	<p>MARCHE N° 2025-1154 (lot 1)</p> <p>Appel d'offres ouvert</p> <p>(En application des articles L2124-2 et R2124-2 du code de la commande publique)</p>
---	---

CAHIER DES CLAUSES TECHNIQUES PARTICULIÈRES

ACCORD-CADRE A BONS DE COMMANDE DE SERVICES

**Mise à disposition temporaire de personnels en ingénierie logicielle pour le Centre
Inria de l'université Grenoble-Alpes**

**Lot 1 – Mise à disposition temporaire de personnel en ingénierie logicielle pour le
projet DeepGreen**

CENTRE INRIA DE L'UNIVERSITÉ GRENOBLE ALPES

Inovallée, Avenue de l'Europe,
38334 Montbonnot Saint Martin

Sommaire

1	Généralités.....	3
2	Contexte de la prestation	3
2.1	Contexte général	3
2.2	Objectif en lien avec l'objet du marché	5
3	Objet du marché.....	6
4	Prestation à réaliser par le personnel mis à disposition	6
4.1	Description et objectifs de la prestation	6
4.1.1	Lieu et durée de la mission	6
4.1.2	Missions confiées	6
4.1.3	Activités principales.....	7
4.2	Compétences à mettre en œuvre dans le cadre de la prestation.....	7

1 Généralités

Le Centre de recherche Inria GRENOBLE RHONE-ALPES est un établissement public de recherche à caractère scientifique et technologique (EPST) sous la double tutelle des ministères en charge de la Recherche et de l'Industrie.

Créé en 1967, Inria a pour mission de produire une recherche d'excellence dans les champs informatiques et mathématiques des sciences du numérique, et de garantir l'impact de cette recherche auprès des acteurs économiques et sociétaux.

Cette recherche s'effectue au sein de 9 centres de recherche répartis dans toute la France (Paris, Rennes, Sophia Antipolis, Grenoble, Lyon, Nancy, Bordeaux, Lille et Saclay). Le siège social de l'institut est situé à Rocquencourt, près de Paris.

Le centre Inria Grenoble Rhône-Alpes compte une trentaine d'équipes de recherche ainsi que des services d'appui à la recherche. Le personnel du centre (750 personnes environ réparties sur cinq campus) est composé de scientifiques de différentes nationalités, d'Ingénieurs, de Techniciens et d'Administratifs.

2 Contexte de la prestation

2.1 Contexte général

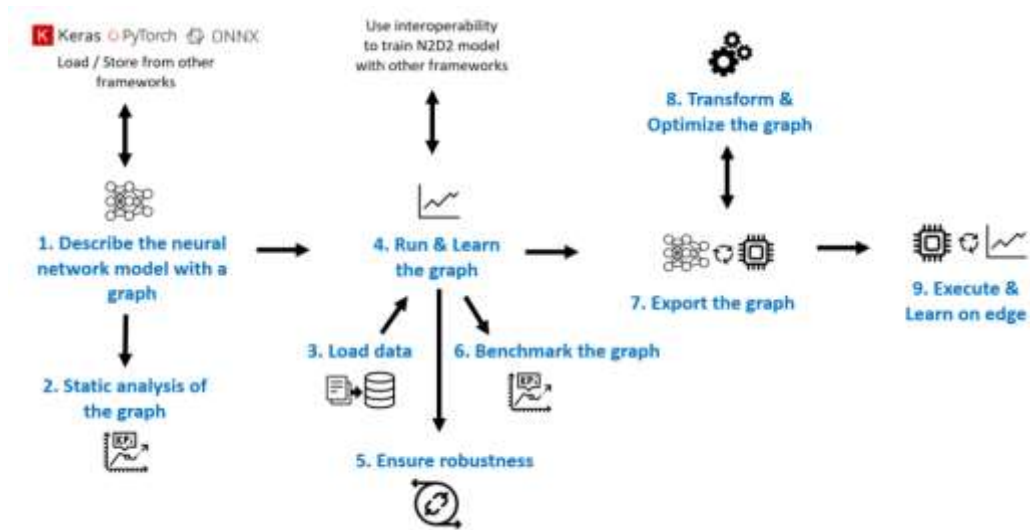
Le projet DeepGreen, piloté par le Commissariat à l'Energie Atomique (CEA) a été lancé pour la période 2024-2027. Ce projet réunit une vingtaine de partenaires de la recherche et de l'industrie, et a pour ambition de procéder à un développement de l'intelligence artificielle embarquée dans des systèmes et appareils intelligents.

L'objectif global du projet est de disposer, à terme, d'un outil performant permettant l'optimisation et le déploiement de l'intelligence artificielle embarquée, laquelle trouve des applications essentielles dans de nombreux domaines de l'industrie et de la société.

A cette fin, AIDGE, une plateforme logicielle de construction et de déploiement de Réseaux de Neurones Multi couches (DNN) est en développement à la suite d'un précédent outil N2D2 développé par le CEA. Cette nouvelle plateforme codéveloppée par les partenaires du projet DeepGreen, est conçue pour :

1. être interopérable avec les sources de description de DNNs communes telles que Keras, PyTorch, ONNX ;
2. être aisément extensible par le biais de composants utilisateurs développés en Python et/ou C++ ;
3. permettre différentes formes d'exécution et de génération de code pour des architecture variées ;
4. intégrer des techniques avancées d'optimisation de réseaux (quantisation, pruning, précision variable). La Figure 2-1 décrit les capacités de la plateforme AIDGE pour le développement de DNNs.

Figure Erreur ! Utilisez l'onglet Accueil pour appliquer Heading 1 au texte que vous souhaitez faire apparaître ici.-1 Flot de développement proposé par la plateforme AIDGE



La plateforme AIDGE est open-source et hébergée par l'Eclipse Foundation, dont le lien vers sa page d'accueil est le suivant : <https://eclipse-aidge.readthedocs.io/en/latest/index.html> .

L'architecture logicielle de AIDGE est composée d'une partie Core implémentée en C++ et fournissant les ponts vers Python pour l'interopérabilité avec les autres composants et pour une interface de programmation simplifiée pour les utilisateurs.

Les autres composants communément utilisés sont les composants d'exécution du graphe (Backend plugins), d'exportation du graphe (Export plugins), en particulier pour la compilation vers des systèmes embarqués. La Figure 2-2 résume à haut niveau l'architecture logicielle de AIDGE.

Figure Erreur ! Utilisez l'onglet Accueil pour appliquer Heading 1 au texte que vous souhaitez faire apparaître ici.-2

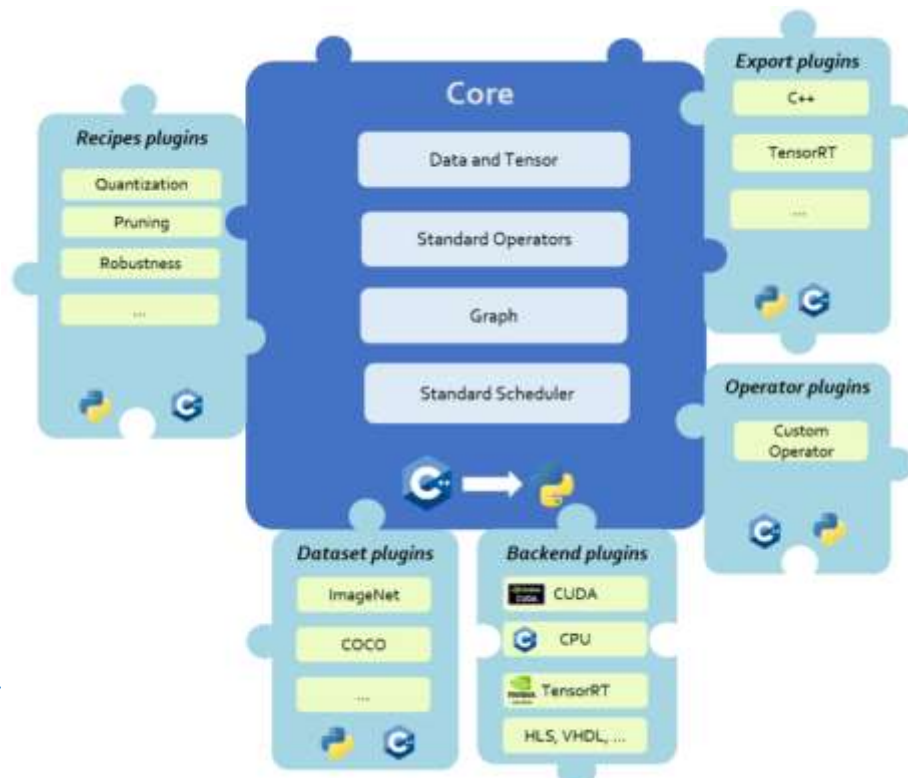


Figure Err

âtre ici.-3

2.2 Objectif en lien avec l'objet du marché

Un des axes du projet DeepGreen a pour but d'intégrer des technologies de compilations avancées pour les opérateurs présents dans les DNNs et de définir les langages domaine spécifiques permettant de décrire ces opérateurs. Ces technologies doivent être intégrés dans une plateforme de génération de code avancée pour un réseau complet et interopérables avec la plateforme AIDGE.

Cet axe de développement est porté par l'équipe de recherche CORSE (Centre Inria de l'Université Grenoble-Alpes), en collaboration avec le CEA pour le développement de la plateforme AIDGE pour la partie exportation et inférence.

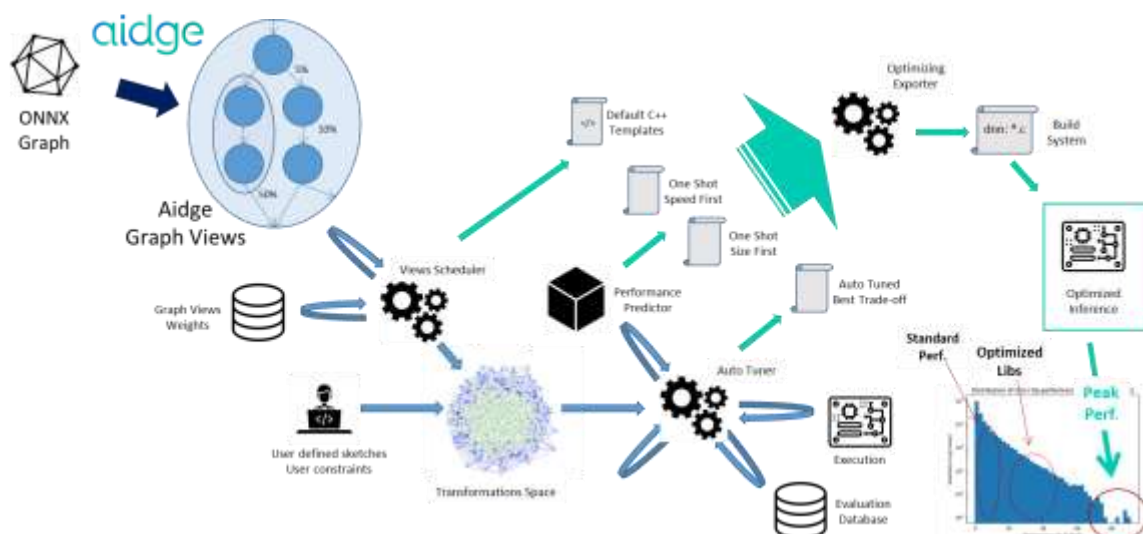
L'objectif de ces contributions est de générer un code pour les réseaux exportés dont la performance est proche du maximum théorique de la machine cible tout en respectant des contraintes éventuelles de taille de code ou de répartition en tâches. Aussi, afin de conserver les capacités d'extensibilité et d'interopérabilité de la plateforme AIDGE qui intégrera ces outils de compilation, l'utilisateur doit pouvoir aisément ajouter des contraintes ou forcer des choix de code génération.

Actuellement, le code généré par la plateforme AIDGE avec son module d'export par défaut (export C++) est non optimisé. Les outils de compilation seront intégrés dans un module d'export avancé pour la génération de code optimisé qui reste interopérable avec l'export C++. La Figure 2-3 donne un aperçu des composants qui contribuent à la chaîne de compilation proposée.

Les composants principaux sont :

1. un sélecteur de vues sur le graphe et ordonnanceur (View Scheduler) pour prioriser le temps de compilation alloué à chaque sous-graphe;
2. un optimiseur automatique (Auto Tuner) qui parcourt l'espace des transformations pour faire la sélection des choix les plus pertinents en fonction des objectifs fixés ;
3. un prédicteur de performance (Performance Predictor) qui permet d'assister l'optimiseur ou de générer immédiatement un code prédit efficace ;
4. un code générateur optimisant (Optimizing Exporter) supporté par des chaînes de compilation éprouvées capables de prendre la description des opérateurs, les sous graphes sélectionnés et les transformations choisies pour générer le code objet ou du code source optimisé.

Figure Erreur ! Utilisez l'onglet Accueil pour appliquer Heading 1 au texte que vous souhaitez faire apparaître ici.-4 Outils de compilation intégrés dans l'export AIDGE
View Scheduler, Auto-Tuner, Performance Predictor, Optimizing Exporter



3 Objet du marché

Afin de participer à l'axe de développement décrit à l'article 2.2, le présent marché a pour objet de poursuivre la mise à la disposition de l'équipe projet un personnel scientifique et technique, hautement qualifié dans le domaine des plateformes logicielles permettant le développement de programmes d'intelligence artificielle.

Le poste sur lequel le personnel sera affecté est intitulé Ingénieur développement et intégration pour technologies de compilation optimisantes dans un *framework* de code génération pour l'inférence.

Le Titulaire devra disposer de solides compétences et expériences dans le domaine de la mise à disposition de personnels scientifiques et techniques

Le personnel mis à disposition sera directement intégré à l'équipe projet au sein du Centre Inria de l'université Grenoble Alpes et assurera une assistance technique concourant directement à la réalisation du projet mené.

4 Prestation à réaliser par le personnel mis à disposition

4.1 Description et objectifs de la prestation

4.1.1 Lieu et durée de la mission

La prestation se déroulera principalement dans les locaux de Minatec Campus, antenne du Centre Inria de l'université Grenoble Alpes, située au 17 Rue des Martyrs (38054 GRENOBLE Cedex). Toutefois, le personnel mis à disposition sera régulièrement amené à effectuer des déplacements sur Paris Saclay (environ 2 jours par mois), lieu de résidence des équipes du CEA.

La durée de la mission est estimée à 2 ans, la durée quotidienne de travail du personnel mis à disposition correspondant à un temps plein. Les modalités concrètes de la mise à disposition sont précisées dans le CCAP.

4.1.2 Missions confiées

Intégré directement à l'équipe projet, et devant assurer un compte rendu de progression, le personnel mis à disposition se verra confiées des missions diverses :

- en architecture logicielle, le développement et l'intégration d'outils pour des transformations de code avancées telles que :
 - les transformations automatiques au niveau cœur (parallélisation, tuilage, packing/padding)
 - la fusion d'opérateurs
 - la génération de code pour opérateurs quantisés
 - l'auto-tuning avec boucle de feedback compilation->exécution
 - la gestion de la distribution sur machines à hiérarchie mémoire distribuée
- La collaboration sur la définition des langages domaine spécifique (DSL) pour l'interopérabilité avec les plateformes de développement pour l'IA, les outils de code génération et les outils de transformation cités plus haut :
 - DSL pour la définition des opérateurs algébriques
 - DSL pour les transformations de code et l'espace de recherche des optimisations
 - interfaces interactives pour l'exploitation de ces DSLs

Ces missions sont susceptibles d'évoluer au fil de l'avancement du projet DeepGreen.

4.1.3 Activités principales

En lien avec les missions confiées, telles que décrites à l'article précédent, le personnel mis à disposition sera principalement affecté aux activités suivantes :

- utilisation des plateformes de génération de code pour l'inférence à partir des *front-ends* classiques d'IA (*pytorch*, *tflite*) et de leurs backends (MLIR, TVM, LLVMIR)
- contributions à la plateforme Aidge
- contributions aux optimisations et langages de transformations développés dans l'équipe
- prototypage et intégration pour compilation optimisée end-to-end native/jit compilation et cross-plateforme compilation/exécution
- benchmarking sur des plateformes cibles telles que x86, ARM CPU ou GPU
- benchmarking sur plateformes embarquées type ARM-Cortex / STM32
-

4.2 Compétences à mettre en œuvre dans le cadre de la prestation

Le personnel mis à disposition aura un diplôme d'Ingénieur Bac + 5 ou équivalent, et présentera une expérience d'au moins 5 ans.

De par sa formation et son expérience, le personnel mis à disposition présentera *a minima* les compétences suivantes :

- architecture logicielle, architecture des ordinateurs
- outils de compilation et de cross-compilation
- développement sous environnement Linux
- C/C++/python/assembleur ARM

Par ailleurs, les compétences listées ci-après, susceptibles d'être utiles au bon déroulement du projet, sont fortement souhaitées :

- analyse/transformation de graphes de calculs
- connaissance des plateformes de développement pour l'IA
- développement pour plateformes embarquées