Guide utilisateurs pour les demandes de ressources sur les moyens nationaux de GENCI

Paris, le 1er mars 2022
Version 1.5

Ce document a pour objectif de guider les utilisateurs lors du dépôt d'un renouvèlement de dossier ou lors de la création d'un dossier pour un nouveau projet sur l'application eDARI, afin de bénéficier de ressources informatiques sur les moyens nationaux de GENCI. Les informations demandées dans les différentes parties du document permettent aux Comités Thématiques et aux directeurs de centres d'évaluer les avancées scientifiques mais aussi de justifier éventuellement une demande de ressources sur un calculateur d'un centre national.

À noter, qu'il est **nécessaire de déposer**

- **<u>Pour un dossier d'accès régulier</u> : une description détaillée du projet pour les nouveaux et anciens utilisateurs (partie 1)** ;
- **<u>Pour un dossier d'accès régulier ou d'accès dynamique</u> : un rapport d'activité pour les utilisateurs ayant déjà obtenu une allocation de ressources** sur un ou plusieurs supercalculateur(s) des trois centres nationaux (CINES, IDRIS et TGCC) **(partie 2)**.

Table des matières

1. Description du projet 2

1.1 Informations générales 2

1.2 Résumé 2

1.3 Présentation générale 2

1.4 Méthode 3

1.4.1 Méthode numérique et implémentations 3

1.4.2 Justification de l'usage des ressources sur la machine demandée 6

1.5 Plans de gestion de données (DMP) 8

1.6 Bibliographie 8

2. Rapport d'activité 9

2.1 Informations générales 9

2.2 Résultats scientifiques 9

2.3 Bibliographie 10

2.4 Conférences et posters 10

1. Description du projet

La description du projet est nécessaire afin de permettre aux Comités Thématiques d'évaluer la qualité scientifique et technique d'un dossier d'accès régulier (ressources supérieures à 50 000 heures GPU / 500 000 heures cœurs CPU). Les différentes parties demandées sont les suivantes : Les informations générales concernant le dossier, un court résumé, une présentation générale du projet, une description de la méthode utilisée, une justification de l'usage des ressources demandées sur le(s) supercalculateur(s), la description d'un Plan de gestion des données (DMP) et une bibliographie. **Il est important de noter qu'une absence des informations demandées aura un impact direct sur la note scientifique et/ou technique attribuée au dossier et donc sur l'attribution d'heures.**

1.1 Informations générales

Dans cette partie, les informations générales d'un dossier telles que son titre et numéro, le nom de la structure de recherche, le responsable scientifique et le nombre d'heures demandées devront identifier une demande.

1.2 Résumé

La longueur typique du résumé scientifique est d'environ 15 lignes et d'une longueur maximale d'une page. Une introduction au projet scientifique est attendue avec le positionnement par rapport au contexte ou état de l'art international, ainsi qu'un bref résumé des derniers résultats obtenus dans le domaine (soit au sein de l'équipe/laboratoire soit au niveau international).

1.3 Présentation générale

Dans cette partie, il s'agit de détailler l'intérêt scientifique du projet, comme :

- **les objectifs** concrets à atteindre grâce à l'obtention de ces ressources et comment y parvenir ;
- **la situation des travaux de l'équipe** sur le thème de recherche proposé vis-à-vis du travail déjà effectué par l'équipe (résultats acquis sur le sujet), et vis-à-vis d'autres travaux sur un plan national et international ;
- **la pertinence des publications de l'équipe** dans ce domaine de recherche, référencées dans la section « Bibliographie » à la fin de la description du projet. Tout article, jugé utile à l'évaluation du dossier, peut être annexé au format PDF sur eDARI. Les articles

qui ne sont pas en *open access* ou en version *preprint* seront effacés après chaque Comité d'Attribution.

La longueur typique de cette partie est d'environ deux pages avec une longueur maximale de quatre pages. Cependant, si le projet se décompose en sous-projets, il est possible de rajouter deux pages additionnelles au maximum par sous-projets.

1.4 Méthode

Cette partie doit être suffisamment précise et argumentée pour permettre au Comité Thématique d'apprécier l'adéquation de l'architecture prévue (scalaire, vectorielle ou parallèle, GPU ...) au problème posé. Une explication détaillée (maximum une demi page) est demandée afin de préciser les raisons pour lesquelles la méthode numérique et l'usage sont adaptés à un supercalculateur national (Tier-1).

1.4.1 Méthode numérique et implémentations

La longueur typique est de quatre pages et la longueur maximale est de six pages. Si la méthode utilisée est très connue dans le domaine scientifique en question ou si le logiciel utilisé est disponible dans le centre, l'explication peut être plus courte.

Voici un exemple de plan à adapter selon le sujet :

- **Algorithme utilisé et adaptation à l'architecture visée**
- **Modalités d'optimisation** (vectorisation, optimisation superscalaire, parallélisation)
- **Structure du programme**
- **Logiciels nécessaires**
- **Langages utilisés**
- **Bibliothèques prévues**
- **Systèmes de gestion de bases de données ou systèmes documentaires utilisés.**

Des exemples ou descriptifs des points attendus sont fournis par des présidents de Comités Thématiques ci-après.

Descriptifs pour le CT6

- Pour des projets nécessitant des résolutions d'équations, préciser les méthodes numériques (éléments finis, volumes finis, maillages, discrétisation temporelle, …). Cela peut être succinct s'il s'agit de solveurs classiques.
- Pour les aspects résolution de systèmes linéaires, là aussi, préciser le type de méthodes (factorisation, solveurs directs/inverses, …), la complexité attendue, la dimension, …

Descriptifs pour le CT7

1) Description précise des systèmes d'étude

Les éléments suivants devront être détaillés :

- La provenance des structures de macromolécules (cristallographique, par homologie, etc.), nombre d'atomes du système de simulation, présence de solvant explicite, de bicouche explicite, etc.
- Le type de modèle (tout-atome, united-atom, gros grain, etc.) et le champ de force utilisé.
- Le code utilisé (GROMACS, NAMD, LAMMPS, OpenMM, etc.).
- Pour les simulations tout-atome de systèmes basés sur un modèle par homologie, il est essentiel que le comité puisse juger de la pertinence du modèle. L'information minimale devra être : Le pourcentage d'identité par rapport au support. Le modèle a-t-il été validé contre des données expérimentales ?

2) Justification numérique de l'approche

- Justifier si les temps simulés demandés permettront d'observer le phénomène attendu compte-tenu des problèmes d'échantillonnage, de convergence et de reproductibilité. Est-il plus intéressant de faire une simulation plus longue ou plusieurs plus courtes ?
- Pour les simulations dont le but est le calcul d'une grandeur (par exemple calcul d'énergie libre), à quelle barre d'erreur peut-on s'attendre sur cette grandeur avec le protocole envisagé.

Pour les deux aspects, les candidats pourront se baser sur des exemples similaires dans la littérature ou sur leur propre expérience si ce n'est pas publié.

Exemple pour le CT8

- Les travaux théoriques développés dans le projet ont nécessité l'utilisation de l'approche de la théorie de la fonctionnelle de la densité (DFT) dans la cadre de calculs en condition périodique (DFT-PBC). La fonctionnelle d'échange-corrélation utilisée est de type GGA PBE. Les calculs électroniques ont nécessité l'étude des effets de polarisation de *spin* et donc les propriétés magnétiques ont été abordées tout au long de l'étude. Afin de décrire de manière effective les interactions faibles de type van der Waals (dispersion) dans les systèmes chimiques considérés, une correction semi-empirique à l'approche DFT de type D3 (Grimme) a été systématiquement appliquée au calcul de l'énergie électronique totale.
- Les calculs ont tous été réalisés au moyen du logiciel de calcul VASP (version 5.4.1) disponible sur les ressources du GENCI. Les calculs ont été menés en environnement parallèle MPI, en utilisant de manière routinière entre 32 et 64 cœurs sur les serveurs alloués, pendant le temps d'exécution.
- Les modes de calcul considérés ont été les optimisations de géométrie (gradient conjugué), le calcul vibrationnel (calcul par différences finies) et la recherche de chemins réactionnels (approche dite de la bande élastique CI-NEB).
- Pour les optimisations de géométrie, les critères de convergence sont relativement forts (EDIFF=10^{-6} et EDIFFG=-0.01 eV.$Â^{-1}$).
- Pour le calcul vibrationnel, l'approche par différences finies a été utilisée avec un critère de convergence encore plus strict (EDIFF=10^{-7}) et deux points de calcul pour évaluer les dérivées secondes de l'énergie électronique totale (NFREE=2). Tous les degrés de liberté optimisés n'ont pas été retenus pour l'analyse vibrationnelle en raison du coût de calcul trop grand. Seuls les degrés de liberté de la molécule ont été considérés pour cette analyse.
- Concernant la recherche d'états de transition, la méthode NEB a été développée en utilisant 8 images intermédiaires par chemin réactionnel étudié. Lorsqu'il n'était plus possible de continuer à raffiner le profil énergétique associé au chemin réactionnel, la détermination de la structure géométrique de l'état de transition a été poursuivie au moyen de l'approche CI-NEB qui relâche la contrainte de force tangentielle/normale sur le point énergétique le plus élevé du profil, puis par une optimisation fine à l'aide de l'algorithme de quasi-Newton.

Descriptifs pour le CT10

Pour le CT10 et les dossiers relatifs aux nouvelles thématiques de l'Intelligence Artificielle, la démarche reste similaire. Les dossiers indiqueront le domaine applicatif concerné par l'IA et détailleront la méthodologie adoptée. Par exemple, *à titre indicatif*, quel type d'apprentissage (supervisé ou non, avec ou sans renforcement…), quel type de réseau de neurones (nombre de couches, mécanisme d'attention, transformer …), quel logiciel associé (open source, spécifique…), quel type de données (images, son, texte…), quelle durée de calcul (selon le nombre de GPU nécessaire…) et quel volume de données en distinguant apprentissage et tests (voir § 1.4.2). Si besoin, évoquer les questions de licence de logiciel, de propriété intellectuelle, de confidentialité des données.

1.4.2 Justification de l'usage des ressources sur la machine demandée

La longueur typique est d'une page et la longueur maximale de deux pages. La justification du nombre d'heures demandées en heure par cœur ou GPU doit inclure les informations suivantes :

Nombre et nature des tâches calculatoires du projet : Pour chaque type de tâche calculatoire, il faudra indiquer les ressources impliquées en mentionnant le temps de présence en machine, la mémoire, le nombre de nœuds, le nombre de cœurs ou GPU et l'espace disque en termes de nombre de fichiers et volumétrie (il est possible de décrire précisément ce point au chapitre suivant). Le nombre de tâches qui seront soumises de manière simultanée sur la machine doit être mentionné. L'ensemble de ces informations devront être présentées dans un tableau.

Descriptifs pour le CT7

Chaque calcul envisagé doit être décrit précisément. Par exemple, pour des simulations de dynamique moléculaire, le système, le nombre de simulations, la durée (en ns/micros), les heures de calcul nécessaires, le nombre de répliques, les détails sur un éventuel échantillonnage augmenté (REMD, métadynamique, ...).

Pour les demandes conséquentes (supérieures ou égales à 1 million d'heures) et si le système n'a jamais tourné sur un centre national, un accès préparatoire est vivement conseillé afin de générer une courbe de *scaling* sur le centre demandé. Celle-ci accompagnera la demande et le temps demandé sera estimé au mieux.

Afin de faciliter le travail des évaluateurs, l'usage d'un tableau pour les simulations envisagées est recommandé.

Exemple pour le CT7

Projet	Système	Nombre de simulations	Temps de simulation	Nombre de noeuds	Nombre de Cœurs	Performance (ns/h)	Temps CPU	Heures demandées
1.1	Protéine 1 + ligand 1 Condition 1	5	500 ns	5	120	5.0	100h	60000
1.1	Protéine 1 + ligand 2 Condition 1	3	500 ns	5	120	5.0	100	60000
1.2	Etc.							

Exemple pour le CT8

Projet	Système	Mode de calcul	Espace disque	Mémoire	Nombre de nœuds	Cœurs	Temps CPU	Heures dépensées
1.1	Molécule	Optimisation	Fonction d'onde WAVECAR 1 Gb	256 Gb	1	32	8 *runs* au total, 20h / *run*	5120
1.2	Molécule + substrat	Analyse vibrationnelle	Fonction d'onde WAVECAR 2 Gb	256 Gb	2	64	1 *run* 20h	1280
1.3	Molécule + substrat	Chemin réactionnel	8 fonctions d'onde WAVECAR 8x2 Gb	8 x 256 Gb	8	8x32	7 *runs* au total, 20h/*run*	35840

L'estimation des ressources : Elle doit, et cela est d'autant plus important si la demande est conséquente (au-delà d'un million d'heures), être basée sur des calculs menés sur un mésocentre ou sur un centre national pour des systèmes ou des tailles de problèmes identiques ou proches de ceux visés dans la demande présente de calcul pour s'appuyer sur des bancs d'essai. Si ce n'est pas le cas, il faudra préciser de combien cela s'écarte du

problème ciblé et expliquer comment le nombre d'heures de calcul a été établi via le nombre de travaux soumis en précisant le nombre de cœurs ou GPU requis pour chacun d'eux.
Les ressources de calcul qui seront attribuées devront de préférences être consommées de manière régulière afin de garantir une utilisation optimale des moyens nationaux. Aussi, il faudra préciser si l'utilisation des ressources attribuées sera régulière et/ou par périodes plus intensives et dans ce cas le justifier. Une planification sur l'année (grosse maille) de la consommation des ressources constitue un plus dans un dossier et peut apparaitre dans un tableau récapitulatif pour les gros projets.

1.5 Plans de gestion de données (DMP)

Seulement à titre d'information actuellement, il faudra décrire dans cette section le plan de gestion des données ou *Data Management Plan* (DMP) si disponible. Un exemple est disponible à cette adresse : https://anr.fr/fileadmin/documents/2019/ANR-modele-PGD.pdf. Cette rubrique sera prochainement obligatoire pour les projets nécessitants de fort volume de stockage ou un grand nombre de fichiers.

1.6 Bibliographie

Il s'agit de lister les articles publiés par l'équipe ainsi que les articles de bibliographie générale portant sur le même sujet en précisant bien si les références proviennent du laboratoire. Le dépôt de ces articles se fera sur eDARI en cliquant sur « Déposer les PDF des publications ».
Par exemple : Les références [2] sont celles de l'équipe. Si des ressources ont déjà été attribuées sur les moyens nationaux de GENCI, il faudra préciser les publications dont les résultats sont issus de ces calculs.

Exemple de Références

[1] W. Kohn. Nobel Lectures, Chemistry 1996-2000. World Scientic Publishing Co., Singapore, 2003.
[2] P. Nom. Titre de l'HDR. Document d'habilitation-Direction des recherches (2009).

2. Rapport d'activité

Le dépôt d'un rapport d'activité d'une page minimum est nécessaire pour tout type d'accès, si des ressources ont déjà été attribuées sur les moyens nationaux de GENCI afin de permettre aux Comités Thématiques et aux Directeurs de centres d'évaluer les avancées scientifiques et de justifier la demande de ressources additionnelles sur un calculateur d'un centre national.

Les différentes parties demandées sont les suivantes : Informations générales, résultats scientifiques, bibliographie ainsi que la présentation des résultats de recherche à des conférences.

2.1 Informations générales

Il faudra indiquer dans cette section les informations générales relatives au projet : Titre du projet, numéro de dossier, nombre d'heures accordées et consommées.

Dans le cas du renouvellement d'un dossier pour un

- **Accès régulier** : En cas de sous-consommation, merci d'en justifier la raison. L'absence de justification d'une sous consommation importante des ressources du projet à renouveler aura des répercussions sur le jugement du dossier (i.e. note scientifique et/ou technique).
- **Accès dynamique** : En cas d'extrême sous-consommation, merci d'en justifier la raison.

Afin de permettre à l'ensemble des utilisateurs de disposer des ressources attribuées, **il est recommandé de consommer les heures le plus régulièrement possible.**

2.2 Résultats scientifiques

Il faudra indiquer les **résultats scientifiques issus des calculs générés sur les supercalculateurs du parc de GENCI,** sur deux pages maximums par sous projet. Il est possible de préciser l'approche et les théories utilisées pour le projet.
L'avancement du projet scientifique devra être indiqué par rapport aux objectifs scientifiques annoncés dans la demande. Il est en particulier attendu de démontrer l'apport des calculs effectués dans l'obtention de résultats scientifiques. Les principaux résultats pourront être résumés avec le support d'une ou plusieurs illustrations.

2.3 Bibliographie

Il faudra indiquer les **articles en cours de préparation ou publication** ainsi que ceux **publiés** pour lesquels les résultats scientifiques obtenus ont été permis grâce à l'allocation de ressources sur les moyens nationaux de GENCI des années précédentes.
Ces articles peuvent être déposés sur eDARI en cliquant sur « Déposer les PDF des publications ».

2.4 Conférences et posters

Il faudra lister les conférences où ont été présenté les **résultats de recherche issus des calculs générés sur les moyens nationaux de GENCI,** en précisant s'il s'agit d'une présentation orale ou d'un poster ; ainsi que la participation à des *Challenges*.