En particulier, la puissance de Jean-Zay a permis de calculer de petites corrections électromagnétiques absolument nécessaires pour atteindre la précision visée de quelques pour mille sur la contribution du HVP. Pour la première fois, un calcul ab initio de cette contribution rivalise en précision avec l'approche usuelle, basée sur l'exploitation théorique de données expérimentales. Notre calcul permet donc de confirmer ou d'infirmer le désaccord entre la mesure du moment magnétique du muon et sa prédiction dans le modèle standard. À notre grande surprise, notre résultat élimine tout besoin d'invoquer une physique fondamentale nouvelle pour expliquer la mesure expérimentale du moment magnétique du muon.
À moyen terme, les méthodes développées et utilisées pour ce travail permettront d'améliorer plus encore la précision de ce calcul, au fur et à mesure que la puissance des supercalculateurs augmentera.

Il est important de noter que ce calcul est d'une complexité rarement égalée en QCD sur réseau, et que la précision atteinte l'est aussi. Avant de tirer des conclusions définitives sur le sort du modèle standard, il est donc impératif que nos calculs soient refaits avec d'autres techniques et vérifiés par d'autres équipes. Si nos résultats sont confirmés, il sera alors très important de comprendre pourquoi ils ne sont pas en accord avec ceux donnés par l'approche traditionnelle. Et, bien sûr, nous sommes tous très impatients de découvrir les mesures expérimentales du moment magnétique du muon, des mois et des années à venir.

[1] *Budapest-Marseille-Wuppertal collaboration, Sz. Borsanyi et al, «Leading-order hadronic vacuum polarization contribution to the muon magnetic moment from lattice QCD,» arXiv:2002.12347 [hep-lat] (2020).*

[2] *Budapest-Marseille-Wuppertal collaboration, Sz. Borsanyi et al, «Hadronic vacuum polarization contribution to the anomalous magnetic moment of leptons from first principles,» Phys. Rev. Lett. 121 (2018) 022002 (Editors' Selection).*

Jean Zay (avec des ressources supplémentaires d'IDRIS-GENCI, du FZ Jülich, du Leibniz Supercomputing Centre München, du High Performance Computing Center Stuttgart).

Collaboration Budapest-Marseille-Wuppertal: CNRS, Aix-Marseille U., U. de Toulon, CPT, Marseille, France; Bergische Universität Wuppertal, Allemagne; Forschungszentrum Jülich, Allemagne; Université Eötvös, Budapest, Hongrie.

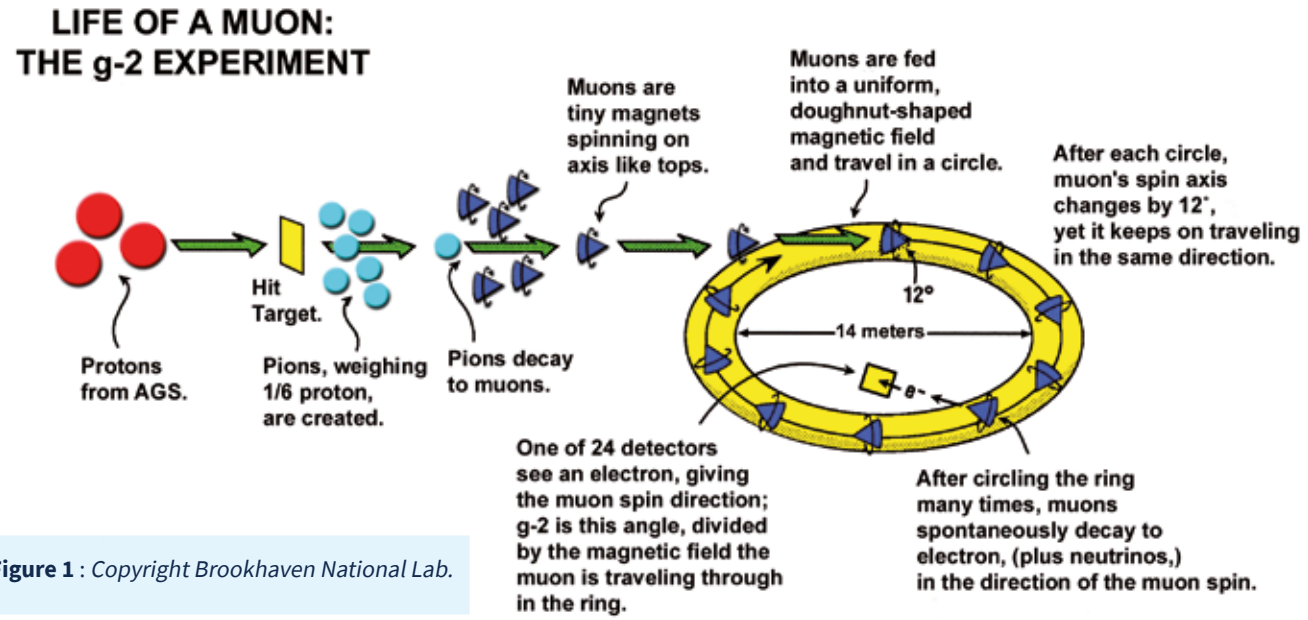


Figure 1 : *Copyright Brookhaven National Lab.*

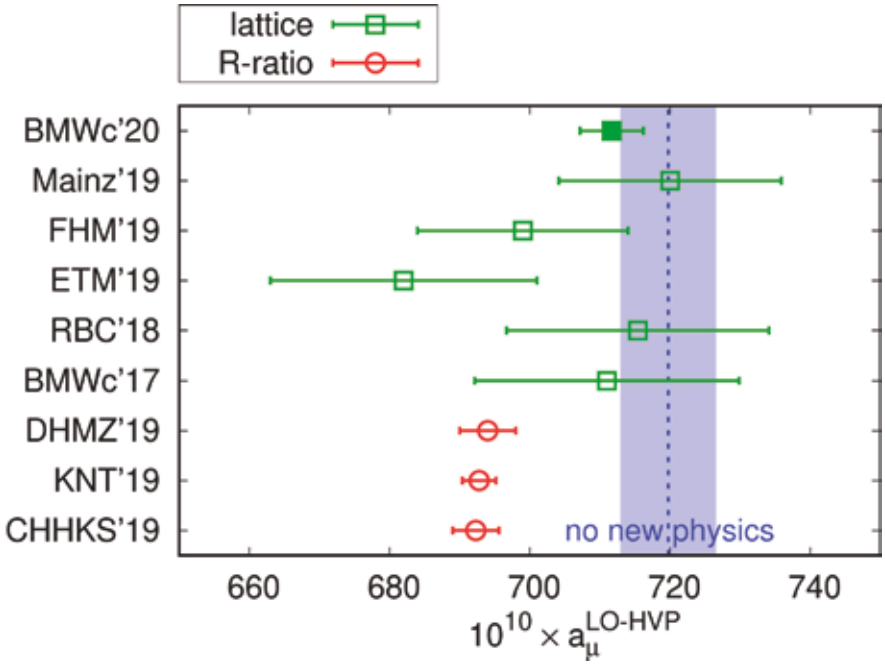
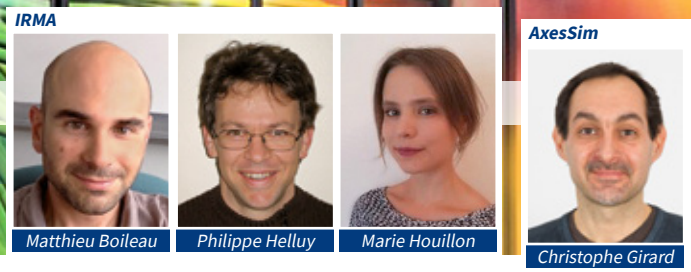


Figure 2 : *Comparaison de résultats récents pour la contribution de la polarisation hadronique du vide au moment magnétique du muon, divisée par deux. Les cercles rouges sont les résultats obtenus par la méthode traditionnelle.*
Les carrés verts sont ceux des calculs ab initio en QCD sur réseau. La bande bleue correspond à la valeur que devrait prendre cette contribution dans le modèle standard pour que celui-ci puisse expliquer la mesure expérimentale du moment magnétique du muon. Le carré vert plein est le résultat obtenu, en partie, grâce au Grand Challenge Jean-Zay 2019 et aux allocations 2018 et 2019 de GENCI [1]. Son incertitude est environ 4 fois plus petite que celle de notre précédent résultat, BMWc '17 [2], et que ceux des autres calculs en QCD sur réseau. Cette incertitude rivalise avec celle de la méthode traditionnelle et celle de la mesure expérimentale actuelle du moment magnétique du muon.



Simulation de l'interaction électromagnétique des objets connectés avec le corps humain

Matthieu Boileau, Philippe Helluy, Marie Houillon, IRMA
Christophe Girard, Nathanaël Muot, Guillaume Prin, Thomas Strub , Bruno Weber, AxesSim

Contexte

Ce projet avait pour but de simuler la propagation des ondes électromagnétiques produites par des petites antennes et leur interaction avec les tissus du corps humain. L'objectif à long terme est de proposer aux fabricants d'objets connectés des outils de simulation leur permettant d'optimiser le design électromagnétique de leurs produits. Une des retombées principales est l'amélioration de la durée de vie des batteries, ce qui ouvrirait la voie à de nombreux domaines d'applications : smartphones, sport, équipement médical, etc.
Les simulations électromagnétiques fournissent une autre information d'intérêt : le niveau d'exposition des tissus biologiques aux ondes. L'absorption par les tissus peut être précisément calculée afin d'évaluer leur échauffement et prédire les éventuels risques liés à l'utilisation d'objets rayonnants par l'homme. L'apport de la simulation numérique dans ce domaine est d'autant plus intéressant que la mesure expérimentale de ce type de grandeur est complexe.

À notre connaissance, une telle simulation du corps humain en temps long constitue une première et ouvre la voie à l'utilisation de l'approche développée dans Teta-CLAC dans un contexte prédictif avec des temps de retour compatibles avec les contraintes industrielles.

De telles simulations représentent un défi pour le calcul. Tout d'abord, gérer des géométries complexes impose des contraintes fortes sur les méthodes numériques qui doivent être adaptées aux cas réels de l'interaction d'une antenne avec le corps humain. Par ailleurs, le grand rapport entre les plus petites échelles (0,25 mm pour l'antenne et les détails du corps humain) et les plus grandes (2 m pour le corps humain) nécessite un grand nombre d'itérations temporelles. Pour répondre à cette problématique, l'IRMA (CNRS UMR7501, Université de Strasbourg) et la société AxesSim ont développé ensemble le solveur Galerkine discontinu Teta-CLAC capable d'exploiter les ressources hybrides multi-noeuds CPU et multi-GPU en utilisant des méthodes d'ordre élevé.
Teta-CLAC résout les équations de Maxwell en 3D pour calculer l'émission d'une antenne Bluetooth Low Energy (BLE) à proximité d'un modèle de corps humain incluant le squelette et les organes avec un grand niveau de détail. Ce solveur utilisé aujourd'hui dans un cadre industriel est issu des travaux de deux thèses **[1-2]** à l'IRMA et en CIFRE avec AxesSim.

Configurations calculées sur Jean Zay

Les cas étudiés dans ce grand challenge s'inscrivent dans la continuité de la thèse de Bruno Weber **[2]**. À partir d'un modèle numérique de corps humain obtenu en scannant par tomodensitométrie un mannequin anthropomorphique complet produit par la société Kyoto Kagaku (**Fig. 1**), il s'agissait de simuler trois configurations :
• placement de l'antenne contre la nuque pour représenter l'utilisation d'un maillot connecté (pratique sportive) ;
• placement de l'antenne sur l'avant-bras pour représenter l'utilisation d'une montre connectée (usage quotidien) ;
• placement de l'antenne dans l'abdomen pour représenter la présence d'un senseur telle une gélule connectée (application médicale).

Taille minimale du problème :
• 26,8 millions de mailles (1,3 milliard de degrés de liberté) ;
• au moins 40 Go de mémoire vive GPU pour une simulation.

Passage à l'échelle sur Jean Zay

Dans un premier temps, une étude d'extensibilité a été effectuée dans la configuration cible du corps humain avec un haut niveau de détails géométriques. Le maillage du mannequin Kyoto comporte 12 organes, dont le squelette (**Fig. 1**). Une antenne BLE *(Bluetooth Low Energy)* de technologie LTCC *(Low Temperature Co-fired Ceramic)* et rayonnant à 2,4 GHz a été placée à proximité du corps.
La fréquence maximale considérée pour cette simulation est de 2,9 GHz (bande de 1 GHz autour de la fréquence nominale). À cette fréquence, les éléments du maillage dont nous disposons sont suffisamment petits pour être interpolés à l'ordre 1 en espace. Les efficacités MPI mesurées pour l'ordre d= 1 sont présentées dans le **tableau 1**.
L'efficacité MPI sur N GPU est le rapport du temps de calcul théorique sur N GPU (évalué à partir du temps de référence, ici obtenu sur 4 GPU) divisé par le temps de calcul effectif sur N GPU. L'efficacité obtenue sur le corps humain est comparée au cas simplifié d'un maillage cubique qui sert de référence.
L'efficacité est très bonne jusqu'à 32 GPU. Au-delà, on constate toujours une accélération mais l'efficacité est limitée par la multiplication du nombre d'interfaces combinée à la réduction du nombre de mailles par processus. En effet, plus d'interfaces signifient plus de noyaux OpenCL à calculer, moins de calculs par noyaux et plus de communications MPI, ce qui diminue les performances.

Influence du placement de l'antenne et de la propriété des tissus

Une étude paramétrique de la propagation des ondes électromagnétiques dans le corps humain (**Fig. 2**) a été réalisée dans le but de quantifier l'absorption par les tissus : rayonnement en champ proche, en champ lointain, calcul de débit d'absorption spécifique (DAS) selon le standard IEEE **[3]** Dans ces simulations, nous avons fait varier les paramètres suivants :

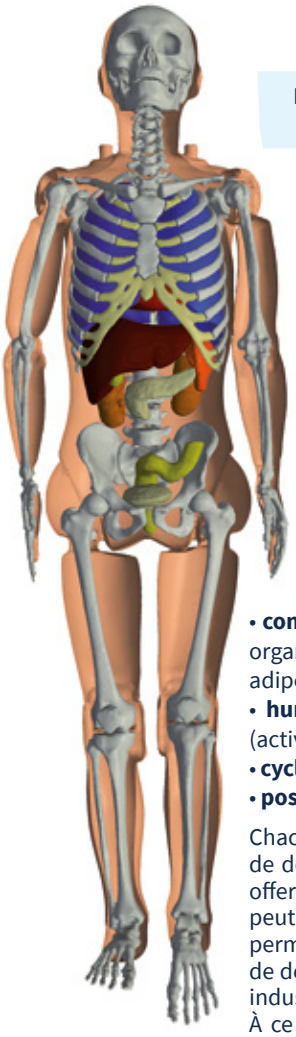


Figure 1 : *modèle numérique du mannequin Kyoto.*

Nombre de GPU (N)	Efficacité sur maillage cube de 256^3 mailles	Efficacité sur le corps humain
4	1	1
8	0.96	0.98
16	0.86	0.74
32	0.75	0.82
64	0.63	0.68
128	0.52	0.48
256	0.36	0.32

Tableau 1 : *Efficacité MPI du solveur Teta-CLAC*

• **composition de l'intérieur du corps** : homogène, avec organes et tissu musculaire ou avec organes et tissus adipeux ;
• **humidité de la peau** : sèche (au repos) ou mouillée (activité sportive) ;
• **cycle de respiration des poumons** : remplis d'air ou vides ;
• **position de l'antenne** : avant-bras, nuque ou ventre.

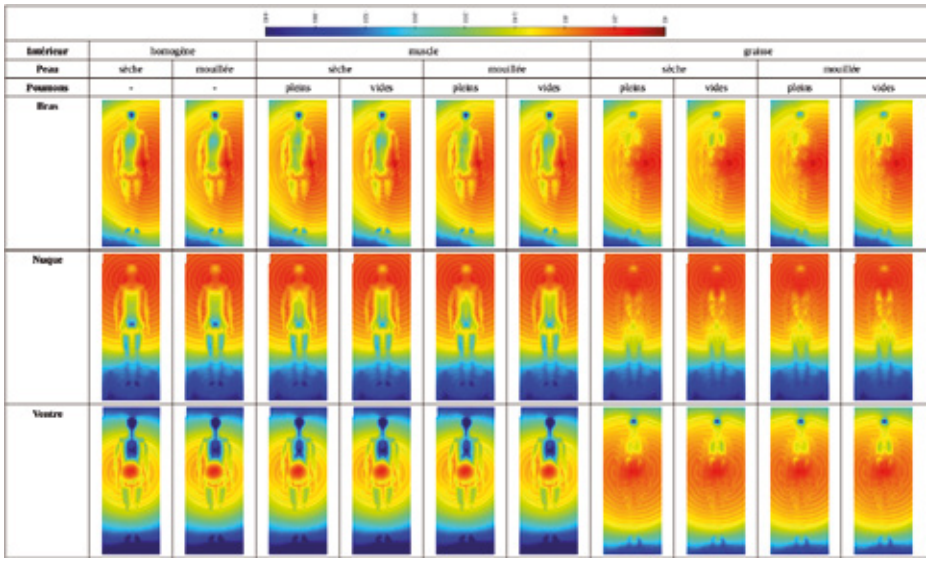
Chacune de ces simulations a pu être réalisée en moins de deux jours sur 64 GPU grâce à la puissance de calcul offerte par le calculateur Jean Zay. Ce temps de calcul peut être amené à moins d'un jour sur 256 GPU, ce qui permet d'envisager la simulation numérique comme outil de design et de validation des normes dans un processus industriel.
À ce jour, les résultats sont encore en cours d'exploitation mais ils devraient nous permettre d'identifier les paramètres ayant un réel impact sur la propagation des ondes ainsi que ceux pouvant être négligés afin de réduire au maximum le temps de calcul des simulations tout en garantissant des résultats de qualité.

Ces simulations ont également rendu possibles des études de DAS cumulé (multi-antennes) sur géométries complexes, résultats actuellement impossibles à obtenir par la mesure.

Conclusion

Du point de vue des mathématiques appliquées et du calcul intensif, ce projet a permis de valider le passage à l'échelle de l'approche numérique et algorithmique de type Galerkine discontinue sur un nombre de GPU encore jamais atteint avec le solveur Teta-CLAC.
À notre connaissance, une telle simulation du corps humain en temps long constitue une première et ouvre la voie à l'utilisation de l'approche développée dans Teta-CLAC dans un contexte prédictif avec des temps de retour compatibles avec les contraintes industrielles. Dans le domaine en pleine expansion des objets connectés, cette approche associée aux ressources de calcul hétérogènes les plus récentes représente une réponse prometteuse aux besoins de l'industrie.

Figure 2 : *Coupes du module du champ électrique (en dBV/m) calculé pour les différentes combinaisons de paramètres simulées.*



[1] Thomas Strub. Résolution des équations de Maxwell tridimensionnelles instationnaires sur architecture massivement multicoeur. Université de Strasbourg, 2015.

[2] Bruno Weber. Optimisation de code Galerkin Discontinu sur ordinateur hybride. Application à la simulation numérique en électromagnétisme. Université de Strasbourg, 2018.

[3] IEEE Recommended Practice for Measurements and Computations of Radio Frequency Electromagnetic Fields With Respect to Human Exposure to Such Fields, 100 kHz–300 GHz.

Les protéines membranaires monotopiques s'accumulent sur la surface des gouttelettes lipidiques

Vincent Nieto,
Luca Monticelli
University of Lyon, CNRS, Molecular Microbiology and Structural Biochemistry (MMSB, UMR 5086), F-69007, Lyon, France

Les gouttelettes lipidiques (GL) sont des organites de stockage de lipides, présentes dans la plupart des cellules vivantes [1]. Celles-ci sont formées quand la cellule a un excédent d'énergie et sont consommées quand la cellule a besoin d'énergie. La formation et la consommation de gouttelettes lipidiques sont cruciales pour le métabolisme cellulaire, et un certain nombre de maladies sont liées au dysfonctionnement du traitement des gouttelettes lipidiques dans les cellules, par exemple l'obésité, la stéatose, le diabète, le cancer et les infections virales [2]. La formation de gouttelettes lipidiques est généralement appelée biogenèse des gouttelettes lipidiques et se produit au niveau de la membrane du réticulum endoplasmique (RE). La biogenèse des GL commence par la synthèse de lipides neutres, tels que les triacylglycérols (TG) ou les esters de stérols, qui, à faible concentration, sont dissous dans la bicouche du RE.

Notre étude montre, pour la première fois, que les protéines hydrophobes ont tendance à s'accumuler de manière non spécifique sur les surfaces GL.

Lors de l'augmentation de la concentration, les lipides neutres se séparent des phospholipides de la membrane pour former une lentille d'huile, ou gouttelette naissante, dans la bicouche [3]. Au fur et à mesure que plus de lipides neutres sont synthétisés dans la cellule, la lentille se développe et émerge finalement dans le cytosol (c'est-à-dire à l'intérieur de la cellule) sous forme de GL mature : une gouttelette d'huile dans l'eau recouverte d'une monocouche de phospholipides avec des protéines incorporées. En effet, lors de l'émergence de la GL, de nombreuses protéines atteignent la surface de la GL [4]. Les protéines ciblant la surface de la GL peuvent être classées en deux groupes : les protéines de classe I, provenant de la membrane du RE, et les protéines de classe II, provenant du cytosol. Les deux classes de protéines sont nécessaires pour assurer un bourgeonnement de la GL approprié et un fonctionnement correct des GL. La manière dont les protéines se lient et s'accumulent spécifiquement à la surface des GL n'est pas bien comprise. La spécificité du ciblage des protéines sur les GL est un intérêt central de la biologie des GL, et la compréhension de ses principes fournirait des connaissances fondamentales sur le métabolisme lipidique et l'homéostasie cellulaire [5].

La composition lipidique et les propriétés physiques de la membrane régulent la distribution et la fonction des protéines à divers organites délimités par dans bicouches. L'interface GL-eau se distingue d'une interface bicouche-eau par plusieurs caractéristiques : elle peut supporter une couverture lipidique plus lâche ; l'épaisseur de la région hydrophobe sous-jacente est beaucoup plus grande que l'épaisseur hydrophobe d'une bicouche (~ 3 nm), jusqu'à des centaines de nm ; le noyau hydrophobe est constitué de lipides neutres, au lieu des chaînes acyles des phospholipides. Compte tenu des propriétés chimiques et physiques des deux interfaces, il n'est pas inattendu que les protéines montrent une préférence pour une interface par rapport à l'autre. La façon dont les propriétés de l'interface déterminent le partitionnement des protéines est largement inconnue.

Dans ce projet, nous avons commencé à répondre aux questions sur le partitionnement des protéines dans les gouttelettes lipidiques. Nous nous sommes concentrés en particulier sur les protéines de classe I, contenant un domaine hydrophobe inséré dans la membrane du RE. L'incorporation de telles protéines hydrophobes dans les bicouches lipidiques peut provoquer une perturbation locale des propriétés de la bicouche, ce qui se traduit par une pénalité énergétique [6]. Quant à l'insertion des protéines à la surface des GL, aucune information n'est disponible concernant le coût énergétique du procédé, ni le type ou l'étendue de la perturbation générée dans les lipides environnants. Nous avons étudié la façon dont des peptides hydrophobes simples se répartissent entre une bicouche et une gouttelette lipidique contiguë. À cette fin, nous avons utilisé des simulations de dynamiques moléculaires (DM) à gros grain, basées sur le champ de force MARTINI [7, 8]. La modélisation gros-grain (CG, « coarse-grain » en anglais) représente une simplification utile des systèmes atomiques, permettant des simulations de plus grandes portions de matière sur des échelles de temps plus longues. Dans les modèles CG, plusieurs atomes sont regroupés dans une particule « virtuelle » qui interagit à travers un potentiel moyen. Compte tenu de la taille typique des GL, et de l'échelle de temps de leur dynamique, l'utilisation de modèles gros-grain a été essentielle pour notre projet. Le champ de force gros-grain MARTINI a été co-développé par le Dr Monticelli, et permet de reproduire semi-quantitativement de nombreuses propriétés des agrégats lipidiques. Il a été utilisé avec succès pour simuler divers systèmes membranaires et protéiques (révisé dans la référence [9]). Les simulations ont été effectuées avec le logiciel GROMACS [9] sur le supercalculateur Jean Zay, en utilisant entre 1 000 et 4 000 cœurs en parallèle.

Nous nous sommes penchés sur le cas des peptides de la famille KWALP ; nos peptides comportent 3 résidus de lysine à l'extrémité N-terminale, puis une séquence de résidus d'alanine et de leucine alternés, et 2 résidus tryptophane à l'extrémité C-terminale. Des protéines analogues ont également été utilisées par nos partenaires expérimentaux, qui ont pu quantifier la distribution des peptides marqués par

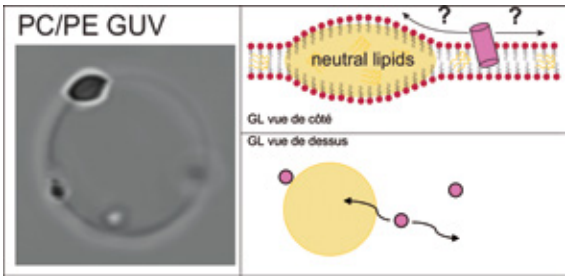


Figure 1 : *Image d'une gouttelette lipidique issue de microscopie à fluorescence. Représentation schématique d'une gouttelette lipidique et des protéines entrant et sortant de la GL.*

fluorescence dans des systèmes de gouttelettes lipidiques artificielles. Dans les simulations, nous utilisons 8 types de protéines différents, différents par la longueur de la portion hydrophobe (20 ou 28 résidus d'acides aminés au total), la présence ou l'absence de charges à l'extrémité C-terminale, l'orientation (parallèle ou antiparallèle) et l'attractivité des interactions peptide-peptide. Pour chacun des 8 types de peptides, nous avons effectué 2 séries de simulations, avec une concentration en peptides faible et élevée. Les protéines ont été insérées dans une bicouche lipidique modèle, constituée de lipides DOPC, et la bicouche comprenait une gouttelette lipidique naissante en forme de lentille (Figure 1). En raison de leur grand noyau hydrophobe et de l'extrémité N-terminal chargée, les peptides sont toujours insérés dans des membranes bicouches dans une orientation transmembranaire, l'extrémité N-terminale exposée à l'environnement aqueux. Les peptides ont été initialement placés dans la région bicouche du système, et ils étaient libres de se diffuser entre la région bicouche et la région GL. Cette configuration permet de déterminer la tendance spontanée de chaque type de peptide à se répartir entre la région bicouche et monocouche, sous condition d'un échantillonnage suffisant. L'échelle de temps typique pour la diffusion d'un peptide transmembranaire simple est de l'ordre de centaines de nanosecondes, nous avons donc exécuté toutes les simulations pour 20 µs chacune.

Nous avons constaté que toutes les protéines avec une extrémité C-terminus chargée restaient toujours dans la région bicouche, et ne se situait jamais dans la gouttelette lipidique. De plus, ces peptides transmembranaires sont accumulés au bord de la gouttelette - le point où la bicouche se divise en deux monocouches. Tous les peptides sont restés mobiles, il semble donc que leur préférence pour la région bicouche ne soit pas due à un échantillonnage limité. Pour les peptides chargés à la fois à l'extrémité C-terminus et à l'extrémité N-terminus, le partitionnement dans la GL nécessiterait que l'un des deux terminaux soit incorporé dans le noyau hydrophobe de la GL, ce qui est énergétiquement très coûteux pour les résidus chargés.

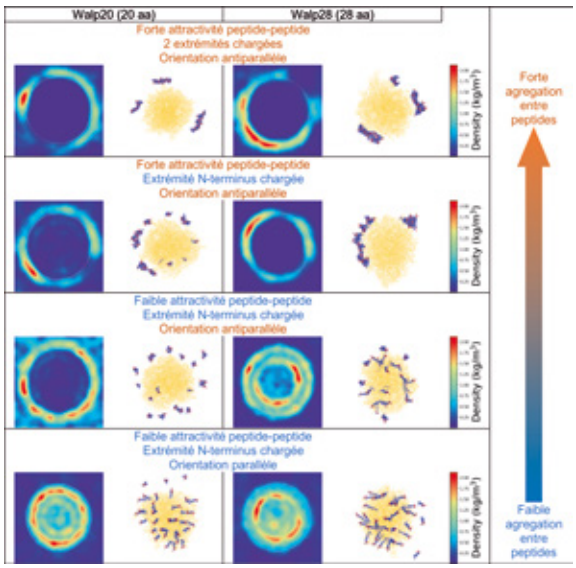


Figure 2 : *Pour les 8 types de peptides, à droite, vue de dessus d'un aperçu de la simulation de la GL et des peptides. À gauche : densité spatiale d'atomes de peptides de la simulation correspondante.*

D'autre part, les peptides avec une seule extrémité chargée (l'extrémité N-terminale) ont montré une nette tendance à se répartir dans la monocouche recouvrant la gouttelette lipidique. Une fois encore, tous les peptides sont restés très mobiles, et la distribution dans la région GL était dynamique : les peptides se sont déplacés à l'intérieur et à l'extérieur de la région GL au long des simulations. La tendance à se distribuer dans la région GL a diminué lorsque les peptides se sont agrégés dans la région bicouche, avant d'atteindre la GL ; les peptides insérés en orientation parallèle montraient une faible tendance à s'agréger, car leurs extrémités N-terminus chargées se repoussaient lorsqu'elles étaient à proximité ; les peptides anti-parallèles se sont agrégés plus favorablement, d'où leur entrée dans la GL moins fréquente. Un tel comportement s'explique facilement, une fois encore, par l'électrostatique : les dimères peptidiques anti-parallèles ont des charges électriques des deux côtés de la membrane, et entrer dans la région GL en tant que dimère implique de faire pénétrer une charge dans l'environnement hydrophobe de la gouttelette lipidique. Nous avons remarqué que plusieurs peptides seuls sont entrés dans la région GL même dans des simulations avec une orientation anti-parallèle. Les peptides à faible attraction mutuelle s'agrègent moins et se répartissent préférentiellement dans la région GL indépendamment de leur orientation ; lorsqu'ils sont insérés dans la membrane dans le même sens, ils se sont partitionnés presque exclusivement en monocouche recouvrant la gouttelette lipidique. L'origine de la distribution des protéines sur les gouttelettes lipidiques semble être dû à la préférence des chaînes latérales des acides aminés pour le triglycérol par rapport aux chaînes acyles des phospholipides. Les résultats expérimentaux sur des protéines analogues ont confirmé que la relocalisation des peptides KWALP dans la bicouche est défavorable, et les peptides se déplacent spontanément de la bicouche vers la région monocouche.

En conclusion, notre étude montre, pour la première fois, que les protéines hydrophobes ont tendance à s'accumuler de manière non spécifique sur les surfaces GL. Dans les cellules, un mécanisme de contrôle ou de modification des protéines au niveau de la membrane du réticulum endoplasmique serait nécessaire pour empêcher la relocalisation non spécifique aux GL des protéines contenant des domaines hydrophobes.

Références

[1] Walther TC & Farese RV, Jr. (2012) Annu Rev Biochem 81 : 687-714.
[2] Welte MA & Gould AP (2017) Biochim Biophys Acta 1862(10, Part B) : 1260-1272.
[3] Thiam AR & Forêt L (2016) BBA - Molecular and Cell Biology of Lipids 1861(8) : 715-722.
[4] Kory N, Farese RV, & Walther TC (2016) Trends Cell Biol 26(7) : 535-546.
[5] Dhiman R, Caesar S, Thiam AR, & Schrul B (2020) Semin Cell Dev Biol.
[6] Andersen OS & Koeppe RE (2007) Ann Rev Biophys Biomol Struct 36 : 107-130.
[7] Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, & de Vries AH (2007) J Phys Chem B 111(27) : 7812-7824.
[8] Monticelli L, et al. (2008) J Chem Theory Comput 4(5) : 819-834.
[9] Marrink SJ & Tieleman DP (2013) Chem Soc Rev 42(16) : 6801-6822.

L'équipe de l'IC2MP (de gauche à droite) : F. Guégan, H. Zhang, R. Larhlimi, G. Frapper et B. Wang (abs. S. Faraji Nafchi)

Prédiction de matériaux : la rencontre de Darwin et Schrödinger

Gilles Frapper, coordinateur du programme « Prédiction de matériaux in silico - Grands Challenges Jean Zay » Groupe de recherche « Chimie Quantique Appliquée », E4 IC2MP, UMR 7285 Université de Poitiers – CNRS Contact : gilles.frapper@univ-poitiers.fr

Erwin Schrödinger a clairement établi que le potentiel externe lié à la structure atomique détermine la fonction d'onde qui décrit les états quantiques d'un système. Aussi, la seule connaissance de la structure cristalline autorise la détermination de nombreuses propriétés physico-chimiques par des calculs en chimie quantique : étude des phonons, capacité calorifique, conductivité électrique, propriétés thermodynamiques, mécaniques (élasticité, dureté…), thermiques et thermoélectriques, de supraconductivité… Encore faut-il avoir accès à l'arrangement spatial des atomes dans la matière. Or, les blocages sont nombreux pour déterminer expérimentalement une structure cristalline. Citons-en trois :

- le taux de cristallinité faible de certains matériaux ne permet pas une caractérisation univoque de la phase…
- l'impossibilité expérimentale de sonder la matière sous certaines conditions de pression et de température. Or l'état de la matière au centre de planètes, d'exoplanètes, voire de la Terre où une pression de 360 GPa règne en son centre, ne répond pas aux règles « atmosphériques » usuelles. Sous haute pression, une nouvelle chimie s'invite, de nouvelles règles doivent être établies ;
- l'analogie structurale pour la prédiction d'une structure d'un composé se limite à l'existant et à l'imaginaire du scientifique. Cette démarche n'autorise pas un balayage exhaustif de l'infernal champ des possibles.

Aussi, la prédiction de nouvelles structures cristallines à partir de la seule connaissance de la composition chimique est un défi majeur en science des matériaux. L'enjeu est donc d'élaborer une méthodologie fiable dans la prédiction de structures cristallines par simulation numérique.

Aussi, la prédiction de nouvelles structures cristallines à partir de la seule connaissance de la composition chimique est un défi majeur en sciences des matériaux. L'enjeu est donc d'élaborer une méthodologie fiable dans la prédiction de structures cristallines (*ab initio* CSP, crystal structure prediction) par simulation numérique, validée par des applications variées.

Notre projet « **Prédiction de matériaux *in silico*** » - Grands Challenges Jean Zay 2019-2020 » s'inscrit dans cet axe de recherche. Nous employons un algorithme évolutionnaire (génétique) implémenté dans le code USPEX combiné à des calculs quantiques en théorie de la fonctionnelle de la densité (DFT, code VASP). Notre méthodologie USPEX/VASP est mise en œuvre sur le supercalculateur Jean Zay par une approche « *embarrasingly parallel* » où une multitude de jobs VASP (CPU) est réalisée en parallèle, sans grand effort de communication nécessaire entre les tâches concurrentes. Cependant, l'accès à un grand nombre de cœurs CPU est primordial. Le supercalculateur Jean Zay répond pleinement à notre demande dès lors que le temps d'attente entre deux tâches concurrente est réduit.

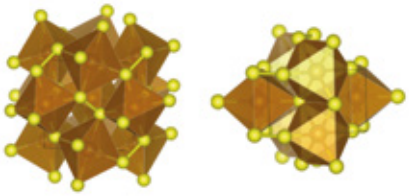
Au cours de cette période d'intenses calculs – 16 millions d'heures consommées sur Jean Zay en 4 mois –, plusieurs familles de matériaux ont été abordées par les équipes de recherche de Poitiers (IC2MP), Nantes (IMN), Rennes (ICR) et Moscou (MIPT et Skoltech). Citons les sels M_xN_y où des motifs polyazotés sont observés sous haute pression (ex. M = yttrium Y) ; la nouvelle phase polymorphique bidimensionnelle du disulfure de fer FeS_2, produit de désintercalation du matériau pour la conversion de l'énergie Li_2FeS_2 (article sous presse à *JPC Letters*, 2020) ; et de nouveaux hydrures $A_xB_yH_z$ et borocarbures $M_xB_yC_z$ explorés dans une gamme de pression de 0 à 50 GPa.

Un autre objectif est clairement affiché : établir par simulations numériques une banque de données de composés cristallins binaires « *in silico* », stables et métastables, dont l'analyse de leurs structures électroniques et géométriques invitera la communauté de la synthèse des matériaux à explorer telle ou telle composition chimique, à traquer le composé solide proposé. Cependant, ce programme nécessitera plusieurs millions d'heures cpu d'où notre intérêt à soutenir un parc informatique puissant et ouvert à la communauté de recherche, des centres de calcul de Service Public (GENCI, PRACE, …).

Avant d'évoquer l'association de concepts issus de la sélection naturelle chère à Darwin à la mécanique quantique de Schrödinger, exposons succinctement la problématique posée, « comment prédire une structure cristalline par simulation numérique ? ».

Accès à la structure cristalline par simulation numérique

Chaque matériau étudié est un cristal : il présente un motif qui se répète selon les trois directions de l'espace. Ce motif, nommé maille cristalline, peut être vu comme une boîte contenant plusieurs atomes arrangés de façon bien déterminée. Par exemple, le disulfure de fer FeS_2 présente plusieurs structures cristallines, appelées phases polymorphiques. La plus stable aux conditions standards de pression et de température est la pyrite ; une autre forme structurale, proche en énergie, est la marcasite (voir Figure ci-après).



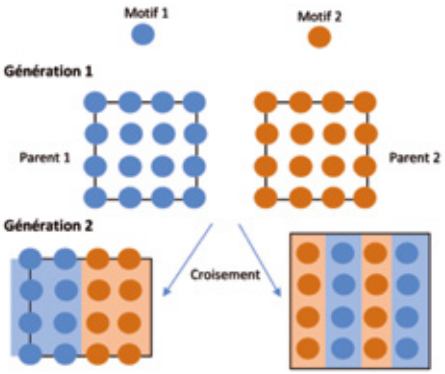
Deux phases polymorphiques de FeS_2. A gauche, la maille cubique de la pyrite ; à droite, la maille orthorhombique de la marcasite.

Dans ces deux phases, l'atome de fer se loge au centre d'un octaèdre de soufre, tandis que deux atomes de soufre sont intimement liés. Leurs structures diffèrent, entre autres, par l'arrangement relatif des octaèdres FeS_6.
Cet arrangement particulier est lié aux liaisons chimiques existantes entre ces atomes, et définit les propriétés macroscopiques de chaque phase FeS_2. Aussi, pour la compréhension d'un matériau cristallin, il est primordial de connaître l'arrangement des atomes, c'est-à-dire leur environnement : combien de voisins ? Forment-ils un octaèdre, un prisme… ? Quelles distances interatomiques ?
Cependant, comment accéder aux structures par simulation numérique ?
Le but est donc de déterminer, pour un nombre et un type donné d'atomes dans la maille, leur arrangement à 3 dimensions, c'est dire leurs positions (x,y,z) et les 3 vecteurs de répétition. Cet exercice est loin d'être trivial : il s'agit d'identifier la ou les structures cristallines les plus basses en énergie (composés thermodynamiquement stables et métastables), parmi des millions de structures possibles. Par exemple, la répartition aléatoire de 20 atomes dans une boîte cubique peut engendrer a priori 10^{25} structures cristallines différentes. Imaginons qu'il faille 1 seconde de temps de calcul pour déterminer numériquement l'énergie associée à chaque structure. Le temps global nécessaire pour déterminer les structures cristallines de plus basses énergies dépasserait alors l'âge de l'univers (10^{17} ans) ! La problématique est donc la suivante : comment accéder au composé cristallin stable le plus rapidement possible, en monopolisant un minimum de ressources informatiques ?

Algorithme (R)évolutionnaire

L'algorithme évolutionnaire USPEX est basé sur la théorie Darwinienne : des structures (motifs chimiques) sont croisées jusqu'à engendrer la structure optimale.

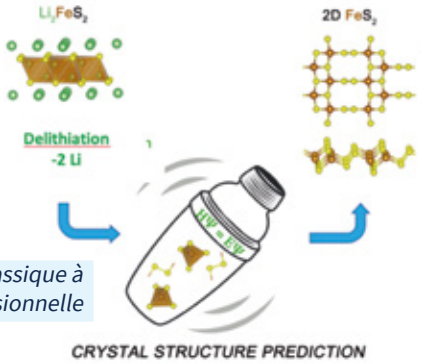
Récemment, A. R. Oganov a proposé une méthode pour déterminer l'arrangement le plus stable, implémentée dans le code USPEX (http://uspex-team.org). Ce programme informatique s'appuie sur un algorithme évolutionnaire basé sur la théorie Darwinienne. Quel est donc son principe ? Un fragment d'une structure cristalline peut s'associer à une brique élémentaire d'une seconde structure (voir Schéma ci-après).



Des liaisons sont créées entre ces deux fragments et, si ce nouvel arrangement est plus stable que ses « parents », cet « enfant » est autorisé « à se reproduire », c'est à dire à se combiner avec un autre motif d'une autre structure. Si cet arrangement est trop haut en énergie – donc instable énergétiquement –, il est écarté du cheptel des « reproducteurs potentiels ». Ainsi, de génération en génération, les meilleurs motifs, présentant des liaisons stables entre atomes du cristal sont rapidement identifiés. Notons que la structure cristalline et son énergie associée sont déterminées par résolution de la célèbre équation de Schrödinger $H\Psi=E\Psi$ (interactions électrons/électrons, électrons/noyaux, etc.) *via* une méthode de chimie quantique adaptée.
C'est l'étape excessivement gourmande en temps de calculs… De même, pour maintenir une grande diversité structurale et autoriser l'exploration complète du champ des possibles, il est intéressant d'injecter quelques structures « hautes en énergie » - donc non « parfaites » - dans la population. L'existence de « défaut », ici structural, est indispensable à la localisation de la structure la plus stable. Par cette approche auto-apprenante, le minimum global est atteint suite à l'optimisation de 1 000 à 3 000 structures cristallines au lieu des milliards de structures potentielles ! Ainsi, le design *in silico*, c'est-à-dire par simulation numérique, de nouveaux composés cristallins aux propriétés physico-chimiques originales est désormais possible.

Vers la découverte de nouveaux matériaux

Par cette approche, de nombreuses phases ont été proposées, dont quelques-unes synthétisées. Récemment, le groupe « Chimie quantique appliquée » de l'IC2MP a prédit des matériaux sous haute pression, TiN_2 et MgN_4 cristallins, qui par la suite ont été caractérisés expérimentalement.
Pour illustrer nos travaux « Grands challenges Jean Zay », revenons au disulfure de fer FeS_2 : une phase métastable – plus haute en énergie que la pyrite – a été synthétisée voilà plus de 30 ans dans les laboratoires nantais (IMN) à partir du précurseur Li_2FeS_2, sans aucune caractérisation DRX du fait de sa faible cristallinité. Notre exploration USPEX/VASP de formes métastables où le fer est connecté à quatre atomes de soufre - données expérimentales EXAFS et Mossbauër - a abouti à la proposition d'une structure cristalline en feuillet avec des haltères de S_2 et des ligands S, représenté sur la figure ci-après à droite. Ce résultat majeur est une étape dans l'étude des profils réactionnels du mécanisme de désintercalation du lithium au sein des phases $Li_{(2-x)}FeS_2$, tant d'un point de vue expérimental que théorique.



De Li_2FeS_2 massique à FeS_2 bidimensionnelle

Au-delà, le programme « Grands Challenges Jean Zay » a permis d'établir de fortes collaborations entre les chercheurs de l'IC2MP, de l'IMN et de l'ICR du pôle Ouest du Réseau Français des Chimistes Théoriciens (RFCT). Cet aspect structurant est à souligner.
Pour conclure, la prédiction de nouveaux matériaux par algorithme évolutionnaire et calculs quantiques permet la découverte de structures originales, par la seule connaissance de la composition chimique, du moins pour des systèmes chimiques de petite taille (<60 atomes par maille). De ces structures, il est ensuite possible de calculer une batterie de propriétés physico-chimiques : dureté, conductivité électrique, supraconductivité, thermoélectricité, réactivité de surface… Pour se faire, il est impératif que notre communauté de recherche puisse avoir accès à des supercalculateurs performants associés à des personnels hautement qualifiés (et en nombre suffisant…) : le supercalculateur Jean Zay y répond, restera à maintenir ce haut niveau de moyens et compétences à GENCI !

Références

[1] B. Huang et G. Frapper. Pressure-Induced Polymerization of CO_2 in Lithium–Carbon Dioxide Phases, J. Am. Chem. Soc. 2018, 140, 413−422.
[2] B. Wang *et al*. Prediction of a New Layered Polymorph of FeS_2 with $Fe^{3+}S^{2-}(S_2^{2-})_{½}$ Structure. The Journal of Physical Chemistry Letters, 2020, sous presse.
[3] Ce texte est largement inspiré d'un article de vulgarisation paru dans la revue Microscoop du CNRS Nouvelle Aquitaine. « Quand Darwin rencontre Mendeleïev ».
R. Larhlimi, F. Guégan et G. Frapper, Microscoop février 2019, p.16-17.

Au cœur des effets quantiques nucléaire : hydrogène à haute pression et ondes à densité de charge en dimension réduite

Matteo Calandra
Institut des Nanosciences de Paris, UMR7588, CNRS, Sorbonne Université.

Les systèmes à onde à densité de charge et les systèmes basés sur des atomes légers sont dans la limite quantique extrême car ils présentent soit (i) des transitions structurelles à des températures proches du zéro (transition de phase quantique), (ii) soit des faibles masses ioniques et donc des forts effets quantiques. La description de ces systèmes nous pousse à aller donc au-delà des connaissances actuelles de physique de matériaux et à développer des nouvelles méthodes de calcul. Notre équipe à l'Institut des Nanosciences de Paris, CNRS et Sorbonne Université, en collaboration avec l'Université des Pays Basques et l'Université de Rome, a développé une méthode pour traiter les systèmes fortement anharmoniques et avec des effets quantiques nucléaires importants.

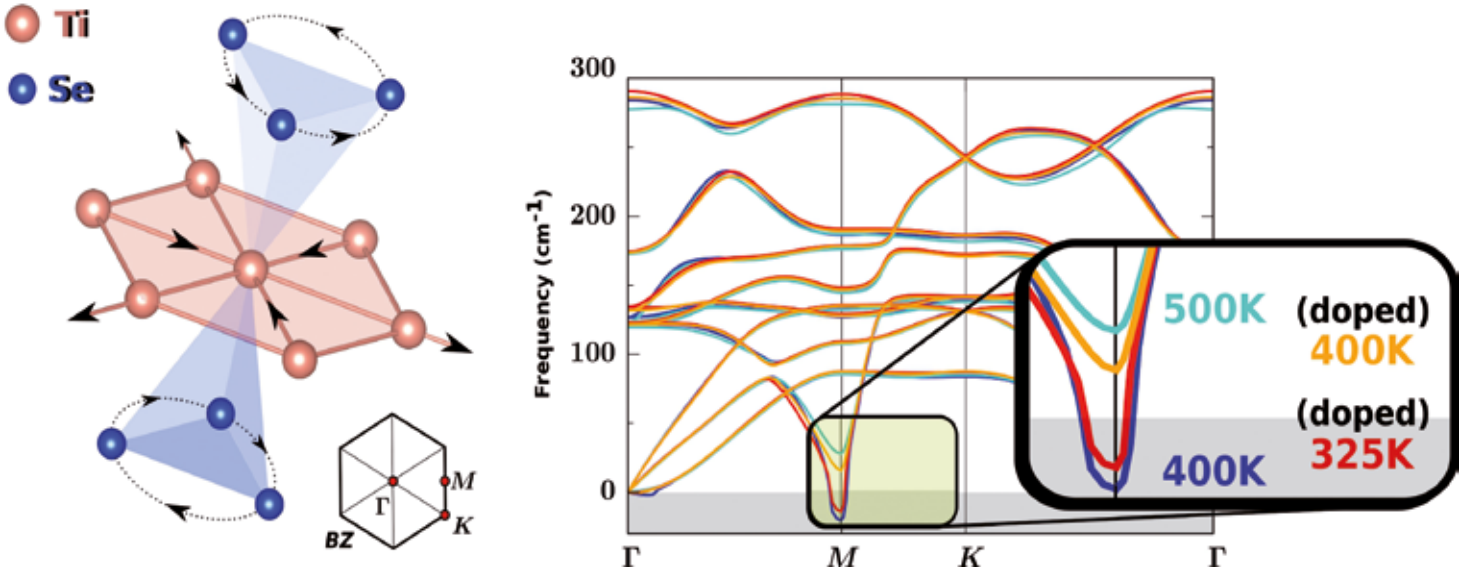
La métallisation de l'hydrogène est actuellement le problème le plus important de la physique des hautes pressions.

Dans le cadre de notre projet Grand Challenge Jean Zay, nous avons appliqué cette méthode à l'Hydrogène à très hautes pression et aux ondes à densité de charge dans les cristaux bidimensionnels. Les résultats les plus marquants sont les suivants.

La métallisation de l'hydrogène est actuellement le problème le plus important de la physique des hautes pressions. Plusieurs expériences récentes ont montré des signes des métallisations à des pressions critiques qui varient entre 360 et 490 GPa, selon le type de mesure (transport, infrarouge). Nous avons calculé les propriétés vibrationnelles, Raman, infrarouges et optiques de la phase III de l'hydrogène en incluant les effets quantiques nucléaires [1]. Nous avons démontré que les fluctuations quantiques des ions réduisent les énergies du vibrons de 25% et le gap optique de 3 eV.

Nous avons aussi montré que la métallisation a lieu à 380 GPa en accord avec les mesure de transport.
Enfin, nos calculs d'optique montrent que la phase métallique de l'hydrogène est noire et transparente dans l'infrarouge jusqu'à 450 GPa. Étant donné que l'hydrogène devient métallique avant de devenir opaque à l'infrarouge, ceci explique les mesures contradictoires.
Nous avons aussi étudié le rôle des effets quantiques nucléaires et de l'anharmonicité non-perturbative dans plusieurs monocouches de dichalcogénures [2,3] et en particulier dans un mono feuillet de $TiSe_2$. Les systèmes à basse dimensionnalité avec des faibles gaps électroniques (< 0.2 eV) et une forte attraction électron-trou ont été proposés comme des possibles candidats pour les phases onde à densité de charge de nature excitoniques. Ceci est dû au fait que, conventionnellement, l'énergie de liaison de l'exciton augmente en dimension réduite (pour des bandes paraboliques). Le $TiSe_2$ massif est souvent considéré comme un cas d'école car à température ambiante c'est un semi-conducteur avec une très faible bande interdite et à 200 K il présente une onde à densité de charge. Récemment le monofeuillet de $TiSe_2$ a été synthétisé par plusieurs équipes. Il a un faible gap et, à des températures qui varient entre 200 et 280 K, il montre une onde à densité de charge avec une périodicité 2x2 de la maille élémentaire (voir Fig.1). Donc, logiquement, on devrait s'attendre à des effets très importants de l'interaction électron-trou sur les spectres vibrationnels par rapport au cas massif. Nous avons démontré, au contraire, que les effets de l'interaction électron-trou dans les spectres vibrationnels sont très faibles. La transition vers l'état onde à densité de charge est dominée par l'anharmonicité et le dopage intrinsèque de type n dû à la stœchiométrie imparfaite. Le dopage joue un rôle très important dans la réduction de la température critique, un résultat totalement inattendu.

Figure 1 : *Structure d'un monofeuillet de $TiSe_2$. Les flèches montrent la distorsion à basse température qui caractérise son onde à densité de charge (à gauche). Spectres vibrationnels d'un monofeuillet de $TiSe_2$ en fonction de la température et du dopage (à droite).*



Références

[1] *Black metal hydrogen above 360 GPa driven by proton quantum fluctuations,*
Authors: Lorenzo Monacelli, Ion Errea, Matteo Calandra, Francesco Mauri
Nature Physics, 10.1038/s41567-020-1009-3 <https://doi.org/10.1038/s41567-020-1009-3>

[2] *Anharmonicity and doping melt the charge density wave in single-layer TiSe2,*
Authors: Zhou, Jianqiang Sky; Monacelli, Lorenzo; Bianco, Raffaello; Errea, Ion; Mauri, Francesco; Calandra, Matteo
Nano Letters. 20, 7, 4809–4815 (2020)
https://doi.org/10.1021/acs.nanolett.0c00597

[3] *Theory of the thickness dependence of the charge density wave transition in 1T-TiTe2,*
Authors: Zhou, Jianqiang Sky; Bianco, Raffaello; Monacelli, Lorenzo; Errea, Ion; Mauri, Francesco; Calandra, Matteo
2D materials sous presse
https://iopscience.iop.org/article/10.1088/2053-1583/abae7a

La lacune dans CdTe, un challenge pour les calculs de structure électronique

Damien Caliste
Department of Physics, IRIG, Univ. Grenoble Alpes and CEA, F-38000 Grenoble, France.

Les alliages II-VI sont des semi-conducteurs à gap direct aux propriétés optiques intéressantes. Le HgCdTe, par exemple, est utilisé comme matériau actif dans les détecteurs infra-rouge. Il peut être facilement dopé type p grâce aux lacunes de mercure. Comme pour les capteurs photo dans le visible, les détecteurs infra-rouge utilisent des matrices de pixels. Vouloir augmenter la résolution, implique souvent une augmentation du nombre de pixel défectueux ou clignotant. Ces pixels clignotants sont thermiquement activés et leur origine est liée aux défauts présents dans HgCdTe.

La possibilité d'effectuer ces N^2 calculs (N étant le nombre de fonctions d'ondes) sur GPU est un atout considérable.

Actuellement, il n'existe que peu de références dans la littérature concernant l'étude à l'échelle atomique du HgCdTe et de ses défauts ponctuels (bien que plusieurs références existent pour CdTe [1, 2]). Les résultats disponibles reposent souvent sur des approximations assez importantes notamment sur les tailles des systèmes simulés. Ainsi, la simulation ab initio des matériaux II-VI reste un challenge de par la capacité des approximations utilisées en théorie de la fonctionnelle de la densité (DFT) à reproduire correctement leur structure électronique. La DFT, introduite à la fin des années 60 et récompensée par le prix Nobel de Walter Kohn en 1998, a rencontré un franc succès avec le développement des machines de calcul parallèle. Elle permet de simuler les matériaux à l'échelle de plusieurs centaines d'atomes en quelques dizaines d'heures sur des machines possédant plusieurs milliers de cœurs. Elle repose sur l'expression en fonctionnelle de la densité d'une partie du terme d'interaction inter-électronique. Cette fonctionnelle exacte est cependant inconnue et seules des approximations de cette fonctionnelle sont disponibles. Elles peuvent être locales (comme la local density approximation, LDA) ou faire apparaître les gradients (comme l'implémentation proposée par Perdew, Burke et Ernzerhof, PBE). Plus récemment, certaines formes dites « hybrides » font apparaître une partie d'échange calculé exactement. Toutefois ce dernier terme est extrêmement coûteux et peut aisément multiplier par 10 le temps de calcul, limitant ainsi la taille et/ou le nombre des systèmes simulés.

Quelques articles récents font état de calculs avec ces fonctionnelles hybrides pour la lacune de cadmium dans CdTe [3, 4], neutre ou chargée négativement. Les auteurs montrent que cette fonctionnelle favorise une hybridation des premiers voisins tellure de la lacune malgré une importante déformation élastique engendrée par cette reconstruction – il s'agit d'un effet Jahn-Teller. Au contraire de fonctionnelles plus classiques (comme la LDA) pour laquelle la lacune conserve une symétrie tétraédrique. Ces résultats posent la question de la prédictibilité des fonctionnelles DFT quant aux structures des défauts. Est-ce qu'une fonctionnelle reproduisant correctement une forte interaction électronique donnera des résultats plus pertinents qu'une fonctionnelle reproduisant parfaitement la nature élastique du matériau ?

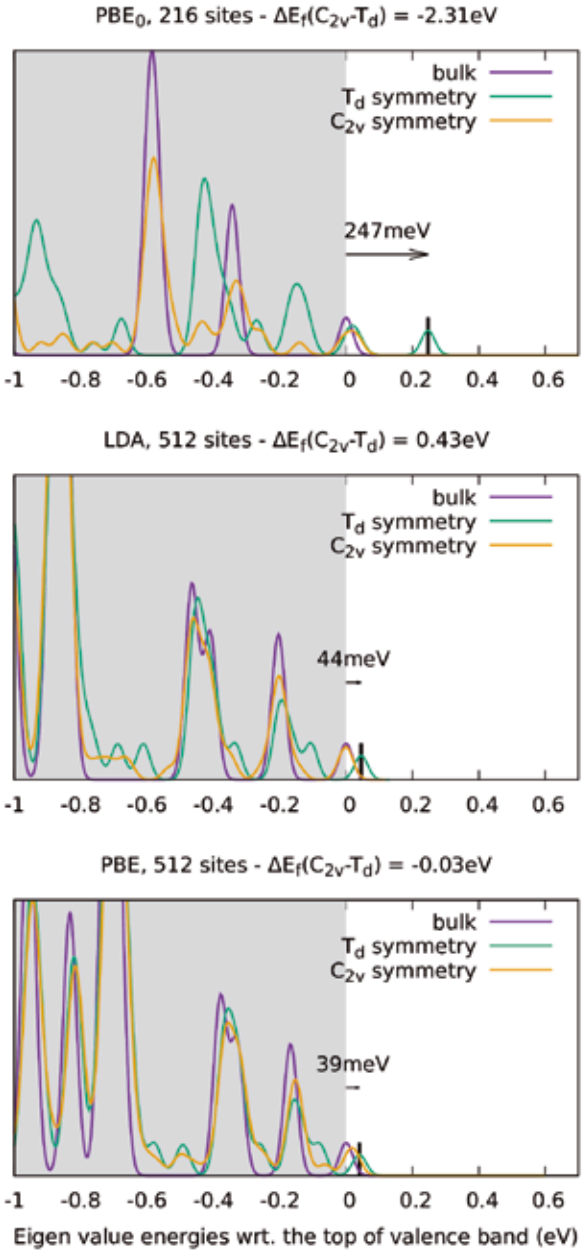
La capacité du code BigDFT à traiter sur GPU le calcul d'échange exact, associée à la disponibilité de nombreux accélérateurs graphiques sur la partition GPU du calculateur Jean Zay, nous offre la possibilité d'étudier cette question. Le terme d'échange exact peut être exprimé comme un calcul de Poisson à partir de densités partielles, obtenues comme le produit des fonctions d'onde entre elles. La possibilité d'effectuer ces N^2 calculs (N étant le nombre de fonctions d'ondes) sur GPU est un atout considérable et permet d'effectuer des calculs DFT en hybride pour un coût à peine supérieur à trois ou quatre fois le coût d'un même système en fonctionnelles standards sur CPU.

Un calcul DFT permet d'obtenir l'énergie totale du système simulé, les forces s'exerçant sur les atomes ainsi que les états électroniques. Durant les simulations, les atomes sont déplacés de façon à minimiser les forces s'exerçant sur eux. La lacune de cadmium peut alors adopter plusieurs géométries méta-stables. L'une est caractérisée par une symétrie tétraédrique (Td), où les quatre voisins tellure de la lacune se rapprochent pour hybrider les électrons présents dans les liaisons pendantes laissées par le manque d'un atome de cadmium. La déformation élastique induite par cette reconstruction est limitée. L'autre géométrie obtenue par relaxation des positions atomiques prend une symétrie C2v, avec un fort rapprochement de seulement deux des voisins tellure, caractéristique d'une hybridation plus importante. Elle est associée à une forte déformation élastique. Pour des fonctionnelles standards (LDA et PBE), les boîtes de simulation contiennent 512 atomes, ce qui correspond à un cube de matière de 2,6 nm de côté. Pour la fonctionnelle hybride utilisée (PBE0), la super-cellule est limitée à 216 atomes (cube de 2,0 nm de côté).

La densité d'état proche du haut de bande de valence est représentée sur la **figure 1**, pour chacune des fonctionnelles utilisées. La densité d'état du matériau massif est représentée par les courbes en violet.

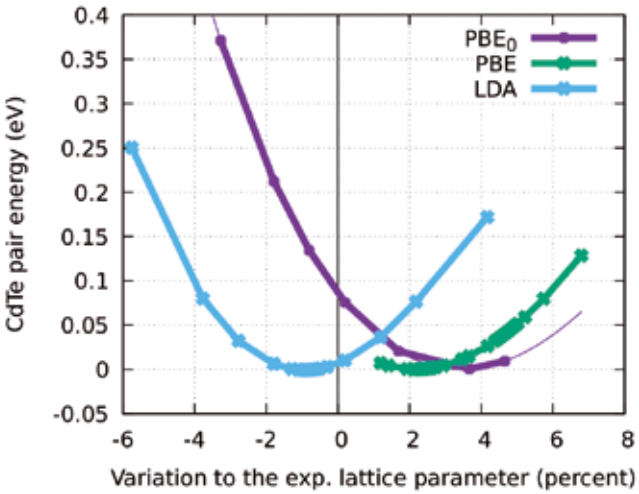
Un calcul DFT fournit les énergies des différents états électroniques. Ici, une dispersion gaussienne de sigma égale à 0,02 eV a été appliquée. Compte tenu de cette dispersion, le dernier état de la bande de valence est un état triplement dégénéré peuplé de six électrons. On constate que la géométrie C_{2v} (courbe jaune) ne rajoute pas d'états dans le gap (ou bien à la marge), au contraire de la géométrie T_d (courbe verte) qui rajoute un état situé entre quelques dizaines de meV (fonctionnelles classiques) et quelques centaines de meV (fonctionnelle hybride). Cet état est trois fois dégénéré mais seulement partiellement occupé par quatre électrons. C'est donc un état doublement accepteur. Comparées aux données expérimentales disponibles qui situent une transition -/2- en deçà de 470 meV [5], les trois fonctionnelles s'accordent pour indiquer que cette transition correspond à la géométrie T_d. Mais seule la fonctionnelle hybride reproduit qualitativement la position de ce niveau. D'autre part, la différence d'énergie totale entre les deux géométries est indiquée sur la **figure 1** dans les titres de chacun des trois graphes. Cette différence indique le rapport de population entre les deux géométries dans un échantillon à une température donnée. Lorsqu'elle est négative, la géométrie C_{2v} est dominante. Avec une différence de -2,3 eV, la fonctionnelle hybride exclue la présence de la géométrie T_d à température ambiante. Le ratio s'inverse en faveur de la géométrie Td en LDA. Cette différence de stabilité s'explique en partie par la façon dont les deux fonctionnelles reproduisent l'absorption de la déformation élastique (forte en géométrie C_{2v} et plus faible en géométrie T_d), avec un avantage certain pour la LDA qui reproduit plus fidèlement le paramètre de maille et le module de compression du CdTe, comme indiqué sur la **figure 2**.

Figure 1 : *Densité d'état proche du haut de la bande de valence pour le CdTe, la lacune de cadmium, dans la géométrie Td et C2v, pour trois fonctionnelles DFT.*



Les calculs réalisés sur Jean Zay grâce aux accélérateurs graphiques, ont permis de comparer dans des super-cellules suffisamment grandes pour absorber les déformations élastiques, les états électroniques et la stabilité de différentes configurations de la lacune de cadmium dans CdTe. Cette lacune est un élément important des dispositifs industriels à base de CdTe, qu'elle soit utilisée pour le dopage, ou au contraire passivée afin de diminuer son effet recombinant [6].
Pour compléter les informations présentées ici, nous préparons actuellement un article sur la façon dont on peut modifier la position des états dans le gap par recombinaison avec un atome de chalcogénure dans le cadre des applications photovoltaïques du CdTe.

Figure 2 : *Écart au paramètre de maille expérimental pour différentes fonctionnelles.*

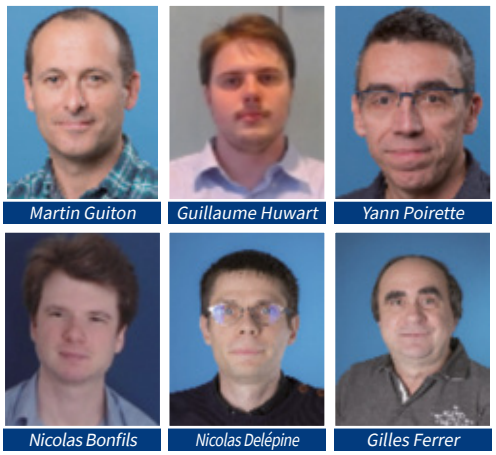


Références

[1] A. Carvalho et al. Phys. Rev. B 81, 075215 (2010)
[2] Y. Wu et al. Comp. Mat. Sc. 98, 18 (2015)
[3] A. Lindström et al. J. Phys. D: Appl. Phys. 49, 035101 (2016)
[4] A. Shepidchenko et al. Scientific reports 5, 1450 (2015)
[5] P. Emanuelsson et al. Phys. Rev. B 47, 15578 (1993)
[6] T. Fiducia et al. Nat Energy 4, 504 (2019)

Estimation de la durée de vie d'une éolienne flottante par calcul aéro-servo-hydro-élastique massif.

N. Delépine, N. Bonfils, M. Guiton, G. Huwart, G. Ferrer Y. Poirette.
IFP Energies nouvelles



Contexte

La France vise un taux d'énergies renouvelables de 32% dans son mix énergétique à l'horizon 2030.
Afin d'atteindre cet objectif, la part de la production d'énergie éolienne doit continuer à croître. L'éolien terrestre joue déjà un rôle important dans la transition énergétique. L'éolien en mer est aussi une technologie mature dans le cas de la technologie dite « posée » où le mât de l'éolienne est fixé sur une fondation ancrée dans le sol du fond marin. En France, cette technologie est en cours de déploiement sur plusieurs sites en Manche et en Atlantique avec une mise en exploitation prévue à partir de 2022. Cependant, la profondeur d'eau limite le développement de la filière éolien en mer posé à des sites peu profonds et proches des côtes. Pour permettre d'installer des éoliennes sur des sites de profondeur d'eau supérieure à 100 mètres et plus loin au large où les conditions de vents sont plus favorables, de nouveaux concepts proposent de fixer l'éolienne sur un support flottant maintenu grâce à un système d'ancrage (figure 1). Différents projets pilotes et pré-commerciaux ont été récemment développés, notamment en Méditerranée avec une technologie présentée dans Caillé et al. (2017) et modélisée dans ce Grand Challenge. Ces nouvelles structures doivent garantir le bon fonctionnement de l'éolienne et prouver leurs résistance et tenue en fatigue sous différents cas de charge imposés par les normes en vigueur.

un tel plan d'expérience massif prenant en compte autant de paramètres environnementaux n'avait jusqu'ici jamais été réalisé.

Motivations de l'étude

Cette étude s'inscrit dans une démarche d'optimisation de plans d'expériences numériques de calculs de fatigue adaptés au design de structures pour l'éolien en mer. Cette optimisation est nécessaire pour deux raisons :

1- Pour chaque condition environnementale considérée, une heure de temps physique requiert 15 heures de temps CPU sur la machine Jean Zay pour ces simulations multi-physiques.

2- Pour un calcul de fatigue, il est nécessaire de prendre en compte un très grand nombre de conditions environnementales représentatives de la vie d'une éolienne.

Description de l'étude

L'étude a utilisé un plan d'expérience numérique qui prend en compte de manière complète l'ensemble des combinaisons de paramètres environnementaux déterminants de mer et de vent obtenus à partir de données réelles. Les paramètres de vent se déclinent en fonction de sa vitesse moyenne, son intensité de turbulence et sa direction. Les paramètres de mer sont la hauteur significative, la période du pic du spectre et la direction de la houle. Les variables environnementales de mer et de vent sont fortement corrélées entre elles, par exemple entre la hauteur et la période de houle, ou le vent moyen et la hauteur significative.
A partir des données d'une bouée au large de la côte Est américaine, un tirage de type échantillonnage par hypercube latin est réalisé respectant la corrélation entre les variables de vent et de mer : 260 000 sextuplets de points sont ainsi obtenus afin de couvrir la totalité du domaine. Chaque sextuplet de paramètres environnementaux permet de générer une réalisation temporelle des conditions environnementales (hauteur de vague, vitesse instantanée de vent) et la réponse mécanique de l'éolienne flottante grâce à un calcul couplé aéro-servo-hydro-élastique. Le code de calcul utilisé, DeepLines Wind, est un logiciel éléments finis qui intègre à la fois les effets combinés des efforts aérodynamiques sur les pales, d'un contrôle du couple électromagnétique de la génératrice et de l'orientation des pales, des efforts hydrodynamiques sur la plateforme flottante et les lignes d'ancrage. Ainsi, environ 30 ans de la vie d'une éolienne flottante sont simulés.

Résultats

Grâce à la capacité du supercalculateur Jean Zay, les calculs ont pu être réalisés en quelques mois, au cours desquelles, plusieurs pics de fréquence d'utilisation ont été relevés : jusqu'à 20.000 cœurs ont pu être simultanément et ponctuellement occupés, soit environ le tiers de la capacité de la machine.

A partir de ces résultats, des surfaces de réponse du dommage en fatigue sont construites, grâce à une prise en compte exhaustive de l'ensemble des domaines de définition de chaque paramètre d'entrée.

Ce Grand Challenge a permis à IFP Energies nouvelles d'estimer la durée de vie en fatigue du support d'une éolienne en mer flottante, en tenant compte d'un grand nombre de conditions de mer et de vent auxquelles elle sera soumise. A notre connaissance, un tel plan d'expérience massif prenant en compte autant de paramètres environnementaux n'avait jusqu'ici jamais été réalisé.

Perspectives

La validation des approches de design par plans d'expérience adaptatifs permettra de justifier les évolutions des méthodes d'ingénierie et de certification. A terme, des analyses de fiabilité et d'optimisation sous contrainte pourront être proposées pour ce type de problèmes d'ingénieries complexes.

Ce type d'optimisation participe également à rendre compétitif économiquement ces technologies. En effet, la cible de coût de l'éolien flottant est ambitieuse puisqu'elle est d'environ €80-100 / MWh pour les premières fermes commerciales d'ici 2023-2025 (WindEurope, 2018).

Références

Caillé, F., Bozonnet P., Perdrizet T., Poirette Y. and Melis C. *Model Test and Simulation Comparison for an Inclined-Leg TLP Dedicated to Floating Wind. Proceedings of the ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering.* Volume 10 : Ocean Renewable Energy. Trondheim, Norway. June 25–30, 2017. V010T09A070. ASME. https://doi.org/10.1115/OMAE2017-61652.

Huchet Q., Mattrand C., Beaurepaire P., Relun N. et Gayton N. Approximation d'intégrales coûteuses par utilisation de métamodèles : Application à la certification en fatigue des structures éoliennes. 23ème Congrès Français de Mécanique, juillet 2017.

Wind Europe. "Floating offshore wind energy: a policy blueprint for Europe", Position paper, Octobre 2018.

Remerciements

Nous remercions le personnel de l'IDRIS pour leur disponibilité et leur aide : Rémy Dubois, Pascal Voury et Alberto Garcia. Nous remercions également la société SBM-Offshore pour le design de la structure flottante.

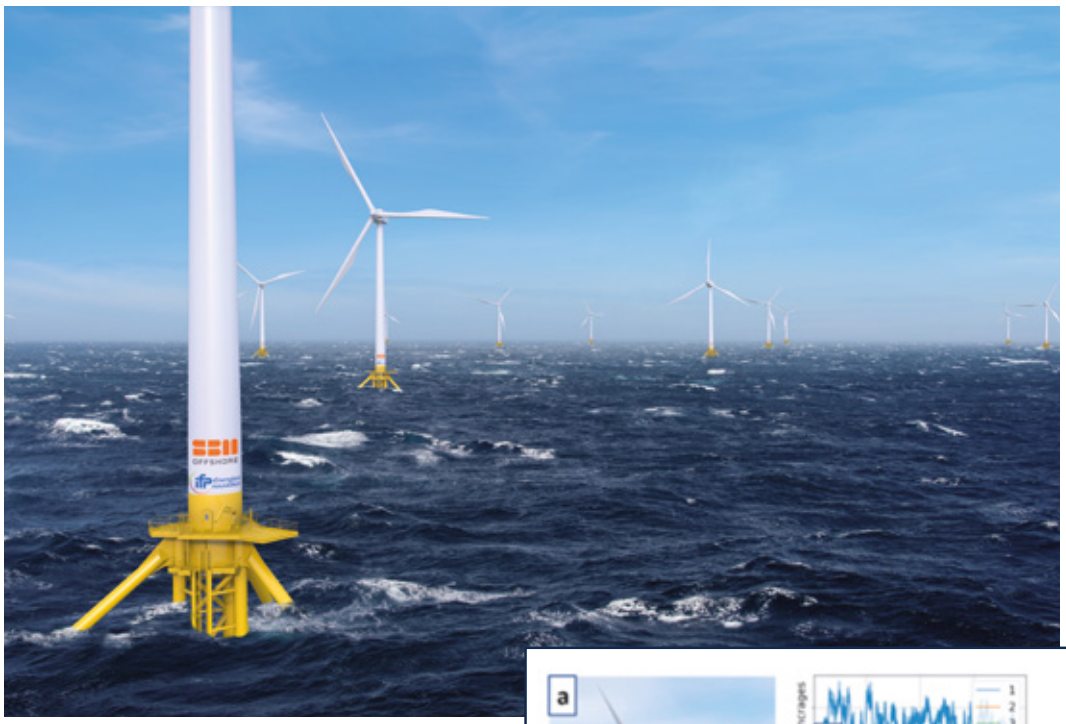


Figure 1 : *illustration d'une éolienne flottante de type « Tension-Leg ».*

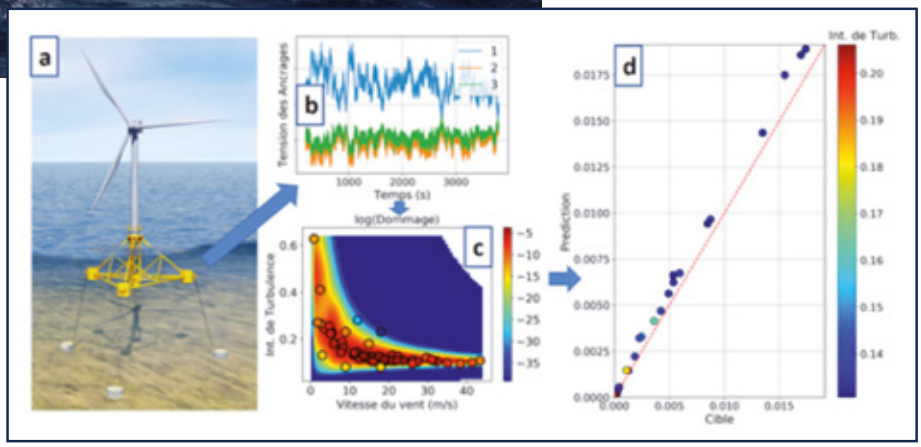


Figure 2 : *Illustration de l'approche pour estimer la fatigue en tête des lignes d'ancrage du support flottant (a) : les tensions obtenues par calcul (b) sont utilisées pour estimer des surfaces de réponse en dommage (c), ici représenté en fonction de la vitesse de vent et de sa turbulence. La qualité de ces surfaces de réponse est vérifiée par comparaison à une cible de référence (d)*

Apprentissage à très grande échelle de représentations vectorielles de documents. Application au domaine biomédical

Porteur du projet :
François Role - francois.role@u-paris.fr
Université de Paris

Contexte et objectifs du projet

La communauté biomédicale a développé un grand nombre de bases de données, d'ontologies et de bibliothèques numériques. A titre d'exemple, en janvier 2020, la base Pubmed contenait près de 30 millions d'articles scientifiques. Ces articles contiennent des gisements de connaissances qui, vue leur ampleur, ne peuvent être vraiment exploitées que par des techniques de *text mining* permettant d'analyser automatiquement de gros volumes de données textuelles et d'en extraire des connaissances utiles.
La maîtrise de techniques efficaces d'extraction automatique à très grande échelle de connaissances médicales représente **des enjeux scientifiques et sociétaux importants**.
L'analyse automatique des textes permet en effet à un chercheur de confirmer des résultats obtenus par des expériences biomédicales ou d'obtenir des suggestions sur l'existence d'associations inconnues entre certaines entités biomédicales.
Le fait de travailler sur de gros volumes de données a un effet de décloisonnement suscitant des mises en relation inattendues entre des gènes, des maladies, des facteurs, ce qui favorise le développement de thérapies innovantes pour mieux soigner les patients. Cette préoccupation prend aujourd'hui encore plus de sens dans le contexte de la pandémie dûe au Covid-19.

La maîtrise de techniques efficaces d'extraction automatique à très grande échelle de connaissances médicales représente des enjeux scientifiques et sociétaux importants

Avec l'essor de l'apprentissage par réseaux de neurones, l'analyse automatique de gros volumes de textes peut rapidement nécessiter des calculs ne pouvant s'effectuer qu'à l'aide de machines puissantes dotées de CPU puissants et de GPU (processeurs graphiques ayant une structure hautement parallèle) performants. Certains chercheurs de l'équipe *Machine Learning for Data Science* **[2]** (MLDS) de l'Université de Paris **[3]** travaillent régulièrement sur des données textuelles médicales depuis plusieurs années mais jusqu'à présent sur des volumes de données compatibles avec les moyens dont ils disposent. Au début de l'été 2019, afin de pouvoir appliquer leurs idées à plus grande échelle, ils ont donc décidé de répondre au « Grand Challenge Jean-Zay » lancé par le CNRS pour tester son nouveau supercalculateur.

Résultats obtenus

Pour pouvoir analyser des textes, une des étapes incontournables est de les représenter sous une forme vectorielle, notamment sous forme de vecteurs denses en basse dimension (des *embeddings*). Avoir des représentations vectorielles de mots ou de groupes de mots (phrases, paragraphes) sur l'ensemble des résumés d'articles de la base Pubmed était l'ambition initiale du projet.

La première étape du projet a eu pour but de tester les capacités de différentes bibliothèques logicielles pour le traitement de très gros volumes de textes. Grâce au package multiprocessing de la bibliothèque scikit-learn **[4]** qui fournit des primitives de parallélisation très simples mais très performantes nous avons pu écrire un programme permettant de découper rapidement et efficacement en phrases les 30 millions de résumés d'articles disponibles dans la base Pubmed.
L'ensemble des phrases a ensuite été analysé avec le package Python scispacy **[5]** qui propose des modèles pré-entraînés sur des textes biomédicaux scientifiques et cliniques. L'objectif était d'extraire automatiquement des expressions significatives dans chaque phrase.
Par exemple dans la phrase :
Surgical practice has been significantly impacted by the COVID-19 pandemic.
Le programme avait pour but d'identifier des expressions comme *surgical practice* et COVID-19 pandemic.

La seconde étape du projet a consisté à calculer des *embeddings* de phrases sur un sous-ensemble des textes lié à un domaine biomédical spécialisé. Compte tenu des circonstances, nous avons choisi de travailler sur des articles relatifs aux maladies provoquées par les différents coronavirus.
Nous avons extrait 15 000 articles recensés par Pubmed, début mars 2020, comme traitant des maladies coronavirus en général notamment les SARS-CoV et MERS-CoV des épidémies de 2002 et 2012 et 1100 articles ciblant spécifiquement le Covid-19 **[6]**. Les *embeddings* d'environ 150 000 phrases ont été calculés en utilisant les modèles BioBERT **[7]**. Nous avons ensuite entraîné et utilisé un modèle de reconnaissance d'entités nommées s'appuyant également sur BioBERT et capable de retrouver dans les phrases des mentions d'entités biomédicales comme des maladies

(co-morbidité associée au Covid-19) ou de composés chimiques et de leur associer des étiquettes adéquates *DISEASE ou CHEMICAL*.

Conclusion et perspectives

A titre personnel, travailler sur Jean-Zay nous a permis de nous familiariser avec une plate-forme de type supercalculateur. Nous avons ainsi pu découvrir les principes de base de l'écriture de scripts Slurm pour lancer des traitements. Notre relative inexpérience en ce domaine ne nous a cependant pas permis surtout dans la phase 2 d'exploiter à fond les puissantes ressources GPU qui étaient à notre disposition. Le Grand Challenge nous a donné envie d'en savoir plus dans ce domaine.
En ce qui concerne les résultats, ils pourraient éventuellement être mis à la disposition de la communauté scientifique dans des conditions qui seraient à définir pas nos institutions de tutelle.

Nous voulons pour conclure remercier l'équipe support de l'IDRIS qui a toujours été très réactive et a répondu rapidement à nos questions et à nos demandes d'installation de modules logiciels supplémentaires dès que nous en avons eu besoin.

Références

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] François Role, Mohamed Nadif, Séverine Affeldt, Lazhar Labiod.
[3] https://u-paris.fr/
[4] https://scikit-learn.org/stable/
[5] https://allenai.github.io/scispacy/
[6] Il s'agissait du nombre d'articles relatifs à ces sujets en mars 2020. Depuis, le nombre d'articles consacré au Covid-19 a énormément augmenté !
[7] https://github.com/dmis-lab/biobert

TrainedBot: entrainement d'agents robotiques à partir de simulations

Edward Beeching [1], Jilles Dibangoye [1], Olivier Simonin [1], Christian Wolf [2]
1 INRIA Chroma team, CITI Lab., INSA Lyon, France
2 LIRIS, INSA-Lyon, France


Edward Beeching


Jilles Dibangoye


Olivier Simonin


Christian Wolf

Contexte

Les dernières années ont été marquées par l'essor du *Machine Learning* (ML), qui a permis des gains en performances significatifs dans plusieurs domaines d'application. Outre les progrès méthodologiques indéniables, ces gains sont souvent attribués à de grandes quantités de données d'entraînement et à la puissance de calcul, qui ont conduit à des avancées dans la reconnaissance de la parole, la vision par ordinateur et le traitement automatique de la langue. Dans ce projet, nous proposons d'étendre ces avancées à l'apprentissage large-échelle d'agents intelligents (robots compagnon, robots de service) évoluant dans des environnements complexes et capables de manifester un raisonnement de haut niveau. Nous ciblons des applications telles que des assistants robotiques dans les hôpitaux, des robots domestiques et des assistants de navigation pour les aveugles. En effet, les futurs robots seront entraînés, plutôt que programmés, pour presque toutes leurs tâches et sous-problèmes : perception, planification et navigation, comportement et contrôle. Cependant, obtenir des gains en apprenant des quantités massives de données ne sera pas aussi facile que dans la vision, la parole, le traitement du langage, où l'apprentissage automatique a été un succès récent. Contrairement à ces problèmes, qui sont souvent résolus par l'apprentissage supervisé, enseigner aux robots et aux agents à agir de manière autonome nécessite d'apprendre à partir d'interactions avec un environnement. Dans ce contexte, nous avons proposé une méthode permettant l'entraînement de comportements robotiques à partir d'un ensemble vaste de simulateurs différents et variés, l'objectif étant d'aboutir à une généralisation du comportement vers des situations non vues et l'apprentissage continu.

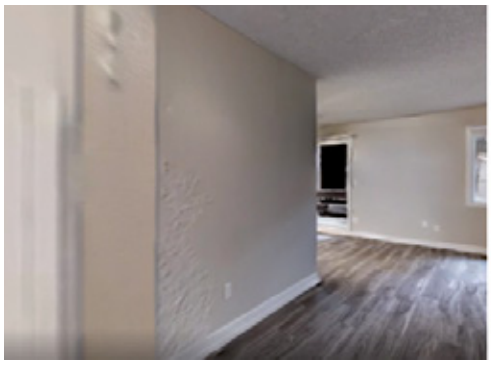
Nous avons entrainé les réseaux de neurones en les laissant interagir avec un vaste ensemble d'environnements simulés.

Nous avons développé de nouveaux modèles neuronaux permettant d'entraîner des agents virtuels à naviguer dans des environnements photo-réalistes, l'objectif étant de déployer le modèle (le comportement appris) sur un vrai robot physique. **La figure 1** (gauche) montre une image observée par un agent en cours de navigation.

Nous entraînons des réseaux de neurones interagissant avec un vaste ensemble de tels environnements simulés. Les ressources de calcul GENCI ont été utilisées pour l'exécution, en parallèle, des simulations photo réalistes et des entraînements (par "Deep Reinforcement Learning"), les deux étant effectués sur cartes graphiques (GPU), avec un entraînement parallèle sur un maximum de 4 GPU. Ces travaux ont été acceptés pour publication dans les actes de la conférence international European Conference on Computer Vision (ECCV), 2020 [6].

Apprendre à planifier avec des cartes topologiques incertaines

Dans ces recherches, nous avons traité le problème de la planification dans un environnement incertain : le robot connaît un modèle de l'environnement sous forme de graphe, c.à.d. des sommets connectés par des arêtes indiquant des liens entre les différents emplacements sur la carte. Ces connexions étant issues d'observations, elles sont incertaines, c.à.d. sujet aux erreurs. Cette représentation incertaine est échantillonnée à partir d'une phase exploratoire dans un simulateur 3D photoréaliste. Nous appliquons cette approche dans le simulateur "Habitat" créé par Facebook AI Research [4], illustré dans la **Figure 1** (droite).
En utilisant un nouveau modèle basé sur un réseau de neurones de type graphes ("graph neural network"), nous entraînons un planificateur neuronal robuste qui surpasse les solutions classiques de l'IA basé sur le calcul du plus court chemin.
Nous considérons qu'un agent puisse explorer un nouvel environnement avant de passer à la navigation, c.à.d. pendant une phase exploratoire, où l'agent construit un graphe des emplacements visités dans l'environnement. En chaque emplacement, nous calculons également un ensemble de caractéristiques visuelles à partir d'une observation monoculaire, c.à.d. une image, et nous générons un graphe avec des estimations probabilistes de la connectivité des nœuds, prédites par un réseau neuronal.



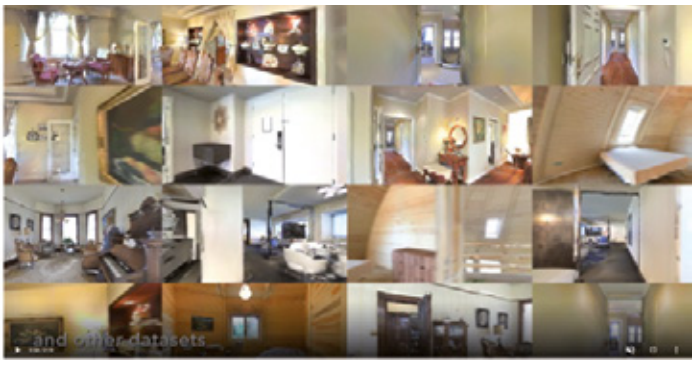


Figure 1 : *Gauche : exemple d'une observation vue par un agent (robot) virtuel navigant en simulation; Droite : un aperçu de la collection d'appartement simulés disponibles dans le simulateur habitat.ai proposé Facebook AI research.*

Nous visons ensuite à effectuer la planification dans ce graphe incertain, avec l'objectif de trouver le chemin le plus court d'un noeud emplacement à un autre afin d'aboutir à un emplacement cible donné. Nous combinons ensuite ce planificateur avec un contrôleur de bas niveau capable de naviguer entres les cibles intermédiaires générés par le planificateur neuronal.
Nous traitons la planification de chemins comme un problème de classification. Nous supposons qu'au cours de l'entraînement, la vérité terrain sur la connectivité des nœuds est disponible et qu'une planification de chemin classique peut être effectuée. Pour chaque sommet dans un graphe à N sommets, nous effectuons une classification à plusieures voies du sommet suivant sur le chemin optimal. Notre contribution principale consiste en une amélioration des graph networks leur permettant de représenter les algorithmes classiques de planification (de type Bellman-Ford), et de les améliorer.

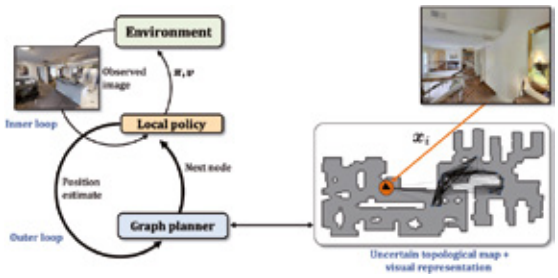


Figure 2 : *Vue d'ensemble du planificateur neuronal graphique couplé à un contrôleur de bas niveau. Lors de la navigation, le planificateur a accès à une carte incertaine sous forme de graphe, et sur des descripteurs visuels associés à chaque sommet du graphe.*

Expériences et résultats

L'entraînement a été mené sur un ensemble de 72 000 graphes donnant lieu à 74 millions d'instances du problème. Nous avons évalué notre contribution sur des environnements non visités lors de l'entraînement [2] afin d'évaluer la capacité du modèle à généraliser.
Une analyse quantitative a été menée avec deux paramètres standard de la littérature [3]: le taux de réussite, c.a.d. les points de départ et d'arrivée qui ont été franchis avec succès et le SPL, c.à.d. le rapport moyen entre la longueur du trajet et la longueur optimale. Nous nous comparons à deux variantes de l'algorithme de Dijkstra (un algorithme classique de planification); nous nous comparons également avec une approche "end-to-end", obtenue avec un apprentissage par renforcement sans utilisation de cartes. Les résultats sont présentés dans le **tableau 1.**

Méthode: planificateur + stratégie locale	Succès	SPL
Agent aléatoire	**0.15**	**0.11**
Agent neuronal classique (sans carte)	**0.55**	**0.53**
Algorithme Symbolique "Dijkstra"	**0.68**	**0.59**
Planificateur neuronal (échantillonnage)	**0.97**	**0.88**

Tableau 1 : *Succès de navigation de l'approche proposée comparés à des approches classiques.*

La **Figure 3** illustre la navigation d'un agent à l'aide de la méthode hiérarchique dans un environnement de test. Gauche: les images observées en couleur et en profondeur. Droite: la carte topologique (le graphe) de l'environnement, avec les sommets du graphe (cyan), le départ de la navigation (vert), le sommet cible (rouge), la position actuelle de l'agent (noir), le voisin le plus proche de l'agent (violet), la prochaine cible intermédiaire (bleu) fourni par le planificateur de haut niveau, et le chemin planifié.

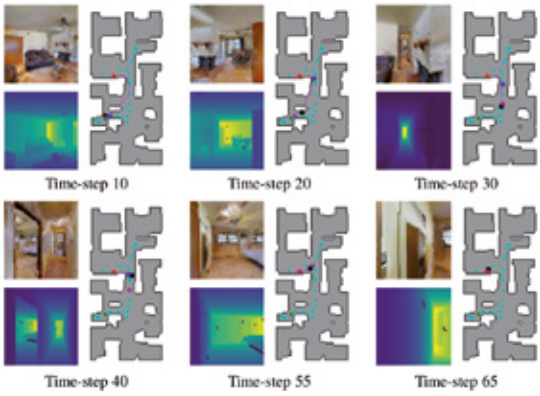


Figure 3 : *Gauche : les images observées en couleur et en profondeur. Droite : la carte topologique (le graphe) de l'environnement, avec les sommets du graphe (cyan), le départ de la navigation (vert), le sommet cible (rouge), la position actuelle de l'agent (noir).*

Conclusion

Dans ces travaux, nous avons proposé un nouveau modèle neuronal permettant d'entraîner des agents virtuels à naviguer dans des environnements photo-réalistes, en introduisant un planificateur hiérarchique basé sur une carte topologique neuronale (un graphe). Nous avons entraîné les réseaux de neurones en les laissant interagir avec un vaste ensemble d'environnements simulés. Les ressources de calcul GENCI ont été utilisées pour l'exécution, en parallèle, des simulations photo réalistes et des entraînements (par "Deep Reinforcement Learning"), les deux étant effectués simultanément sur les cartes graphiques (GPU) d'un sommet de calcul du cluster Jean-Zay. Ces travaux ont été acceptés pour publication dans les actes de la conférence international European Conference on Computer Vision (ECCV), 2020 [6].

[1] Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
[2] Xia, F., R. Zamir, A., He, Z.Y., Sax, A., Malik, J., Savarese, S.: Gibson env: realworld perception for embodied agents. In: Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on. IEEE (2018)
[3] Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta,S.,Koltun,V.,Kosecka,J.,Malik,J.,Mottaghi,R.,Savva, M., Zamir, A.R.: On evaluation of embo died navigation agents (2018)
[4] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, and structured egocentric memory for Deep RL. To appear in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2020. Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
[5] Edward Beeching, Jilles Dibangoye, Olivier Simonin and Christian Wolf. EgoMap: Projective mapping and structured egocentric memory for Deep RL. To appear in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2020.
[6] Edward Beeching, Jilles Dibangoye, Olivier Simonin and Christian Wolf. Learning to plan with uncertain topological maps. To appear in European Conference on Computer Vision (ECCV), 2020.

FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français rendus disponibles grâce au supercalculateur Jean Zay

Léquipe (en haut à gauche à en bas à droite) : *Benoît Crabbé (en noir &blanc); Laurent Besacier, Didier Schwab, Benjamin Lecouteux, LE Thi Phuong Hang, Alexandre Allauzen, Loïc Vial, Jibril Frej, Vincent Segonne*

Didier Schwab, Univ. Grenoble Alpes, CNRS, LIG didier.schwab@univ-grenoble-alpes.fr

Introduction

Nous présentons ici le projet FlauBERT qui n'aurait jamais vu le jour sans le supercalculateur Jean Zay (programme « Grand Challenge Jean Zay » – projet 100967) mis à disposition à l'IDRIS (Institut du développement et des ressources en informatique scientifique).

Ces recherches sont le fruit de la collaboration entre :

— l'Université Grenoble Alpes (Laboratoire d'informatique de Grenoble, CNRS) – Hang Le, Loïc Vial, Jibril Frej, Maximin Coavoux, Benjamin Lecouteux, Laurent Besacier, Didier Schwab ;

— l'Université Paris Diderot – Vincent Segonne, Benoît Crabbé

— E.S.P.C.I, CNRS LAMSADE, PSL Research University – Alexandre Allauzen

Le projet FlauBERT n'aurait jamais vu le jour sans le supercalculateur Jean Zay ...

En 2018, l'introduction de représentations linguistiques profondes contextuelles, obtenues à partir de textes bruts, a conduit à un changement de paradigme pour plusieurs tâches du TALN. Alors que les approches fondées sur des représentations continues comme word2vec (Mikolov et al., 2013) apprennent un vecteur unique pour chaque mot, les modèles introduits récemment produisent des représentations contextuelles qui dépendent de la séquence de mots d'entrée complète. Initialement fondées sur des réseaux neuronaux récurrents, ces approches ont peu à peu intégré des modèles Transformer (Vaswani et al., 2017) comme c'est le cas pour BERT (Devlin et al., 2019), ou RoBERTa Liu et al. (2019). L'utilisation de ces modèles pre-entraînés a permis des avancées de l'état-de-l'art pour de nombreuses tâches du TALN. Cependant, ceci a surtout été montré pour l'anglais, même si des variantes multilingues sont également disponibles, prenant en compte plus d'une centaine de langues dans un seul modèle : mBERT (Devlin et al., 2019) par exemple. L'objectif de ce projet était de construire et partager FlauBERT (*French Language Understanding via Bidirectional Encoder Representations from Transformers*), un modèle BERT pour le français. FlauBERT surpasse le modèle multilingue mBERT dans plusieurs tâches. Nous avons également proposé à la même occasion un référentiel d'évaluation nommé FLUE (French Language Understanding Evaluation) similaire au benchmark GLUE (Wang et al., 2018) pour l'anglais. FlauBERT et FLUE sont disponibles en ligne. [1]

Apprentissage du modèle FlauBERT

Données d'apprentissage et pré-traitements nous agrégeons 24 sous-corpus de types divers (wikipedia, livres, *Common Crawl*, . . .). Nos trois sources principales sont (1) les textes monolingues des campagnes d'évaluation WMT19 (Li et al., 2019, 4 sous-corpus), (2) les textes en français de la collection OPUS (Tiedemann, 2012, 8 sous-corpus), (3) le projet Wikimedia (8 sous-corpus). La taille totale (non compressée) des textes ainsi agrégés est de 270 GB.

Après un prétraitement consistant en différents filtrages (enlever les phrases très courtes, les séquences de numéros ou d'adresses électroniques, etc.), une normalisation de l'encodage des caractères, et une tokenisation, nous obtenons un corpus de 71 GB. Notre code pour télécharger et pré-traiter les données est publiquement disponible.

Objectif de l'entraînement et optimisation

Le pré-entraînement de FlauBERT consiste à réaliser un modèle de langue masqué (MLM) qui apprend à prédire des jetons masqués de façon aléatoire. Pour optimiser cette fonction objectif, nous avons suivi Liu et al. (2019) et utilisé l'optimiseur Adam (Kingma & Ba, 2014) avec les paramètres suivants :

— $\text{FlauBERT}_{\text{BASE}}$: étapes de mise en route (ou *warm up*) de 24 k, taux d'apprentissage maximal de 6e−4, $\beta 1 = 0,9$, $\beta 2 = 0,98$, $\varrho = 1e{-}6$ et perte de poids de 0,01.

— $\text{FlauBERT}_{\text{LARGE}}$: étapes de mise en route de 30 k, taux d'apprentissage maximal de 3e−4, $\beta 1 = 0,9$, $\beta 2 = 0,98$, $\varrho = 1e{-}6$ et perte de poids de 0,01.

Modèles et configuration d'apprentissage

Nous utilisons la même architecture que BERT (Devlin et al., 2019). Un vocabulaire de 50 K unités sous-lexicales est construit en utilisant l'algorithme *Byte Pair Encoding* (Sennrich et al., 2016, BPE).

Nous entraînons deux principaux modèles (transformers bi-directionnels) : $\text{FlauBERT}_{\text{BASE}}$ (12 blocs de dimension cachée 768, 12 têtes pour l'attention) et $\text{FlauBERT}_{\text{LARGE}}$ (24 blocs de dimension cachée 1024, 12 têtes).

Le critère d'apprentissage est de type *masked language model* : il consiste à prédire des tokens d'une phrase ayant été préalablement et aléatoirement masqués. $\text{FlauBERT}_{\text{BASE}}$ est appris sur 32 GPU Nvidia V100 SXM2 32 GB en 410 h et $\text{FlauBERT}_{\text{LARGE}}$ est appris sur 128 de ces mêmes GPU en 390 h.

Tableau 1 : *Comparaison entre FlauBERT et d'autres modèles de langue pré-entraînés.*

	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	CamemBERT	$\text{FlauBERT}_{\text{BASE}}$ / $\text{FlauBERT}_{\text{LARGE}}$
Langue	Anglais	Anglais	Français	Français
Données d'apprentissage	13 GB	160 GB	138 GB†	71 GB‡
Objectifs de pré-entraînement	NSP et MLM	MLM	MLM	MLM
Nombre total de paramètres	110 M	125 M	110 M	138 M/ 373 M
Tokenisation	WordPiece 30K	BPE 50K	SentencePiece 32K	BPE 50K
Masque	Statique + sous-mots	Dynamique + sous-mots	Dynamique + mot entier	Dynamique + sous-mot

Tableau 2 : *Résultats finaux sur les tâches de FLUE.*

Section Mesure	Livres Acc.	DVD Acc.	Musique Acc.	Acc.	Acc.	F1	POS	UAS	LAS	Noms F1	Verbes F1
État de l'art ant.	91.25c	89.55c	93.40c	66.2[d]	80.1/85.2[e]	87.4[a]		89.19[b]	85.86[b]	-	43.0[h]
Sans pré-entr.	-	-	-			83.9	97.5	88.92	85.11	50.0	-
FastText	-	-	-			83.6	97.7	86.32	82.04	49.4	34.9
mBERT	86.15[c]	86.9[c]	86.65[c]	89.3[d]	76.9[f]	87.5	98.1	89.50	85.86	56.5	44.9
CamemBERT	93.40	92.70	94.15	89.8	81.2	88.4	98.2	91.37	88.13	56.1	51.1
FlauBERT$_{BASE}$	93.40	92.50	94.30	89.9	81.3	89.1	98.1	91.56	88.35	54.9/57.9[g]	47.4

Expérience et résultats sur FLUE

Le référentiel d'évaluation FLUE est composé de 7 tâches correspondant à différents niveaux d'analyse (syntaxique, sémantique) du traitement automatique du français.

— Classification de texte qui consiste ici à établir si un texte est positif, neutre ou négatif quant à son sujet ;

— Identification de paraphrases, tâche qui consiste à identifier si des paires de phrases sont sémantiquement équivalentes ou non ;

— Reconnaissance d'implications textuelles, tâche qui considère une prémisse (p) et une hypothèse (h) et consiste à déterminer si p implique, contredit ou ni n'implique ni ne contredit h ;

— Analyse syntaxique et étiquetage morphosyntaxique, deux tâches qui consistent à analyser en constituants ou en dépendances les textes ;

— Désambiguïsation lexicale des verbes et des noms, deux tâches dont l'objectif consiste à assigner un sens, parmi un inventaire donné, à des mots d'une phrase.

Nous comparons les performances de FlauBERT avec BERT multilingue (Devlin et al., 2019, mBERT) et CamemBERT (Martin et al., 2020) sur toutes les tâches.

Nous comparons également avec le meilleur modèle non contextuel pour chaque tâche. Nous utilisons les bibliothèques open-source XLM (Lample & Conneau, 2019) et Transformers (Wolf et al., 2019). Nous renvoyons à Le et al. (2020) pour une description détaillée des expériences et des résultats.

Conclusion

Nous avons présenté et partagé FlauBERT, un ensemble de modèles de langues pre-entraînés pour le français, accompagné de FLUE, un dispositif d'évaluation. FlauBERT obtient des résultats à l'état de l'art sur un certain nombre de tâches de TALN. Nous espérons que cette contribution et la mise à disposition du supercalculateur Jean Zay stimuleront les recherches sur le traitement automatique des langues et en particulier du français.[2]

[1] https://github.com/getalp/Flaubert

[2] FlauBERT est notamment disponible sur https://huggingface.co/models.

Références

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171–4186.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

LAMPLE G. & CONNEAU A. (2019). Cross-lingual language model pretraining. In Advances in neural information processing systems.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : Unsupervised language model pre-training for french. In Proceedings of The 12th Language Resources and Evaluation Conference, p. 2479–2490, Marseille, France : European Language Resources Association.

LI X., MICHEL P., ANASTASOPOULOS A., BELINKOV Y., DURRANI N., FIRAT O., KOEHN P., NEUBIG G., PINO J. & SAJJAD H. (2019). Findings of the first shared task on machine translation robustness. Fourth Conference on Machine Translation (WMT19), p. 91–102.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, p. 3111–3119, USA : Curran Associates Inc.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 1715–1725.

TIEDEMANN J.(2012). Parallel data, tools and interfaces in opus. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Éds., Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey : European Language Resources Association (ELRA).

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In Advances in neural information processing systems, p. 5998–6008.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. doi : 10.18653/v1/W18-5446.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M. & BREW J. (2019). Huggingface's transformers : State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Recherche sur le dialogue : génération de réponses et prédiction de la satisfaction client


Eric Clavier


Lina Rojas-Barahona

Eric Clavier et Lina Rojas-Barahona.
Orange-Labs. Equipe DATA-AI/NADIA. Lannion France.

Les systèmes de dialogues en langages naturels sont de plus en plus présents dans la relation entre Orange et ses clients. Que ce soit à travers un e-chat avec un téléconseiller ou via un bot (chatbot, voicebot), les clients d'Orange disposent de nouveaux moyens pour joindre et échanger avec les services support. Autour de ces outils se posent beaucoup de questions et de challenges : comment comprendre l'intention de l'utilisateur ? Comment trouver une réponse pertinente ? Comment exprimer cette réponse ? Comment s'assurer que l'utilisateur a bien compris la réponse ?
Comment mesurer la qualité de l'interaction ? Toutes ces problématiques sont au cœur d'un système de dialogue et forment sa base fonctionnelle **(cf. Figure 1).**

La multiplication des méthodes basées sur le deep learning, couplée à l'explosion des volumes de données a fortement complexifié les architectures nécessaires à leur mise en œuvre et décuplé les besoins en puissance de calcul

Les approches employées font appels aux différentes disciplines du *Machine Learning*. Néanmoins la multiplication des méthodes basées sur le *deep learning*, couplée à l'explosion des volumes de données a fortement complexifié les architectures nécessaires à leur mise en œuvre et décuplé les besoins en puissance de calcul.
C'est pourquoi, dans le cadre des grands challenges, Orange a pu s'appuyer sur les infrastructures offertes par le supercalculateur Jean Zay pour mener des travaux sur différentes problématiques : l'extraction et génération de la réponse dans un système de dialogue ouvert et la mesure de la satisfaction client dans les systèmes de dialogue dédiés à la résolution de tâches.
Les travaux relatifs à la génération de réponses consistait à évaluer le BLEU score [1] dans le cadre d'une tâche de questions/réponses conversationnelle. Nous avons comparé les résultats entre des modèles extractifs et modèles génératifs sur le corpus CoQA [2] **(cf. Figure 2)**.
La Figure 2 montre que le modèle extractif (*extr*) a un BLEU aux alentours de 18.6 car les réponses extractives sont plus longues que les vraies réponses. Le modèle (subextr), qui est un sous-ensemble des réponses extractives, a un BLEU de 80. Le modèle génératif (Gen) réussit à avoir un BLEU d'environ 50 à la fin de l'entrainement.
Concernant la satisfaction client, nos travaux ont consisté à évaluer si la prise en compte de la structure du dialogue améliorait la capacité des modèles neuronaux à prédire la satisfaction client.

Nous avons testé trois approches sur le corpus Let's Go [6] (cf. Table 1): des réseaux de neurones hiérarchiques récurrents bidirectionnels (BiGRUs), une seconde basée sur les Transformers pour représenter les tours de paroles et une autre basée seulement sur les Transformers pour faire la prédiction. Nous avons obtenu les meilleurs résultats avec les réseaux de neurones hiérarchiques récurrents bidirectionnels (BiGRUs) car la complexité et longueur (200 tours de parole) des dialogues fait que les modèles basés sur les Transformers sont moins performants et plus gourmands en termes de ressources GPU et mémoire. Nous avons pu tester différents modèles de Transformers : BERT [3], DistilBERT [4] ainsi que les Transformers extra large (Trans XL) [5].

Tableau 1 Evaluation des réseaux neuronaux pour prédire la satisfaction client. UAR (Unweighted Average Recall), est la moyenne arithmétique des rappels per classe.

Prédiction de la qualité de l'interaction			
Modèle	UAR*	Cohen's (κ)	Spearman's (ρ)
BiGRUs	**55%**	**74%**	**83%**
DistilBERT+GRU	35%	57%	42%
Trans-XL	43%	54%	68%

Ces travaux se sont poursuivis en utilisant ces modèles à l'intérieur de la fonction de récompense qui est au cœur de l'apprentissage par renforcement. Nous avons pu intégrer les BiGRUs qui gèrent un contexte dialogique de 100 tours de parole dans un framework de dialogue à base d'apprentissage par renforcement [7]. Au final nous avons obtenu le meilleur taux de succès (97%±3.98) avec la fonction de récompense qui utilise ce modèle pour prédire la satisfaction de l'utilisateur par rapport au modèle classique de calcul de récompense (81%±7.78) .

Dans le modèle classique chaque tour de parole est pénalisé avec -1 et si les informations correctes sont données par le système à la fin du dialogue une récompense finale de +20 est donnée dans le cas contraire la récompense finale est de 0.

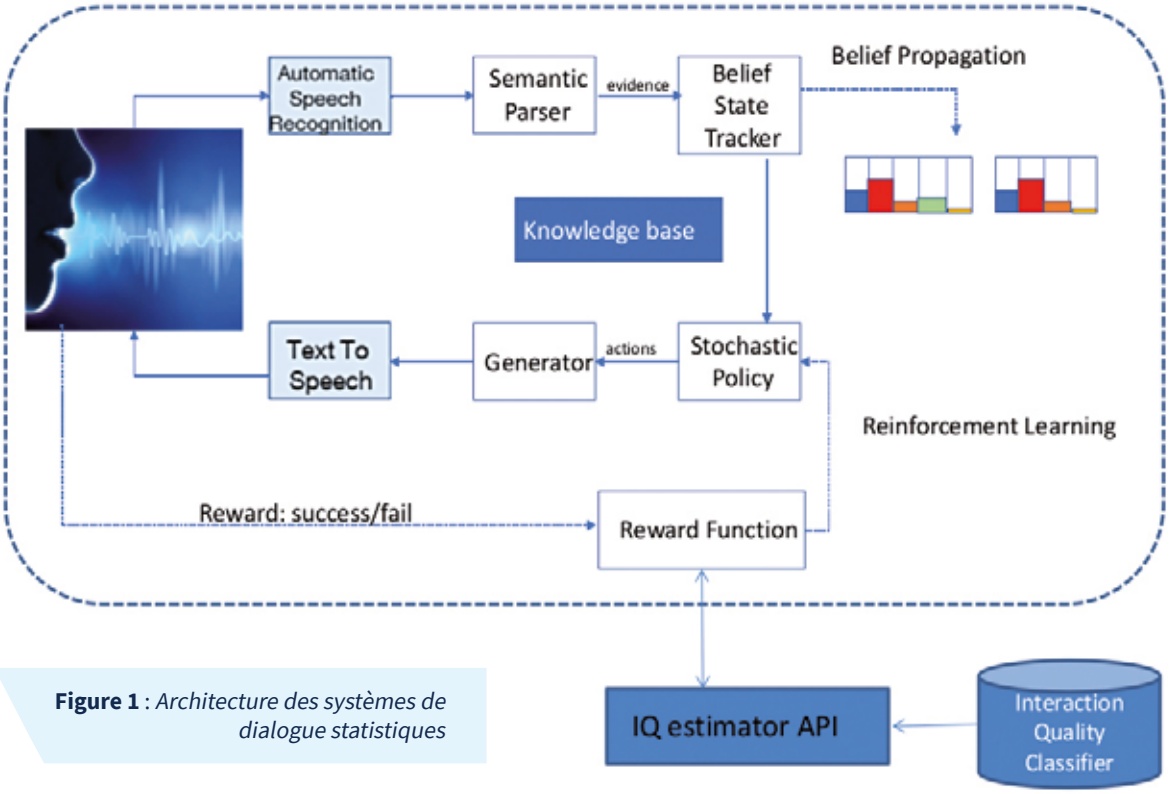


Figure 1 : *Architecture des systèmes de dialogue statistiques*

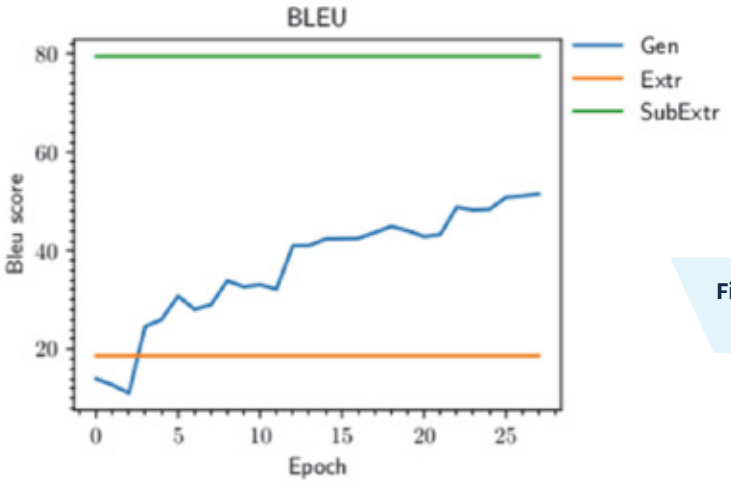


Figure 2 : *BLEU score des modèles génératifs vs modèles extractifs.*

[1] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics 2002 Jul 6 (pp. 311-318). Association for Computational Linguistics.

[2] Reddy S, Chen D, Manning CD. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics. 2019 May;7:249-66.

[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11

[4] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019 Oct 2

[5] Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860. 2019 Jan 9.

[6] Raux A, Langner B, Bohus D, Black AW, Eskenazi M. Let's Go Public! Taking a spoken dialog system to the real world. In Ninth European conference on speech communication and technology 2005.

[7] Ultes, Stefan, Rojas-Barahona, Lina M., SU, Pei-Hao, et al. Pydial: A multi-domain statistical dialogue system toolkit. In: Proceedings of ACL 2017, System Demonstrations. 2017. p. 73-78.

AutoDL Grand Challenge

Romain Egelé, Zhengying Liu,
Herilalaina Rakotoarison, Isabelle Guyon
TAU, LRI-CNRS–INRIA, Université Paris-Saclay, France


Romain Egelé


Zhengying Liu


Herilalaina Rakotoarison


Isabelle Guyon

L'Intelligence Artificielle (IA) moderne se base de plus en plus sur les réseaux de neurones artificiels. Ceux-ci, inventés au milieu du XX^{ème} siècle, découlent d'une analogie assez sommaire avec le cerveau. On attribue souvent leur essor récent à l'accroissement des moyens de calcul, combiné avec le développement des moyens de communication et de stockage, et, plus particulièrement, avec l'apparition des GPUs. La recherche industrielle met au service de l'IA des systèmes de calculs distribués gigantesques, comme les fameux hangars de serveurs des GAFAs aux Etats-Unis ou des BATX en Chine. Cependant, ces moyens restent inaccessibles à la recherche académique. Fort heureusement, apparaissent des institutions nationales telles que le supercalculateur Jean Zay, mettant au service des chercheurs académiques des moyens leur permettant de se lancer dans des recherches jusqu'ici hors de leur portée.

Fort heureusement, apparaissent des institutions nationales telles que le supercalculateur Jean-Zay, mettant au service des chercheurs académiques des moyens leur permettant de se lancer dans des recherches jusqu'ici hors de leur portée.

A l'heure actuelle, il existe une grande variété de types de réseaux de neurones : denses, convolutions, récurrents, Bayésiens, et bien d'autres. La qualité de l'architecture d'un réseau de neurones permet souvent une suprématie technologique forte tel que dans le domaine médical, de l'assistance vocale ou de la traduction linguistique. Le travail d'optimisation de l'architecture des réseaux neuronaux est le résultat d'une ingénierie couteuse en effort humain,résultant la plupart du temps d'essais et erreurs. Ce phénomène se reflète par ailleurs dans l'embauche croissante de « data scientists ». Notre projet de recherche s'articule autour de cette problématique en visant à AUTOMATISER la génération de réseaux de neurones pour l'apprentissage de modèles prédictifs. Ce genre de programme peut aider la recherche et l'industrie en accélérant les boucles itératives qui leur permettent de construire de nouveaux modèles prédictifs. Une première étape fut de pouvoir comprendre l'état de l'art des algorithmes décrits dans la littérature depuis 2016, quand le chercheur Quoc V. Le introduisit le terme de ***Neural Architecture Search*** (NAS: recherche d'architectures de réseaux de neurones) en arrivant à générer automatiquement le meilleur modèle d'analyse d'image du moment. Ce but fut atteint par l'utilisation de 800 GPUs pendant 28 jours ce qui équivaut 537 600 heures d'utilisation de GPUs.

A la suite de ces travaux un intérêt croissant pour NAS se fit ressentir. D'autres méthodes apparurent et arrivèrent à atteindre d'aussi bons résultats tout en utilisant environ 100 heures de GPUs. Nous nous sommes donc particulièrement intéressés à ces nouvelles approchent et avons remarqué deux aspects clef :

- les méthodes de la littérature s'appliquent exclusivement à l'analyse d'images et demandent un effort considérable pour être généralisés à d'autres domaines tel que l'analyse du texte, de la parole, du trafic routier, etc.
- la possibilité de réelles découvertes est toujours fortement bridée par l'omniprésence de l'intuition humaine pour la résolution du problème sous-jacent.

Dans ce contexte, nous avons utilisé le supercalculateur Jean Zay dans plusieurs projets, ce qui a mené à 2 publications potentielles. Les informations sur ces projets se trouvent dans le tableau suivant :

Nom du projet	Heures GPU consommées	Publication soumise à
AutoDL Benchmark	**1000**	**TPAMI**
Apprentissage des meta-features	**30000**	**NeurIPS 2020**

AutoDL Benchmark.
Pour comprendre et comparer les différentes méthodes de NAS (ou plus générale d'AutoML, pour *Automated Machine Learning*), nous avons formaté ~100 *datasets* sous un format tensoriel générique qui permet de représenter des données de presque toutes les modalités : image, vidéo, audio, texte, tabulaire et encore plus. (**Voir figure 1**)
Ensuite, des méthodes de NAS sont exécutés sur tous ces *datasets* et les résultats sont structurés et enregistrés. Ces résultats peuvent ensuite être réutilisés par de nombreux chercheurs, par exemple pour faire de la recommandation d'algorithmes, sans avoir à refaire des calculs monstrueux.

Apprentissage des *meta-features* :
l'objectif du projet AutoDL est de munir l'ordinateur de la capacité d'apprendre par l'expérimentation, mémoriser quelle est la bonne architecture pour une tâche, et évaluer si deux tâches sont similaires. Dans ce but, nous cherchons à apprendre une représentation des tâches les plongeant dans un espace qui induit une notion de proximité pertinente. En pratique, on cherche à définir des descripteurs de tâches, ou "*meta-features*". Ces meta-features sont apprises de telle sorte que, pour deux tâches similaires (ayant des *meta-features* voisines), les architectures neuronales les résolvant sont aussi similaires. Cela permettra, pour une nouvelle tâche T, de rechercher la tâche T' la plus similaire dans notre base de données, afin de reprendre ou adapter la meilleure architecture de T' pour résoudre T. L'étude comprend deux jalons : **1)** L'apprentissage des *meta-features* (prenant en entrée une « tâche » et retournant un vecteur de valeurs); et **2)** une preuve de concept de l'efficacité de la méthode.

Tous ces projets nous ont mené à une meilleure compréhension dans ce domaine de recherche très actif. Aucun de ces projets n'aurait été possible sans l'aide du supercalculateur Jean Zay. Nous n'avons aucun doute que le projet Jean Zay aidera dans le futur de nombreux chercheurs à avancer l'IA et la science en général.

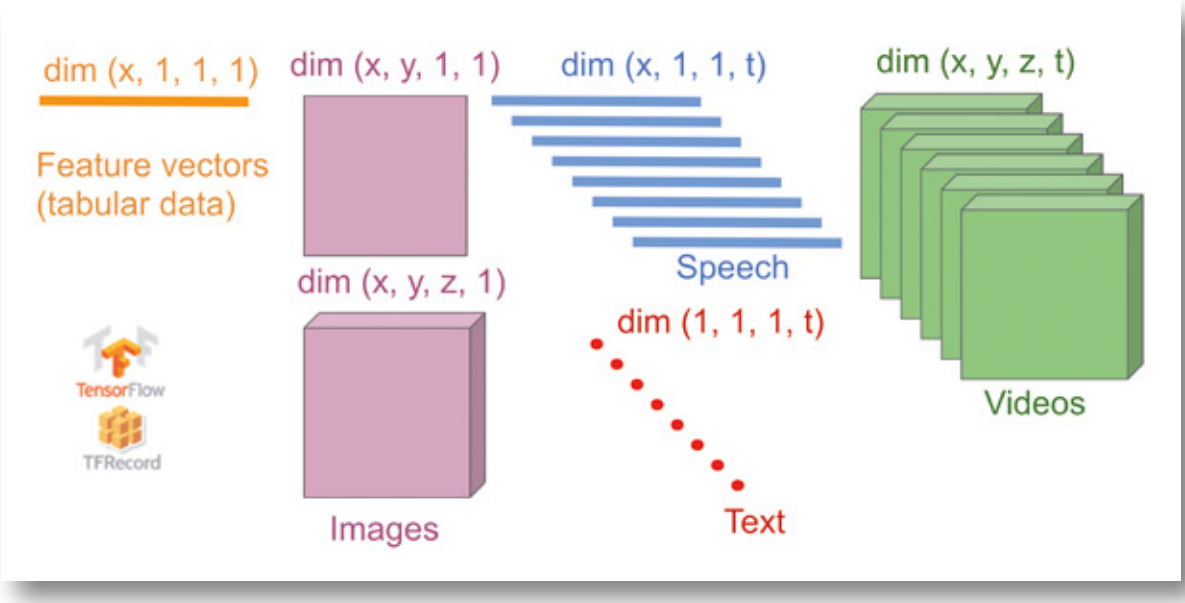


Figure 1: *Données brutes au format Tensor*

Illustration : *Artiste: Gerd Altmann Freiburg/Deutschland (source: Pixabay)]*

Analyse vidéo : génération de vidéos pour la reconnaissance d'activités


François Brémond


Yaohui Wang

François Brémond, Yaohui Wang
STARS, INRIA

Création d'images

La créativité est considérée comme un domaine très important de l'intelligence humaine. Nous utilisons la créativité pour produire de nouveaux arts, des théories scientifiques et des outils du quotidien. Les systèmes actuels d'IA (Intelligence Artificielle) de vision par ordinateur se concentrent principalement sur les tâches de reconnaissance (par exemple, la classification des images, la reconnaissance des actions dans les vidéos), la créativité n'ayant pas encore été largement explorée. Depuis qu'Alan Turing a posé la fameuse question « Les machines peuvent-elles penser ? » dans son article de 1950, nous pensons qu'une machine intelligente devrait avoir la capacité de créer et ce genre de capacité sera au centre de la prochaine étape de la communauté de l'intelligence artificielle.

Nous pensons que la capacité de la machine à créer de nouvelles vidéos est une des orientations futures de l'intelligence artificielle

Au cours de ces dernières années, Ian Goodfellow a proposé un puissant outil, les GANs (*Generative Adversarial Networks*). Cet outil peut générer/créer de nouvelles données en mettant en opposition à l'intérieur de son propre modèle, deux de ses réseaux de neurones pour atteindre un équilibre de Nash. De nombreuses méthodes ont été proposées par la suite, afin de synthétiser une meilleure qualité d'images générées et d'augmenter leur résolution. En 2019, deux méthodes très réussies, StyleGAN **[1]** et BigGAN **[2]** ont obtenu des résultats impressionnants sur les visages humains et également sur un ensemble d'objets généraux.

StyleGAN peut générer des images jusqu'à « 1024x1024 » pixels pour un visage humain sans utiliser d'annotations humaines, (voir Fig 1), qui sont très difficiles à distinguer des images réelles. En utilisant uniquement l'œil humain, il devient impossible de faire la différence entre les données réelles et les données synthétisées.

De même, BigGAN se concentre principalement sur la génération d'images conditionnelles (voir Fig 2), ce qui signifie générer une image en fournissant l'étiquette de catégorie (par exemple, chat, chien, avion ...). Ce type de génération est très utile, car il permet de générer de nouveaux ensembles d'images avec plus de variété pour résoudre de nouveaux problèmes, tels que les troubles mentaux où les données images avec des personnes âgées ne sont pas souvent disponibles. Pour qu'un réseau de neurones puisse apprendre efficacement, il lui faut suffisamment de données (e.g. images) correspondant au problème traité. Ainsi, BigGAN est entraîné sur un ensemble de données à grande échelle - ImageNet, qui contient 1 000 catégories avec plus de 14 millions d'images. Comparé à StyleGAN, BigGAN nécessite un réglage de configurations beaucoup plus complexe, mais peut produire des résultats plus diversifiés.

Il existe également de nombreuses autres méthodes (par exemple, CycleGAN **[3]**, MUNIT **[4]**, ...) se concentrant sur différentes tâches de génération d'images. Grâce à ces méthodes, les machines peuvent créer des images simples (correspondant à des nombres) jusqu'à des scènes très complexes (e.g. église). La production d'images synthétisées simples est une étape très importante pour concevoir une machine créative, car les peintres ont également commencé à apprendre le dessin en dessinant des scènes allant du simple au complexe. L'utilisation du même processus d'apprentissage dans l'apprentissage automatique peut nous aider à mieux comprendre comment les êtres humains apprennent et quelle devrait être la bonne façon pour la machine d'apprendre.

Création de Vidéos

Cependant, nous vivons dans un monde dynamique où tout bouge selon ses propres motifs. La machine peut-elle apprendre également à générer des scènes dynamiques ou, en d'autres termes, des vidéos ? C'est une question très intéressante, car une fois que cela est possible, nous pouvons utiliser la machine pour faire de courtes vidéos jusqu'aux films à long métrage, directement sans intervention humaine. Cela peut réduire le coût de réalisation d'un film tout en maintenant une grande variété.

Nous avons commencé à explorer ce problème à partir de courtes vidéos (visages avec émotions et actions humaines élémentaires). Nous avons proposé une méthode simple, mais efficace pour résoudre ce problème **[5]**. Nous pensons que toute vidéo doit avoir deux facteurs principaux importants, l'apparence (qui vous êtes) et le mouvement (comment vous vous déplacez). Cependant, il est extrêmement difficile de traiter simultanément ces deux facteurs, car le

Figure 1 : *StyleGAN*

Figure 2 : *BigGAN*

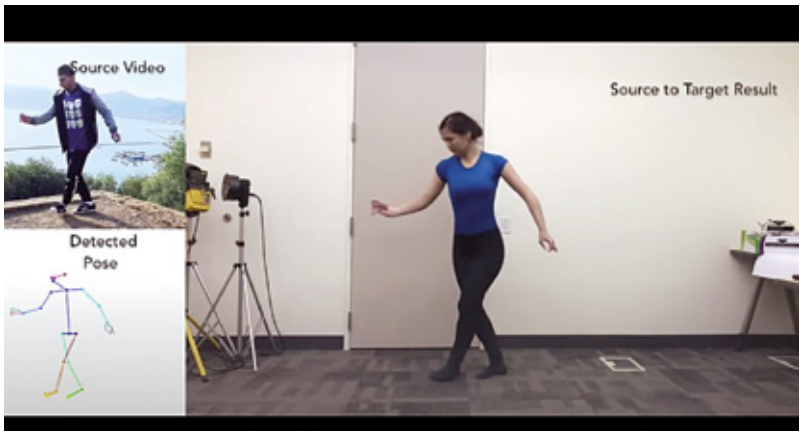


Figure 3 : *EveryBody Dance Now*

modèle doit apprendre à préserver l'apparence et à modéliser la cohérence temporelle de l'action en même temps. Afin de faciliter la tâche, nous décomposons ces deux facteurs et laissons le modèle les gérer en parallèle. Cette méthode donne de bons résultats et permet de générer des vidéos humaines réalistes avec suffisamment de diversité. Ces résultats n'ont pu etre réalisés que dans le cadre des Grands challenges GENCI et ils ont nécessité plus de 1 000 heures de calcul sur 4 GPU.

Cependant, des défis demeurent. Actuellement, nous ne pouvons générer que des courtes vidéos de faible résolution. Afin de simplifier l'apprentissage, des chercheurs de l'Université de Californie à Berkeley ont proposé une nouvelle façon de générer une vidéo synthétique à partir d'une autre réelle (voir Fig 3). Par exemple, cette technique permet de générer une vidéo mimant la danse de Micheal Jackson. Pour y parvenir, nous n'avons besoin que de deux choses, quelques photos de votre corps et le clip vidéo « Dangerous » de Micheal Jackson. En introduisant ces deux éléments dans le modèle, la technique proposée peut automatiquement apprendre à calquer votre apparence sur le mouvement de Micheal Jackson. Bien que cette technique ne considère la génération d'une vidéo qu'avec une seule personne ayant des mouvements répétés, les résultats sont assez prometteurs et réalistes. Grâce à ce travail, la génération de vidéos haute résolution devient possible.

Conclusion

Ce court article a abordé la synthèse d'images et de vidéos. Nous pensons que la capacité de la machine à créer de nouvelles vidéos est une des orientations futures de l'intelligence artificielle, car elle permet d'entrainer la machine sur de nouveaux sujets, grâce à ces bases de vidéos synthétiques. Nous espérons que des méthodes de plus en plus efficaces verront le jour et amélioreront ainsi notre quotidien.

Références

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial net-works," in CVPR, 2019.

[2] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in ICLR, 2019.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in ICCV, 2017.

[4] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in ECCV, 2018.

[5] Y. Wang, P. Bilinski, F. Bremond and A. Dantcheva. G^{3}AN: Disentangling appearance and motion for video generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020, online, June 14-19, 2020.

Réduction des biais pour la tâche de comptage d'objets visuels à partir de questions


Matthieu Cord


Corentin Dancette


Rémi Cadène

Matthieu Cord, Corentin Dancette, Rémi Cadène
Sorbonne Université, CNRS, LIP6.

Notre projet s'inscrit dans le cadre du développement de modèles de raisonnement visuel. Il s'agit de concevoir des modèles informatiques capables de prendre des décisions complexes, nécessitant parfois un processus itératif, ayant pour objet des données visuelles. Par exemple, la tâche de *Visual Question Answering* (VQA) consiste à répondre à des questions sur des images, telles que "De quelle couleur est le tee-shirt de l'homme dans cette image ?" Les modèles de VQA pourraient par exemple servir, à l'avenir, à guider des personnes malvoyantes disposant d'une caméra. Cette tâche nécessite des capacités de raisonnement visuel.

Il s'agit de concevoir des modèles informatiques capables de prendre des décisions complexes ...

Cette capacité de raisonnement visuel est nécessaire pour communiquer efficacement avec une personne, à propos de son environnement, et de prendre des instructions (instructions vocales, textuelles, gestuelles, ou d'autres formes). Le raisonnement permet de comprendre l'instruction et de donner la réponse adéquate.

L'approche majoritaire aujourd'hui pour modéliser ces tâches est le *Deep Learning*, une branche du *Machine Learning* ou apprentissage statistique. Une problématique récurrente dans les tâches de raisonnement visuel à laquelle nous nous sommes intéressés est celle des biais [1] : il s'agit de corrélations superficielles entre l'entrée du modèle et la réponse. Ces biais sont particulièrement présents dans les *datasets* collectés dans le monde réel (non synthétiques), et il est plus simple pour un modèle d'utiliser ces biais pour apprendre à prédire que d'apprendre le comportement désiré. Par exemple, un modèle répondra toujours jaune à la question "De quelle couleur est la banane dans l'image ?" sans analyser la véritable couleur, car il a observé en grande majorité des bananes jaunes durant son apprentissage. Toutefois, lors de l'utilisation dans le monde réel, dans des situations variées, ces corrélations ne seront pas toujours les mêmes, ce qui induira des réponses erronées. Une autre problématique des modèles de *Deep Learning* est leur interprétabilité. En effet, ces modèles sont constitués d'un enchaînement de fonctions complexes, rendant difficile de comprendre leurs prédictions. Ce manque de compréhension du raisonnement du modèle peut entraîner un manque de confiance de la part des utilisateurs.

Pour étudier ce problème, deux approches sont majoritaires :
D'une part, l'explication des prédictions a *posteriori*, où il s'agit d'expliquer les décisions d'un modèle existant, déjà entraîné.
D'autre part, l'intégration *a priori* de structure dans les modèles permettant de rendre leurs prédictions interprétables par un humain. Dans ce cas, l'explication est fournie en même temps que la prédiction, et peut prendre différentes formes (textuelle, visuelle ...). Notre projet s'inscrit dans cette deuxième approche.

En particulier, nous nous sommes intéressés à la tâche de comptage d'objets liés à une question. Ainsi, étant donné une image et une question, par exemple "Combien y a-t-il de chats gris dans l'image ?", l'objectif est d'analyser l'image donnée, et de renvoyer un nombre en réponse. Nous utilisons le *dataset* de questions de comptage TallyQA [2] pour nos expériences.
Notre objectif est, d'une part, d'évaluer les biais présents dans les modèles ayant appris à résoudre cette tâche. D'autre part, nous proposons une approche plus robuste aux biais, et plus interprétable, grâce à l'intégration d'*a priori* de structure.

Protocole d'évaluation des modèles

Notre première contribution est la création d'un protocole d'évaluation permettant de pénaliser les modèles ayant appris des corrélations superficielles, sans avoir appris le mécanisme de comptage. Ce protocole est fondé sur un changement de distribution entre les données d'entraînement et les données de test. Il s'inspire du protocole VQA-CP [1], utilisé pour évaluer les biais pour la tâche de *Visual Question Answering*. L'objectif est d'introduire des différences importantes entre les biais présents en entraînement et en test. Cela permet de pénaliser un modèle qui s'appuie sur les corrélations superficielles.
Nous allons entraîner notre modèle majoritairement sur des exemples ayant une réponse paire (0, 2, 4, 6 objets à compter dans l'image) et le tester majoritairement sur des questions ayant une réponse impaire (1, 3, 5, ...).
Plus précisément, pour les exemples d'entraînement, nous allons retirer une proportion p des exemples ayant une réponse impaire. Pour les exemples de test, au contraire, nous allons retirer une même proportion p d'exemples ayant une réponse paire. Ainsi, si p est proche de 100%,

l'ensemble d'entraînement sera majoritairement constitué d'exemples pairs, alors que l'exemple de test majoritairement constitué d'exemples impairs. Ainsi, un modèle s'appuyant sur des corrélations superficielles sera pénalisé, alors qu'un modèle robuste ayant appris le mécanisme de comptage, pourra généraliser le comportement appris sur des réponses paires à des questions aux réponses impaires.

Spatial Counting Network

Notre deuxième contribution est le développement d'un modèle à la fois plus robuste que les modèles de l'état de l'art, sur notre protocole, et qui soit également davantage interprétable.
Nous avons utilisé le supercalculateur Jean Zay pour mener une recherche de l'architecture optimale. Notre modèle final est représenté en **Figure 1**. Il intègre des a priori de structure importants pour cette tâche de comptage d'objets. Premièrement, ce modèle fonde sa prédiction finale sur les objets présents dans l'image. Ainsi, au lieu de compter de manière globale le nombre d'objets comme la plupart des modèles de comptage, notre modèle sélectionne les objets importants individuellement, leur assigne un score proche de 0 ou 1, et additionne ces scores pour donner un résultat final.
Ce mécanisme permet au modèle d'être plus interprétable, et d'être moins soumis aux biais présents dans les données. Il possède également une modélisation relationnelle permettant de comparer les objets de l'image entre eux, ce qui est utile pour les questions complexes (par exemple "Combien y a-t-il de chats à gauche de l'homme ?"). Ce modèle est entraîné par descente de gradient stochastique, sur le supercalculateur.

Résultats

Les résultats sont affichés sur la **Figure 2**. Notre modèle (SCN) est comparé à RCN [2] et RCN Regression, deux variations d'un modèle état de l'art pour la tâche de comptage.
En abscisse est indiqué le pourcentage d'exemples pairs et impairs retirés respectivement des données d'entraînement et de test.
Nous pouvons voir que notre modèle est plus performant que ces deux modèles pour les valeurs élevées qui correspondent à un gros changement de distribution entre l'entraînement et le test.
Ceci confirme la meilleure robustesse de notre modèle face à ces corrélations superficielles.

Perspective

Une perspective de développement est la généralisation de notre protocole d'évaluation à d'autres tâches de *machine learning* qui contiennent également de nombreux biais superficiels. Nous pensons également que le développement de modèles plus interprétables est un sujet de recherche important.

[1] Agrawal, Aishwarya, et al. «*Don't just assume; look and answer: Overcoming priors for visual question answering.*» *Proceedings of the IEEE Conference on Computer Vision and Pattern* Recognition. 2018.

[2] Acharya, Manoj, Kushal Kafle, and Christopher Kanan. «TallyQA: *Answering complex counting questions.*» *Proceedings of the AAAI Conference on Artificial* Intelligence. Vol. 33. 2019.

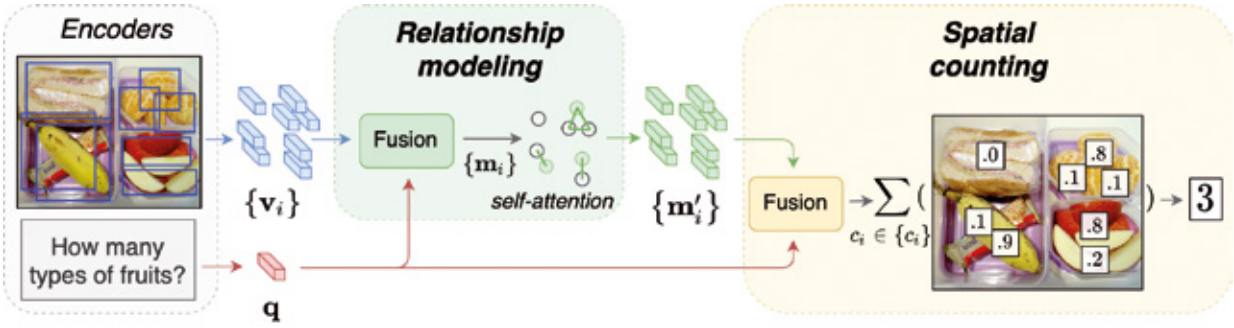


Figure 1

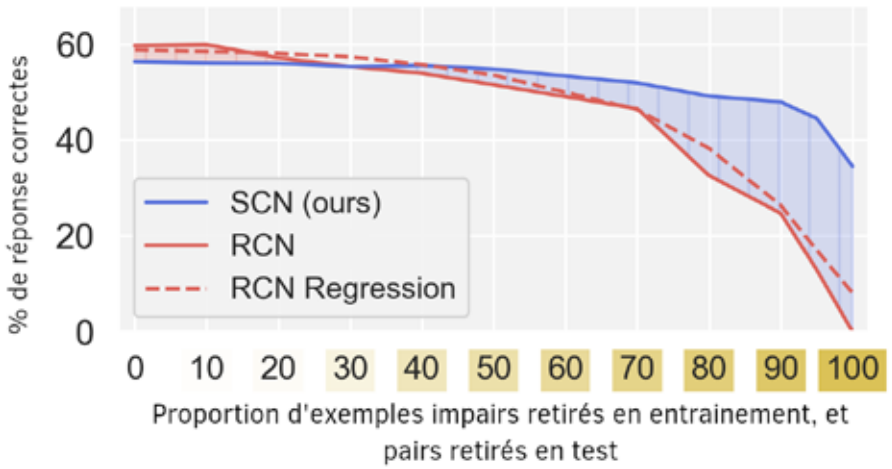


Figure 2

Le *deep learning* à l'épreuve des pirates informatiques.

L'équipe (de gauche à droite) : Benjamin Negrevergne, Yann Chevaleyre, Rafaël Pinot, Jamal Atif, Laurent Meunier, Geovani Rizk, Alexandre Araujo, Florian Yger, Céline Beji

Alexandre Araujo, Équipe Miles, Lamsade, Paris Dauphine, Université PSL, Wavestone
Benjamin Negrevergne, Équipe Miles, Lamsade, Paris Dauphine, Université PSL

Malgré de nombreux succès récents, le *Deep Learning* n'est pas infaillible. Il est possible de fabriquer une image dont le contenu, facilement identifiable par n'importe quel être humain, sera inévitablement confondu par un réseau de neurones avec une tout autre chose. Pire, les utilisateurs malintentionnés qui fabriquent des images dans le but de tromper les réseaux de neurones peuvent altérer la perception de ces réseaux, pour les amener délibérément à confondre les chats avec des armes à feu, ou les cyclistes avec des bornes kilométriques. Les problèmes que peut poser l'existence de telles images sont nombreux. On imagine sans difficulté les dégâts que pourrait provoquer une image manipulée sur un panneau publicitaire, installée au bord de la route pour tromper des voitures autonomes. À l'instar des systèmes informatiques des années 1980, les réseaux de neurones actuels sont vulnérables et doivent être mieux maîtrisés avant de pouvoir être déployés auprès du grand public. L'équipe MILES de l'Université PSL – Paris Dauphine, s'emploie à mieux comprendre les représentations et les mécanismes internes qui permettent aux réseaux de neurones de prendre une décision, ainsi qu'à développer des techniques pour entraîner des réseaux de neurones robustes aux attaques. Grâce à la mise en place du supercalculateur Jean Zay, et dans le cadre d'une allocation Grands challenges de 26 000 h GPU, l'équipe a pu contribuer à améliorer la compréhension des attaques, et proposer de nouvelles techniques de défense.

les réseaux de neurones actuels sont vulnérables et doivent être mieux maîtrisés avant de pouvoir être déployés auprès du grand public.

Le principe qui permet aux utilisateurs malintentionnés de générer des images manipulées est relativement simple à condition d'adopter un point de vue géométrique pour analyser le fonctionnement des réseaux de neurones. Pour un réseau de neurones, chaque image est un point dans un espace. C'est la couleur des pixels de l'image qui défini les coordonnées du point qui lui correspond. Initialement tous les points sont mélangés, mais le réseau applique une série de transformations successives qui tordent et déforment l'espace de départ, de sorte à former des groupement d'images sémantiquement cohérents dans l'espace d'arrivée.

Le défi pour un utilisateur mal intentionné, consiste donc a construire une image,qui serait proche de l'image cible dans l'espace de départ (c'est-à-dire : visuellement similaire), mais éloignée, dans l'espace transformé par le réseau (c'est-à-dire sémantiquement différente). Même s'il peut sembler difficile de trouver de telles images autrement que par accident, l'exercice peut être résolu avec une simple procédure d'optimisation qui exploite les imperfection de la procédure d'apprentissage. Finalement, presque toutes les images naturelles peuvent être altérées de la sorte.

Parmi les techniques à l'étude pour construire des réseaux de neurones robustes : les réseaux de neurones aléatoires. L'approche peut paraître surprenante, mais elle fonctionne : on injecte du bruit dans un réseau de neurones au moment de faire une prédiction pour perturber son comportement et le rendre moins déterministe. Face à l'incertitude, le mécanisme qui cherche à attaquer le réseau est obligé de faire des compromis sur la qualité de son attaque, ce qui la rend moins efficace. Nous avons pu montrer que les mécanismes de défense basés sur l'injection de bruits issus de la famille exponentielle permettaient d'offrir des garanties de fiabilité, ce qui constitue un progrès important par rapport aux techniques de défense empiriques qui n'offrent aucune garantie. Le problème : l'injection de bruit dégrade aussi la qualité de la prédiction avec des images naturelles, non perturbées. Pour que les réseaux soient à la fois précis avec des images naturelles et résistants aux manipulations, ils doivent être entraînés avec des données bruitées. Il s'agit d'un équilibre subtil **[1].**

Malgré l'effort de recherche considérable investi dans le développement de techniques qui permettent aux réseaux de neurones de se défendre, un problème fondamental subsiste : ce sont les attaquants qui choisissent leurs armes. Les défenseurs, eux, doivent se préserver d'une hypothétique attaque dont ils ne peuvent pas connaître la nature à l'avance. Ainsi, pour présenter un véritable intérêt pratique, les mécanismes de défense développés pour les réseaux de neurones doivent offrir une protection solide, non pas contre une seule, mais contre toutes les menaces. Or pour l'instant, l'essentiel des mécanismes de défense existants est développé pour répondre à un seul type d'attaque.

Un aspect important qui diffère d'une attaque à une autre, c'est la définition formelle utilisée pour caractériser le concept proximité visuelle. Ces différences subtiles sont souvent sans importance dans un espace classique avec 2 ou 3 dimensions, mais elles deviennent primordiales dans un espace avec plusieurs centaines ou plusieurs milliers de dimensions comme c'est le cas de l'espace des images. C'est un problème classique : les difficultés liées au passage d'un espace avec peu de dimensions vers un espace avec beaucoup de dimensions sont

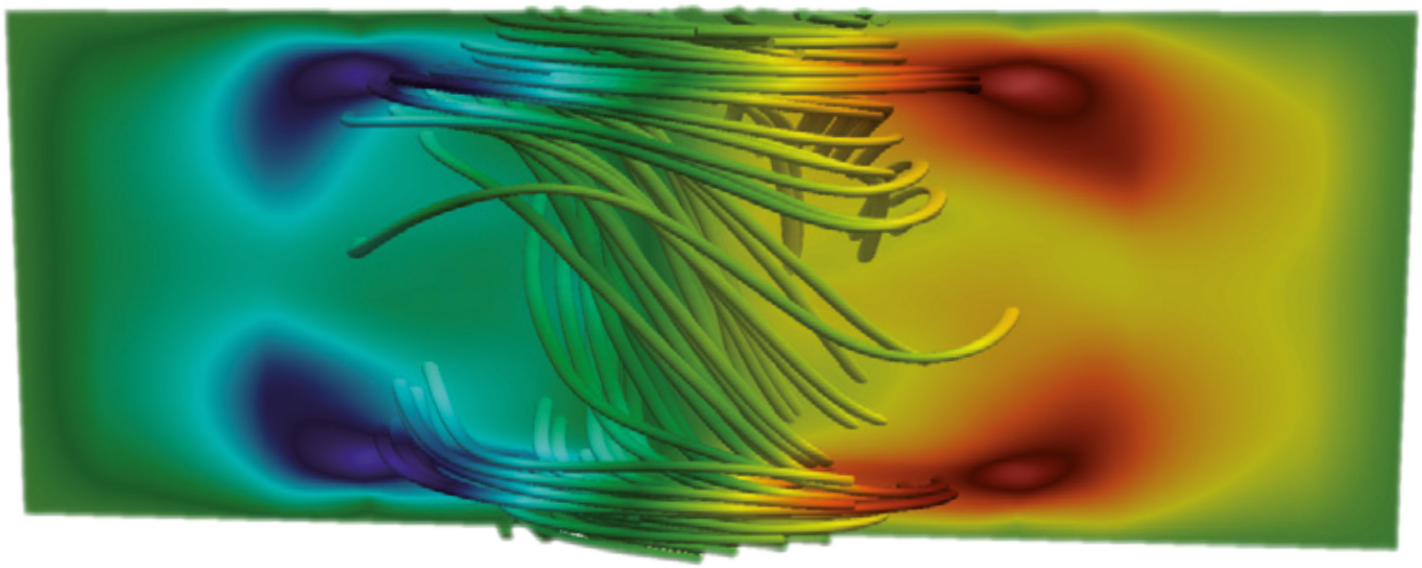
tellement fréquentes que les statisticiens leur ont donné un nom : il parlent de « la malédiction de la dimensionnalité ». En essayant de proposer des mécanismes de défense plus généraux, nous nous sommes rendu compte que les différences subtiles dans la définition du concept de proximité visuelle, avaient des conséquences majeures et largement sous-estimées, sur la nature et les propriétés des images générées. Conséquence directe de cette observation : les mécanismes de défense conçus pour contrer une attaque particulière se révèlent remarquablement inefficaces contre les autres attaques **[2].**

Existe-t-il un mécanisme de défense qui assure une robustesse optimale contre toutes les attaques ? Nous avons tenté de répondre à cette question en abordant le point de vue de la théorie des jeux **[3]**. Grâce à cette nouvelle perspective, nous avons pu montrer que la confrontation entre les attaques et les défenses forment un jeu à somme nulle infinie où les résultats classiques (tel que le théorème min-max de Sion) ne s'appliquent pas. Dans ce contexte, nous avons démontré qu'il n'existait pas d'équilibre de Nash, lorsque le défenseur et l'attaquant sont tous deux déterministes, donnant ainsi une première réponse (négative) à la question posée ci-dessus, pour le régime déterministe. La question reste cependant ouverte lorsque le défenseur et l'attaquant peuvent tous les deux montrer des comportements non-deterministes sur toutes les tâches.

[1] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, J. Atif. Theoretical evidence foradversarial robustness through randomization. Neurips 2019.

[2] A. Araujo, L. Meunier, R. Pinot, B. Negrevergne. Robust neural networks using randomized adversarial training.

[3] R. Pinot, R. Ettedgui, G. Rizk, Y. Chevaleyre, J. Atif. Randomization matters. How to defend against strong adversarial attacks. ICML 2020.



Simulation magnétohydrodynamique d'un écoulement turbulent de sodium liquide dans la géométrie de von Kármán Sodium.
Visualisation des lignes de champ magnétique et plan de coupe de ce champ dans le volume de fluide. Code SFEMaNS, calculs parallèles effectués à l'IDRIS.
Caroline Nore : LIMSI, CNRS, Université Paris–Saclay ; **Daniel Castanon Quiroz :** IMAG, CNRS, Université de Montpellier ; **Loïc Cappanera :** DCAM, Rice University ; **Jean-Luc Guermond :** DM, Texas A&M University

Directeur de la publication : Pierre-François Lavallée

Coordination : Geneviève Morvan et Thierry Goldmann

Crédits photos :
Page de couverture : Gerd Altmann de Pixabay et CDC/ Alissa Eckert, MS; Dan Higgins, MAM / Public domain

4e de couv, pages 5 et 6 : © Cyril FRESILLON / IDRIS / CNRS Photothèque

«Avec Jean Perrin nous créâmes le Centre national de la recherche scientifique»
Jean Zay, 29 avril 1942, Souvenirs et solitude

Institut du
Développement et des
Ressources en
Informatique
Scientifique

IDRIS
Rue John von Neumann, BP 167, Bâtiment 506,
91403 Orsay Cedex - France
www.idris.fr

Réalisation & impression Créaprod -01 69 09 94 27 - RCS 482 246 170