PLATEFORME DANTE

MULTI DATA ANALYSIS AND COMPUTING ENVIRONMENT FOR SCIENCE

SOMMAIRE

PREAMBULE

Dans le cadre de l'Agenda Numérique du MENESR, le comité de pilotage des infrastructures et services numériques (Copil InfraNum) a élaboré la feuille de route « Modernisation des infrastructures et services numériques des établissements de l'enseignement supérieur et de la recherche (ESR) » définissant les axes de rationalisation des infrastructures. Le projet DANTE s'inscrit pleinement dans cette feuille de route.

Il est construit sur un effort proactif des établissements concernés pour commencer à implémenter cette feuille de route « depuis la base » dans une stratégie « research driven ». Ce projet concerne principalement la fédération et *in fine* la mutualisation, dans un lieu d'hébergement commun, des infrastructures de calcul et d'analyse de données de l'IPGP (geosciences) et de l'APC (astrophysique et cosmologie), ainsi que la fédération autour d'un *instrument scientifique unique* des équipes et des expertises associées afin de répondre aux nouveaux enjeux scientifiques interdisciplinaires. *Ce projet entend enrichir et s'inscrire dans un Réseau Francilien en Sciences Informatiques.*

Les spécificités et les principaux axes forts du projet DANTE, détaillés dans ce document, sont :

Une offre commune et fédérée d'infrastructures et de compétences. Il entend fédérer les ressources et les services de calcul et d'analyse de données au sein d'une plateforme commune, et d'un pôle de compétences multidisciplinaires. Il entend ainsi construire un modèle organisationnel garantissant le « stewardship » de cet instrument scientifique, qui rendra *in fine* possible son hébergement et son fonctionnement dans un des futurs « Data Centres » labélisés par le ministère. Cet instrument scientifique fédéré (au travers de liens rapides) avec les grands centres nationaux de calcul, fournira un environnement et des services facilitant l'exploitation des données générées par les applications dimensionnantes de simulation, d'inversion et d'assimilation de données.

Une interdisciplinarité « research-driven ». Le projet permettra de construire autour d'enjeux scientifiques clairs, des méthodologies interdisciplinaires de calcul et d'analyse de données impliquant des domaines séparés jusqu'à récemment mais qui partagent des objets et des pratiques de recherche. Il constitue une première étape ambitieuse et garantie sa faisabilité. Il fournira le cadre permettant de s'ouvrir vers d'autres domaines plus éloignés thématiquement. Le projet, au travers des équipes de recherche et des thématiques associées, traverse aujourd'hui les frontières entre plusieurs instituts du CNRS (INSU, IN2P3, INP, INSIS). Il entend élargir cette dynamique au sein de l'USPC en étroite collaboration avec d'autres communautés, comme la médecine et la santé (Paris Descartes), la bioinformatique (Paris Diderot) et les sciences économiques et sociales (Science PO), en synergie avec d'autres efforts en Île de France.

Un pôle d'expertises et de connaissances multidisciplinaires. Le projet entend construire un « hub » d'expertises multi- et interdisciplinaires afin d'accélérer les projets numériques («machine shop»). Il fournira un environnement pour des « sprints » entre experts des domaines applicatifs, des sciences des données, de la recherche informatique et les « data providers » afin d'accélérer de bout et bout les diverses chaînes d'utilisation des données. Il constituera un « laboratoire numérique » pour l'émergence des nouvelles thématiques et méthodes interdisciplinaires d'inférence statistique afin d'exploiter les informations contenues les volumes et la grande diversité de données issues des systèmes d'observation et des simulations numériques.

Recherche et formation. Le projet est construit en forte synergie avec le Labex UnivEarths, au sein duquel de nouvelles approches interdisciplinaires se sont construites avec succès, et avec le projet d'Ecole Universitaire de Recherche porté par les établissements du Labex, liant ainsi cet instrument scientifique à la recherche et à la formation d'une nouvelle génération d'étudiants et de jeunes chercheurs à ces nouvelles pratiques de recherche associant calcul et analyse de données. Plusieurs activités du Labex (MOOCs, initiatives de science citoyenne, écoles et ateliers pour les professeurs

des lycées et collèges, etc.) bénéficieront et s'appuieront également sur cette plateforme et ce pôle de connaissance en calcul et en analyse de données.

Un lien privilégié avec le spatial. La simulation et l'observation des systèmes Terre-Planètes-Univers proches ou lointains, sont aujourd'hui, à tous les niveaux, des enjeux cruciaux et un défi auquel ce projet entend répondre en synergie avec le projet de Campus Spatial PRG. Les chercheurs et les équipes des établissements proposants participent à plusieurs missions spatiales importantes et grands instruments internationaux d'observation qui demandent un environnement collaboratif (« Concurrent Design Facility ») pour l'élaboration et le suivi de ces opérations, avec les services et les ressources de calcul et d'analyse de données : environnement de Virtualisation, conception et intégration logicielle des chaînes de traitement et d'analyse, simulation et moissonnage des données, etc.

Une ouverture claire vers l'Europe et l'international. Les communautés scientifiques associées au projet sont de longue date structurées au travers de la conception et de la construction de grandes infrastructures de recherche et systèmes d'observation à l'échelle européenne et internationale, et au sein des agences spatiales, qui demandent une envergure interdisciplinaire. Le projet en tant que plateforme multidisciplinaire de calcul et d'analyse de données et pôle d'expertises interdisciplinaires constituera un atout important pour les équipes franciliennes dans la construction rapide de réponses à ces appels d'offre et initiatives, dans un environnement très compétitif.

PRESENTATION DU PROJET ET DE SA FINALITE

Le projet DANTE entend construire une *plateforme et un environnement multidisciplinaire de calcul et d'analyse de données* et constituer dans un premier temps un pôle d'expertise et de connaissance interdisciplinaire en sciences de la Terre, des Planètes et de l'Univers (TPU). La stratégie du projet est « research-driven », et la finalité du projet est d'établir progressivement *in fine* de nouvelles synergies avec d'autres communautés scientifiques au sein de la ComUE USPC et de la région Île-de-France, comme médecine et santé (Paris Descartes), bioinformatique (Paris Diderot), sciences économiques et sociales (Science PO), autour de pratiques de recherche et d'expertise dans les domaines du calcul et de l'analyse de données.

Contexte scientifique du projet. Les sciences de la Terre, des Planètes et de l'Univers partagent une culture scientifique et des pratiques de recherche fondées sur l'observation (spatial, sol, mer) couvrant un large spectre d'échelles spatiales et temporelles et incluant : le développement et l'exploitation de grands instruments et de systèmes d'observation générant des données complexes et multi-types (évènements, séries temporelles, images); des méthodes innovantes de traitement et d'analyse de ces données, de simulation, et de modélisation afin d'en extraire de nouvelles connaissances.

La recherche dans ces domaines adresse des problèmes fondamentaux concernant la compréhension de la formation et de l'évolution des structures des systèmes Terre-Planètes-Univers, dans leur environnement, ainsi que des applications d'intérêt sociétal comme la prévision et la prévention des aléas volcaniques et sismiques, l'exploration et la gestion des ressources énergétiques, l'évaluation des changements environnementaux.

Les communautés TPU sont fortement intégrées au niveau national et international au travers de grands instruments, observatoires, systèmes d'observation, ainsi que des infrastructures distribuées d'archivage et de distribution des données associées. L'importance des données a conduit ces communautés à jouer un rôle pionnier dans la promotion et l'implémentation de l' « Open Data » avec des standards internationaux de données et de méta données, d'accès aux données, de formats d'échange de données et de métadonnées incluant les informations de provenance, et d'interopérabilité des données.

Le projet DANTE s'appuie sur la dynamique interdisciplinaire fructueuse amorcée par le LabeX UnivEarthS[1], associant les laboratoires d'Astroparticules & Cosmologie (APC), d'Astrophysique-Instrumentation-Modélisation (AIM) et l'Institut de Physique du Globe de Paris (IPGP). Cette dynamique a permis entre autres des avancées importantes autour de problèmes scientifiques où les méthodes d'analyse des données de type multi-messagers deviennent de plus en plus importantes : formation et dynamique des planètes ; signature gravitationnelle des tremblements de Terre ; utilisation des neutrinos pour l'imagerie de la Terre profonde et des muons pour la tomographie des volcans ; détection et reconstruction multi fréquence et multi longueur d'onde de sources radiatives complexes.

Un aspect de cette convergence réside dans l'étude de systèmes naturels complexes, multi-échelles, qui repose sur l'acquisition de longues séries temporelles, avec une métrologie extrême du temps et de l'espace et l'analyse statistique d'une grande diversité d'observables (évènements, séries temporelles, images). Un autre aspect réside dans l'importance croissante des approches multi-messagers pour l'analyse conjointe et croisée de différentes données issues d'expériences et de domaines différents. Enfin un dernier aspect réside dans des pratiques de recherche qui associent développement et suivi de grandes missions spatiales, de grands instruments, d'observatoires et de systèmes d'observation avec des méthodes innovantes de traitement et d'analyse, de simulation et de modélisation incluant des méthodes avancées d'inférence probabiliste et quantification des évènements extrêmes.

Le projet DANTE entend renforcer et élargir ces synergies entre pratiques de recherche et expertises scientifiques, méthodologiques et technologiques en sciences de la Terre, des Planètes et de l'Univers autour de grands enjeux scientifiques disciplinaires et interdisciplinaires.

Les nouveaux enjeux. Le taux de production, la complexité et la diversité des observations générées par ces systèmes d'observation et par les simulations numériques, défient nos capacités à les analyser, à les modéliser et à en inférer/extraire de nouvelles informations. Ces enjeux scientifiques et méthodologiques résultent de l'évolution rapide : des technologies d'acquisition ; de la résolution des simulations numériques (déterministes et stochastiques) multi-physiques et multi-échelles ; de nouvelles méthodes statistiques de traitement et d'analyse combinant en particulier des avancées récentes en « machine learning » et en inférence probabiliste; ainsi que de l'évolution rapide des technologies de calcul et de stockage.

Ces enjeux traversent les différentes disciplines des sciences de la Terre, des Planètes et de l'Univers dont les applications ont aujourd'hui franchi l'échelle du Péta et défient l'organisation et les modèles classiques de calcul, d'analyse, de stockage et de communication. Ils ne se résument pas à un problème de capacités informatiques mais requièrent de nouvelles approches intégrant :

- la diversité des chaînes complètes d'utilisation et d'analyse des données orchestrant capture, traitement, analyse, simulation numérique et modélisation ;
- des environnements qui permettent d'accélérer de bout en bout ces chaînes (data streaming « workflows ») et d'optimiser le flux de données entre leurs différentes phases de calcul et d'analyse ;
- une architecture permettant de fédérer différentes technologies de calcul et de stockage, différents modèles de programmation et d'exécution, et d'organiser les services pour l'instanciation des différentes phases des chaînes d'utilisation de données ;
- un espace collaboratif pour le développement et l'intégration logiciel ces chaînes de simulation, de traitement et d'analyse de données associés à la conception et le suivi de grands systèmes d'observation ;

[1] http://www.univearths.fr/

- des nouvelles technologies de Virtualisation permettant de façonner différents environnements applicatifs (logiciel, système) sous forme de machines virtuelles et conteneurs Linux, de les déployer et d'optimiser différents niveaux (matériel, logiciel) de fédération et de mutualisation, tout en respectant les politiques d'accès et d'utilisation des données et des ressources associées aux différents types d'analyse et pratiques de recherche.

La plateforme DANTE a pour but de:

- **fédérer et in fine mutualiser les ressources matérielles et logicielles du Service de Calcul Parallèle et d'Analyse de Données (S-CAPAD) de l'IPGP et du Centre François Arago (FACe) de l'APC au sein de l'IPGP,**
- **créer un centre innovant de type « Concurrent Design Facility » (CDF), au sein de l'APC, en particulier pour la conception et le suivi des opérations de missions spatiales en association avec le projet Campus Spatial PRG,**
- **mettre à niveau les capacités de l'ensemble de ces ressources et services en coévolution avec les besoins et les pratiques de recherche à l'IPGP et à l'APC.**

Le projet complet correspond à un budget d'investissement de l'ordre de 1,3 M€.

DANTE est un instrument scientifique et un pôle d'expertise multidisciplinaire innovant. Il stimulera des approches multi- et interdisciplinaires autours de grands défis scientifiques en TPU en fournissant :

- un *laboratoire numérique* pour « observer » (explorer, analyser et modéliser) les volumes et la diversité des données issues des systèmes d'observation et des simulations numériques, ainsi que concevoir et exploiter des missions spatiales et des grands instruments d'observation;
- un pôle d'expertise, de logiciels et de connaissances, favorisant un partage multi- et interdisciplinaire pour accélérer les projets numériques (« machine shop ») autour de grandes questions scientifiques ;
- un lieu de « sprints » entre domaines applicatifs, experts en sciences des données, recherche informatique et « data providers » afin d'accélérer de bout en bout les diverses chaînes d'utilisation des données.

Il intégrera :

- une architecture « data-intensive » et un ensemble de ressources et de services organisés pour accélérer de bout en bout les chaînes d'utilisation de données depuis l'acquisition, la capture, le traitement, l'analyse et la modélisation jusqu'à la production de connaissance pour la recherche et l'aide à la décision ;
- les nouvelles technologies de virtualisation pour permettre un meilleur contrôle par l'utilisateur de la gestion d'applications complexes et de leur environnement applicatif, et de leur instanciation sur diverses plateformes (HPC, Cloud) ;
- un environnement pour l'amélioration et l'implémentation de procédures de qualité de logiciel entre ces différentes disciplines.

Il permettra aux chercheurs de l'IPGP et de l'APC, et plus largement aux chercheurs de l'USPC et de l'Île-de-France :

- d'explorer de nouvelles pratiques de recherche collaborative pour le développement de méthodes innovantes d'analyse et d'inférence statistique, de simulation et d'assimilation de grandes masses de données ;
- de pouvoir stocker des jeux de données à l'échelle du Po pour les analyser et les modéliser sur des périodes de plusieurs mois à plusieurs années en disposant d'une plateforme de calcul et d'analyse de données adaptée ;

- de disposer d'un ensemble d'outils logiciels et de services organisés de manière agile et flexible pour l'exploitation et la valorisation de ces données par des équipes de recherche distribuées.

L'environnement de recherche et de production de DANTE entend:

- répondre et accompagner les changements de pratique de la recherche associant aujourd'hui calcul et analyse statistique des flux de données dont le volume et la diversité ne cessent de croitre ;
- lever un certain nombre de verrous technologiques et méthodologiques pour une transformation des méthodes, et une utilisation des nouvelles technologies « data-intensive » afin de répondre aux nouveaux enjeux en TPU ;
- fournir un « proof-of-concept » scientifique, technologique et organisationnel (« stewardship ») fédéré avec les centres distribués d'archivages des grands systèmes d'observation nationaux et internationaux en TPU.

Expertise et visibilité du projet. Le projet DANTE s'appuie sur une longue expertise de l'IPGP et de l'APC dans les domaines du calcul et de l'analyse de données, des technologies de type Grille et Cloud, ainsi que de la conception et du suivi de missions spatiales, de grands instruments et d'observatoires multi-capteurs. Les différentes composantes du projet DANTE participent en particulier à la conception et à l'exploitation :

- de grandes missions spatiales en liaison avec le CNES, l'ESA, la NASA, ainsi que les agences japonaises et chinoises (e.g. LISA, Euclid, SVOM, InSight, SWARM, GAIA, COPERNICUS) ;
- de grands instruments et observatoires internationaux inclus dans les feuilles de route nationales (SNRI) et européennes (ESFRI) (e.g. EPOS, CTA, LSST, KM3Net) où le CNRS et le CEA jouent un rôle important.

Elles portent également ou contribuent à de nombreux projets ERC (e.g. SM-GRAV, WHISPER, WAVETOMO, SLIDEQUAKES, TRANSATLANTIC, PRISTINE, SIREAL, EDIFICE).

En retour, ces projets assureront une visibilité internationale au projet DANTE et à la région Île-de-France dans le domaine du calcul et de l'analyse des données.

PORTEUR DU PROJET

Le projet DANTE est porté conjointement par l'Institut de Physique du Globe de Paris (IPGP), qui assurera la gestion du projet, et le laboratoire Astroparticules & Cosmologie (APC), au sein de la ComUE Université Sorbonne-Paris Cité (USPC) qui soutient le projet.

L'IPGP[2] est un grand établissement d'enseignement supérieur et de recherche partenaire de l'université Paris-Diderot et de l'université de la Réunion. L'IPGP est une UMR unique du CNRS-INSU composé de plus de 500 personnes (un tiers d'enseignants-chercheurs, un tiers d'ITA et IATOS, un tiers de doctorants et de post-doctorants). À ses missions de création[3] et de transmission[4] du savoir dans les champs des géosciences (Terre, Planètes, Environnement) s'ajoute une mission d'observation et de surveillance des phénomènes naturels, au sein de ses observatoires nationaux[5] dont l'IPGP a la responsabilité en sismologie globale, magnétisme, volcanologie et études spatiales.

[2] http://www.ipgp.fr

[3] http://www.ipgp.fr/en/research

[4] http://www.ipgp.fr/en/education

[5] http://www.ipgp.fr/fr/observation

L'activité de recherche est structurée autour de grands programmes[6] pluridisciplinaires, et 17 équipes de recherche où géologie, géophysique, physique, chimie, mathématiques, biologie participent à l'analyse et à la compréhension des systèmes Terre-Planètes dans leur environnement. Les chercheurs et ingénieurs de l'IPGP préparent également les observatoires de demain, en installant des observatoires permanents au fond des océans et en participant à l'élaboration et au suivi de missions spatiales, en collaboration avec l'ESA et le CNES.

La recherche à l'IPGP est principalement menée en métropole sur l'îlot Cuvier, les campus de Paris Rive Gauche et de Saint-Maur, et en outremer (Antilles et Réunion) dans les observatoires volcaniques, sismologiques et magnétiques.

L'Institut joue un rôle majeur en sciences de la Terre, aux niveaux national et international, en particulier dans : le développement de méthodes innovantes pour l'exploration et l'analyse statistique des données générées par les systèmes d'observation et de surveillance ; la modélisation et la simulation numérique de ces systèmes naturels à toutes les échelles ; l'inversion et l'imagerie haute-résolution ; la détection et la reconstruction des sources radiatives complexes; ainsi que l'assimilation statistique de données.

L'IPGP a également joué un rôle pionnier dans le développement du calcul intensif en sciences de la Terre. Depuis plusieurs décennies, l'IPGP a démontré une expertise et une capacité à concevoir, déployer et opérer des infrastructures de calcul intensif et d'analyse de données, au travers du Centre National de Calcul Parallèle en Sciences de la Terre et plus récemment du Service de Calcul Parallèle et d'Analyse de Données[7](S-CAPAD).

S-CAPAD est aujourd'hui une plateforme de calcul et d'analyse de données fédérée avec le Centre de Données de l'IPGP, et partagée par les 17 équipes et programmes de l'IPGP (près de 150 chercheurs) et leurs collaborateurs. Cet instrument scientifique commun, qui bénéficie du service informatique de l'IPGP, intègre un ensemble de ressources pour la simulation numérique de systèmes multi-échelles et multi-physiques, et le développement d'applications innovantes dans le domaine du traitement, de l'analyse statistique et de la modélisation (inversion, assimilation) de grandes masses de données.

La configuration actuelle est de 1600 cœurs CPU pour le calcul intensif, auxquels s'ajoutent 28 nœuds spécifiques (mémoire, GPU, SSD) pour l'analyse intensive de données, avec un système de fichiers parallèles de 700 To effectifs, le tout connecté par un réseau Infiniband, et un système de stockage *persistant* de 250 To effectifs. S-CAPAD est un nœud de la plateforme CIRRUS[8] de l'USPC et a développé un ensemble de formations ouvertes aux chercheurs et aux étudiants pour l'utilisation de ces ressources.

L'APC[9]**,** créé en 2005 a pour objectif de comprendre l'origine, les structures et l'évolution de l'univers à l'interface entre l'étude de l'infiniment grand et de l'infiniment petit, entre physique des particules et astrophysique. L'APC est associé à l'Université Paris Diderot, au CNRS (IN2P3, INSU, INP), au CEA (DSM/IRFU) et à l'Observatoire de Paris. L'APC est composé de plus de 200 personnes (75 enseignants-chercheurs, 60 ITA/IATOS, 70 doctorants et post-doctorants) et rassemble expérimentateurs, observateurs et théoriciens dans ces domaines.

---

6 http://www.ipgp.fr/fr/programmes

7 http://webpublix.ipgp.fr/rech/scp/

8 http://cirrus.uspc.fr

9 http://www.apc.univ-paris7.fr/APC_CS/fr/labo

L'activité de recherche à l'APC est structuré autour de quatre grandes thématiques à forte composante expérimentale et observationnelle (cosmologie, ondes gravitationnelles, astrophysique des hautes énergies, et neutrinos) et une thématique transversale qui porte sur l'ensemble des aspects théoriques de la cosmologie et de la physique des astroparticules. La simulation numérique constitue également un outil complémentaire pour l'analyse des phénomènes complexes.

L'attention particulière que reçoit le traitement des données spatiales à l'APC a conduit à l'ouverture en 2010 du centre François Arago[10] (FACe) étroitement connecté au centre de calcul du CNRS de Lyon (CC-IN2P3[11]), afin de soutenir les missions spatiales et les grands instruments nécessitant des traitements complexes de données. Il fournit un certain nombre de services à ses utilisateurs, tels que : l'accès à des calculateurs ; des routines d'analyse ; des serveurs de stockage ; la distribution de nouveaux logiciels d'analyse ; et surtout un CDF avec une équipe experte en informatique et science des données et le soutien du service informatique de l'APC. La puissance de calcul et le volume de stockage fournis par le FACe représentent : 652 cœurs et 42 To de stockage pour le cluster Arago et près de 80 serveurs virtuels pour les clusters de machines virtuelles dédiés aux besoins des projets spatiaux et astroparticules (Euclid, LPF, Lisa, CTA…). L'ensemble de ces services s'appuie sur un volume de stockage total de près de 180 To. Dans le cadre des différents projets instrumentaux et mission spatiales, environ 10 ETP contribuent au développement et à l'expertise dans ces domaines.

Les laboratoires APC et IPGP ont enfin développé depuis 2010 des collaborations interdisciplinaires au sein du LabeX UnivEarths[12], dont les buts principaux sont la compréhension de la formation et de l'organisation des systèmes Terre-Planètes-Univers dans leur environnement, et des crises cataclysmiques amenant des changements radicaux dans leurs évolution. Parmi les axes de cette collaboration ont peut citer : la muographie, les géoneutrinos et neutrinos cosmiques, les détecteurs gravitationnels en sismométrie, et les méthodes de détection des signaux associés à des sources radiatives complexes et leur restauration. Au niveau de la formation, l'APC et l'IPGP portent également en commun, au sein de Paris Diderot, le projet d'Ecole Universitaire de Recherche Terre-Planète-Univers dans le cadre de la ComUE USPC.

CONTEXTE NATIONAL ET INTERNATIONAL

La nature de la recherche en sciences de la Terre, des Planètes et de l'Univers, et dans de d'autres disciplines, change. Les nouvelles découvertes et innovations dépendent de la capacité à :

a) extraire et inférer dans un cadre probabiliste de nouvelles informations grâce à des méthodes innovantes de traitement, d'analyse et de modélisation statistiques de la masse et de la diversité des données générées aujourd'hui par ces systèmes d'observation, et les simulations numériques (instruments virtuels) de ces système naturels multi-échelle et multi-physique.
b) concevoir et calibrer de nouvelles missions spatiales, des grands instruments et systèmes d'observation de plus en plus complexe, et des grands observatoires intégrant multi-capteurs et différents systèmes d'acquisition ;

Ainsi au niveau national, le projet DANTE s'inscrit pleinement dans la feuille de route nationale du numérique. Il est construit sur un effort des établissements partenaires pour commencer à implémenter cette feuille de route « depuis la base » dans une stratégie « research driven ». Il entend contribuer également au projet d'Institut des Sciences des Données porté par l'USPC, dont le but est de créer de nouvelles synergies entre les expertises existantes au sein de l'USPC dans les domaines de la recherche informatique, des sciences des données, de l'intelligence artificielle, des

[10] http://www.apc.univ-paris7.fr/FACe/

[11] http://cc.in2p3.fr

[12] http://www.univearths.fr/fr/accueil-2/

mathématiques, de la physique des milieux complexes, des sciences des matériaux, de la Terre et de l'Univers, de l'environnement, de la linguistique, de la médecine et de la santé, et des sciences humaines et sociales. *Le projet DANTE entend ainsi enrichir et s'inscrire dans un Réseau Francilien en Sciences Informatiques*

L'écosystème des infrastructures numériques en Île-de-France ne répond pas complètement aujourd'hui aux besoins des applications dans les domaines de l'analyse et de la modélisation de grandes masses de données issues des systèmes d'observation et de la simulation numérique. Ces grands centres de calcul (Tiers 1 et 2) sont aujourd'hui indispensables pour le passage à l'échelle de certaines applications du projet DANTE, comme les applications « dimensionnantes » de simulation numérique en cosmologie, sismologie et magnétohydrodynamique, et d'inversion/assimilation de gros volumes de données. Cependant ils restent encore aujourd'hui construits comme des infrastructures généralistes privilégiant les cycles CPU et sont peu adaptés aux chaînes « data-intensive » d'utilisation des données qui requièrent en particulier : un stockage (cache) des données à l'échelle du péta et *persistant* sur des durées d'utilisation de plusieurs mois à plusieurs années ; le transfert et l'agrégation de données issues des centres d'archivages distribués ; des environnement de Virtualisation, et des politiques d'accès et d'identification souples, via des systèmes de certificats compatibles entre les diverses infrastructures de calcul et de données.

Les plateformes publiques de type Cloud ne sont pas non plus une solution satisfaisante, du moins à l'heure actuelle, en raison de leur modèle économique (coûts d'accès et de mouvement de données), et de leur politique d'accès et de propriété peu adaptés au cycle de vie de ces données et au développement et à l'adaptation de méthodes dans le cadre de projet collaboratif. Les plateformes académiques de type Cloud évoluent rapidement avec des capacités et des services intéressants (FG-cloud, et EI Fed cloud). L'écosystème de ces infrastructures et leur modèle économique restent encore aujourd'hui très incertains et extrêmement mouvant en Europe (EOSC) et à l'international, ce qui ne permet pas aujourd'hui aux communautés scientifiques engagées dans de grands projets nationaux et internationaux de se projeter facilement dans le temps.

Par ailleurs, la stratégie du projet DANTE est celle d'une ouverture claire vers l'Europe et l'international. Les communautés scientifiques associées (APC, IPGP) au projet sont de longue date structurées au travers de grandes infrastructures de recherche et de systèmes d'observation à l'échelle européenne et internationale, et au sein des agences spatiales, qui demandent une envergure interdisciplinaire. Le projet en tant que plateforme multidisciplinaire de calcul et d'analyse de données et pôle d'expertises interdisciplinaires constituera un atout important pour les équipes franciliennes dans la construction rapide de réponses à ces appels d'offre et initiatives, dans un environnement très compétitif. A titre d'exemple, plusieurs grands projets sont brièvement présentés ci-dessous pour donner une idée des compétences et des opportunités en jeu dans la future plateforme.

LISA/LISAPATHFINDER[13]. LISA (Laser Interferometer Space Antenna) est une mission spatiale de l'ESA, à l'horizon 2030, dont l'objectif est de détecter en continu les ondes gravitationnelles avec des implications fondamentales pour la physique et l'astronomie. La pertinence de la « Concurrent Design Facility » a été démontrée par l'APC pour le design préliminaire et les exercices préparatoires de la mission LISAPathfinder, démonstrateur technologique de LISA lancé en décembre 2015, ainsi que pour l'analyse des données de cette mission. L'APC a aujourd'hui la responsabilité du « Data Processing Center » (DPC) du consortium LISA, qui s'appuiera sur la plateforme DANTE pour fournir les outils de développement et d'intégration logiciel du futur pipeline de la mission. L'analyse des données utilisera les ressources de calcul et d'analyse de la plateforme et un environnement hybride

[13] http://sci.esa.int/lisa/

cluster/Cloud afin de gérer les pics d'activité et fournir l'environnement pour les Mock LISA Data Challenge (MLDC) (entrainement des chaînes d'utilisation des données).

EUCLID[14]. Euclid est une mission de l'ESA dont l'objectif est de déterminer la nature de l « Energie noire » source hypothétique de l'accélération de l'expansion de l'Univers. Le projet EUCLID s'appuiera sur un Science Data Center (SDC) distribué sur 8 pays, et intégrés à la plateforme CODEEN (COmmon DEvelopment ENvironment), créée par les ingénieurs de l'APC en collaboration avec différentes agences européennes (CNES, ISDC, ROE). CODEEN permet le développement collaboratif des éléments d'une chaîne de traitement en facilitant la production et l'intégration des logiciels d'analyse des données et s'appuiera sur les ressources et les services de plateforme DANTE.

INTEGRAL[15]**/SVOM**[16]**/ATHENA**[17]**.** Dans le domaine des hautes énergies, la mission INTEGRAL (INTErnational Gamma-Ray Astrophysics Laboratory) de l'ESA vise à étudier le rayonnement le plus énergétique qui provient de l'espace. Une partie de l'analyse des données d'INTEGRAL a été réalisé sur le cluster APC et sur les Clouds fédérés par France-Grilles. La mission spatiale Franco-chinoise SVOM qui est la suite d'INTEGRAL, utilisera également les technologies de conteneurs et de Cloud dans l'environnement de la plateforme DANTE en réutilisant le pipeline d'analyse de données de la précédente mission. SVOM est une mission qui aura un potentiel exceptionnel pour des analyses multi-messager (photons, neutrinos, ondes gravitationnelles) des structures cosmiques.

SWARM[18]. La mission Swarm est la 5ème mission du programme Earth Explorer de l'ESA lancée en 2013. Les trois satellites en orbite polaire basse seront opérationnels jusqu'en 2024 au moins. La mission a pour objectif l'étude de l'ensemble des composantes du champ magnétique terrestre ainsi que de l'environnement ionosphérique. Les données sont également exploitées pour des applications sociétales, comme la météorologie spatiale. L'IPGP est responsable des magnétomètres absolus embarqués. Les moyens de la plateforme S-CAPAD (IPGP) sont déjà largement utilisés pour produire des modèles de champ, manipulant simultanément l'ensemble des données produites par la mission, et des méthodes avancées d'analyse. Les données s'accumulant, et la possibilité d'exploiter de manière plus systématique les données synchrones des observatoires magnétiques au sol (plus d'une centaine, produisant chacun autant de données que les trois satellites), ainsi que la découverte récente de nouveaux signaux produits par les éclairs atmosphériques, nécessite de franchir un nouveau palier en termes de capacités et de méthodes d'analyse. La plateforme DANTE permettra d'avoir accès aux ressources et aux services de calcul et d'analyse pour ces volumes de données, ainsi qu'à l'expertise nécessaire à la réalisation de ce projet ambitieux.

InSight[19]. Sous la responsabilité du Jet Propulsion Laboratory (JPL), InSight (Interior Exploration using Seismic Investigations, Geodesy and Heat Transport) est une mission ambitieuse de la NASA à l'horizon 2018. Véritable observatoire géophysique (activité sismique, flux de chaleur, champ magnétique de surface, rotation de la planète), elle a pour objectif d'étudier la structure interne de Mars afin de mieux comprendre sa formation et son évolution unique. L'IPGP a la responsabilité du

[14] http://sci.esa.int/euclid/

[15] http://sci.esa.int/integral/

[16] https://svom.cnes.fr

[17] http://sci.esa.int/cosmic-vision/54517-athena/

[18] http://www.esa.int/Our_Activities/Observing_the_Earth/Swarm

[19] https://insight.jpl.nasa.gov/home.cfm

développement de l'instrument SEIS (Seismic Experiment for Interior Structures), en collaboration avec le CNES et Paris Diderot-Sorbonne Paris Cité, qui sera déployé au sol et enregistrera pendant au moins deux années l'activité sismique, les vibrations générées par les impacts de météorites et le phénomène de marées (accompagnant le satellite Phobos). L'IPGP a également la responsabilité de l'analyse et de l'exploitation des données sismologiques qui s'appuiera sur les ressources et les services de la plateforme DANTE.

LSST[20]. LSST est un télescope au sol dont la première lumière est à l'horizon 2020. Il sera pour plusieurs années le relevé astronomique le plus volumineux jamais fait (100 Po in fine), et 50% du traitement de ces données sera fait en France (CC-IN2P3). Très complémentaire à Euclid pour la cosmologie, les données des deux instruments devraient être combinées lors de la mise à disposition des catalogues des projets. L'expertise reconnue de l'APC sur l'utilisation et l'administration des plateformes d'intégration continue partagées par les trois projets que sont LISA, Euclid et LSST, lui donne un leadership dans les discussions et la plateforme de calcul et d'analyse DANTE renforcera celui-ci.

CTA[21]. CTA est un observatoire global de rayons gamma qui réunit plus de 100 laboratoires provenant de tous les continents. Il offre plusieurs services, comme l'accès aux données scientifiques intégrées dans l'Observatoire virtuel (VO), le support utilisateur, une plate-forme de gestion des propositions pour les observations scientifiques (Proposal Handling Platform -PHP-), une passerelle web centralisée (« Science Gateway ») avec un système centralisé d'authentification et d'autorisation (A & A) pour toutes les applications CTA. Dans ce contexte, l'équipe APC s'est engagée à fournir au Consortium la plate-forme de gestion des demandes d'observations scientifiques, intégrée au Data Model Global de l'observatoire CTA. Cette interface et les fonctionnalités d'intégration de différents types d'observations, enrichira la plateforme DANTE et intéresse d'autres projets spatiaux (SVOM).

EPOS-IP[22]. Le projet ESFRI « European Plate Boundary Observatory System », est aujourd'hui en phase d'implémentation (2015-2019). Il a pour but de fédérer au sein d'une infrastructure pan-européenne de recherche, dotée d'une gouvernance, l'ensemble des systèmes d'observation en Terre solide, d'offrir un ensemble de services thématiques et transdisciplinaires pour l'accès à ces données, leur analyse et leur modélisation, intégrés avec des plateformes de calcul et d'analyse de données. La contribution française est constituée par l'Infrastructure de Recherche Nationales RESIF[23], auquel le Centre de données des observatoires de l'IPGP et la plateforme DANTE sont une contribution importante. L'expertise acquise dans ce cadre par l'IPGP, avec le projet FP-Infrastructure VERCE[24] « Virtual Earthquake and Seismology Research Community e-science Environment », dans le développement d'architecture et de plateforme orientée services, intégrant infrastructures de données, HPC, Grilles et Cloud, est un atout important.

Un atelier interdisciplinaire réunissant les agences de financement des communautés Astroparticule (e.g. « Astroparticle Physics European Consortium » APPEC) et Géosciences (GEO-8) est organisé par les établissements porteurs de ce projet à la rentrée 2017. Un des attendus est une feuille de route

[20] https://www.lsst.org

[21] https://www.cta-observatory.org

[22] https://www.epos-ip.org

[23] http://www.resif.fr/?lang=fr

[24] http://www.verce.eu/index.php

pour la fédération au niveau européen des infrastructures de données, et de infrastructures de calcul et d'analyse de données dans ces communautés, pour l'accompagnement des grandes infrastructures de recherche et systèmes d'observation inscrites sur la feuille de route Européenne (ESFRI) par exemple EPOS, CTA, KM3Net etc.

Les établissements partenaires du projet ont une longue histoire de relations d'échange avec des acteurs internationaux comme les agences spatiales (NASA, ESA, CNES), et de grands Instituts et Universités, e.g. Earthquake Research Institute et KAvili Institute for Physics and Mathematics of the Universe de l'Université de Tokyo, l'Université de Californie Berkeley, le MIT, Université du Chili (Santiago), etc. En particulier l'APC, suivant une incitation de l'IN2P3, construit un projet d'Unité Mixte Internationale entre Paris intra-muros et Berkeley, dont une composante importante est le Berkeley Institute of Data Science, présidé par le prix Nobel Saul Perlmutter, et dans laquelle la plateforme DANTE aura une contribution importante.

La plateforme DANTE entend également développer des synergies avec l'initiative du European Open Science Cloud[25] (EOSC), impulsée par la commission européenne, et à laquelle les équipes de l'APC et de l'IPGP contribuent déjà en liaison avec le CNRS, le CEA et le CNES, et en particulier contribuer dans la suite du projet EOSCpilot[26]. Plusieurs appels seront lancés en 2018-2020 dans le cadre de la phase finale de H2020 et en vue de la préparation de l'appel Européen suivant. La plateforme DANTE constituera un atout important pour les proposants franciliens.

ACTIVITES SCIENTIFIQUES ET COMMUNAUTES UTILISATEURS

Il existe aujourd'hui une masse critique de chercheurs et d'ingénieurs à IPGP et à qui partagent une culture scientifique et des pratiques de recherche communes, et qui dépendent de la nouvelle plateforme multidisciplinaire du projet DANTE.

Activités scientifiques. Les activités scientifiques à l'IPGP et de l'APC partagent des pratiques de recherche communes combinant le développement et le suivi de misions spatiales, de grands instruments et systèmes d'observation, et d'observatoires à l'échelle nationale et internationale, avec des développements méthodologiques dans les domaines de :

- la modélisation physique et numérique (simulation, assimilation de données) de l'histoire, des structures et de l'évolution des systèmes Terre-Planètes-Univers, et des interactions avec leur environnement (milieux stellaires, enveloppes fluides), ainsi que des phénomènes transitoires qui modulent ces structures, dans un cadre statistique (caractérisation des incertitudes et des évènements extrêmes) ;
- le traitement et l'analyse statistique complexe, en particulier dans un cadre bayésien, de grands jeux d'observation (évènements, séries temporelles, images) issus des systèmes d'observation, intégrant la complexité et la diversité des observables (multi-messagers) ;
- la détection et l'analyse statistique de rayonnements (gravitationnel, sismique et électromagnétique, photons, X, gamma, particules haute énergie) associés aux phénomènes transitoires qui modulent l'évolution et la formation des structures dans les systèmes Terre-Planètes-Univers, ainsi que la caractérisation et la reconstruction multi-échelles des diverses sources associées, internes ou aux interfaces externes (milieux stellaires, enveloppes fluides, processus de surface) ;
- l'imagerie haute-résolution, dans le cadre des méthodes d'inversion et d'inférence probabiliste, de l'intérieur de la Terre et des Planètes, de leurs enveloppes fluides externes, et des milieux

[25] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[26] http://libereurope.eu/blog/2017/02/23/eoscpilot-european-open-science-cloud-research-pilot/

stellaires avec des applications fondamentales et sociétales comme l'exploration et l'exploitation de ressources énergétiques, et la surveillance des risques naturels ;
- La quantification des incertitudes directes et inverses, et des évènements extrêmes dans un cadre probabiliste ;
- la surveillance et l'évaluation, dans un cadre physique et probabiliste, des aléas et risques naturels (séismes, volcans, glissements de terrain, tsunami).

Ces activités se déclinent au travers de :

Astrophysique haute énergie *:* compréhension et modélisation des phénomènes violents dans l'univers (processus d'accrétion et d'éjection, explosions stellaires) qui modulent la formation des structures cosmique, des rayonnements associés (photons, neutrinos, ondes gravitationnelles) et des phénomènes physique d'accélération et de propagation des particules jusqu'aux énergies relativistes de l'Univers.

Cosmologie : compréhension de l'histoire et de la structure de l'Univers, de la température, de l'anisotropie et des fluctuations du fond diffus cosmologique diffus (CMB) source d'information sur l'univers et le champ magnétique primordial, de la matière et de l'énergie noire ainsi que leur impact sur la formation des structures et les possibles modification des théories de gravité.

Ondes gravitationnelles : compréhension des différentes sources et de la propagation des ondes gravitationnelles, couplée à des méthodes statistique détection de ces signaux et de reconstruction des sources, ouvrant de nouvelles informations sur les objets astrophysiques massifs et denses comme les trous noirs, les étoiles à neutrons, les étoiles naines, et sur les premiers âges de l'Univers, difficilement accessibles à l'astronomie conventionnelle (rayonnement électromagnétique).

Cosmochimie, astrophysique et géophysique expérimentale **:** compréhension des processus de formation du système Solaire et de son évolution précoce ; de la formation de la Terre et des Planètes, de leur différentiation et de leur histoire.

Planétologie, Gravimétrie, Géodésie et sciences spatiales **:** compréhension des processus de couplage Terre-Océan-Atmosphère depuis les perturbations atmosphériques transitoires associées aux séismes et tsunamis, jusqu'aux interactions long terme entre dynamique interne et atmosphère des planètes ; de la structure et de la dynamique internes de la Terre et des Planètes en combinant sismologie, gravimétrie et géodésie spatiale, et des mesures complémentaires au sol.

Géomagnétisme **:** compréhension de la formation, de la structure, et de l'évolution du champ magnétique principal et secondaires (lithosphérique, ionosphérique et magnétosphérique) ainsi que des processus magnétohydrodynamiques (dynamo) associés, avec des applications sociétales (recherche ressources minérales, systèmes de navigation et de communication, météorologie spatiale).

Dynamique des Fluides Géologiques**:** compréhension à partir des outils de la thermomécanique des fluides réactifs de : la dynamique de la Terre (manteau, noyau) et des Planètes du système Solaire et du couplage entre leurs différentes enveloppes internes et externes ; de la dynamique des éruptions volcaniques ; et des processus géomorphologiques à la surface de la Terre, avec des applications sociétales (risques naturels) et environnementales (changement climatique).

Sismologie **:** compréhension, et reconstruction à partir de l'analyse statistique des signaux sismiques cohérents et non cohérents (bruit sismique) de : l'intérieur de la Terre et des planètes telluriques (structures, propriétés, dynamique) dans leur environnement ; des sources sismiques internes (tectoniques, volcaniques, induites) et externes (enveloppes fluides externes et processus de surface) ; de la propagation du rayonnement associé (sismique, acoustique, infrason, gravitationnel) et de leur processus de couplage ; avec des applications sociétales (risques naturels), industrielles (ressources énergétiques) et environnementales (changement climatique).

Géosciences marines **:** compréhension de la formation et de l'évolution des structures et de la composition de la lithosphère océaniques, en particulier aux niveaux des rides et des zones de subduction actives, avec des applications sociétales (risque sismiques, volcanique, tsunamis) et économiques (exploration et suivi de production de nouvelles ressources énergétiques).

Tectonique et mécanique de la lithosphère **:** compréhension des processus thermomécanique de déformation de la croûte et de la lithosphère terrestre, et des interactions entre processus profonds et de surface, depuis les échelles transitoires à l'échelle (séismes, cycle sismique) jusqu'aux échelles géologiques (reliefs, réseaux de faille, paysage), impliquant l'analyse de signaux géodésiques (GPS, InSAR) et d'images optiques, avec des applications sociétales (risques naturels, environnement).

Paléogéographie et Paléoclimat **:** compréhension de l'histoire et des changements climatiques de la Terre sur des échelles géologiques en combinant enregistrements climatiques, informations tectoniques et paléo-magnétiques, et modèles climatiques et géochimiques, avec des applications sociétales et environnementales.

Enjeux interdisciplinaires. Parallèlement, un nombre croissant d'enjeux scientifiques exige aujourd'hui des approches qui traversent ces différents domaines et expertises, combinant des données multi-types issues d'expériences et de domaines différents avec des nouvelles méthodes de type multi-messagers. On peut citer aujourd'hui par exemple :

- La formation et la dynamique de la Terre et des planètes du système Solaire, en combinant modèles astrophysiques, études de la matière extra-terrestre, géochimie isotopique, et modélisation physico-chimique (expérimentation haute-pression et haute-température, méthodes ab-inito).
- La recherche de signaux dans les données de neutrinos et les expériences d'ondes gravitationnelles, déclenchée par des bouffées de rayonnement gamma détectées dans les rayonnements X dur ;
- La signature gravitationnelle et ionosphérique de grands tremblements de terre et de leurs précurseurs, analysée conjointement avec les données au sol (sismologiques et géodésiques) afin d'améliorer les méthodes de surveillance et d'alerte ;
- L'utilisation des neutrinos pour l'imagerie interne de la Terre, et des muons pour la tomographie des édifices volcaniques, conjointement avec les données sismologiques et gravimétriques.
- De nouvelles méthodes d'analyse multi-échelles (temps, fréquence, nombre d'onde) d'une diversité d'observations, en particulier pour la reconstruction des sources de rayonnement et des milieux.

Nombre de ces applications et méthodes issues de la recherche académique ont d'importantes applications pour la recherche industrielle, en particulier dans le domaine de l'exploration géophysique, et se traduisent par des collaborations industrielles, avec la chaire ANR industrielle associant l'IPGP, l'Ecole des Mines de Paris, TOTAL et Schlumberger, AirBus et Thales, et le consortium LITHOS associant l'IPGP, Cambridge et BP.

Missions spatiales et grands systèmes d'observation. Par ailleurs, les chercheurs de l'APC et de l'IPGP sont directement impliqués et ont une position stratégique dans l'élaboration et le suivi de missions spatiales, de grands instruments, et d'observatoires, dans le cadre de consortium nationaux (IR et TGIR) et internationaux, avec des activités qui nécessitent de combiner :

- simulations numériques, traitement et analyse de données complexes pour la caractérisation de la réponse, la capture (calibration, qualité) des données de ces expériences, dont les instruments sont de plus en plus grands et de plus en plus complexes (multi-détecteurs) ;
- développement de chaînes de traitement et d'analyse de données, associés à des méthodes d'apprentissage, pour le suivi et l'exploitation de ces expériences ;
- conception et opération de plateformes collaboratives (e.g. plateforme CODEEN du projet EUCLID, le « Data Processing Centre » du projet LISA) facilitant le développement et l'intégration

de ces différentes chaînes logiciels, combinées à des technologies de Virtualisation (docker, OpenStack), et intégrées aux centres d'archivage distribués de ces systèmes d'observation.

L'intégration de la plateforme DANTE en tant que nœud de la plateforme numérique CIRRUS de l'USPC pour la recherche et la formation, ouvrira ces ressources à une large communauté d'utilisateurs et de disciplines scientifiques.

BESOINS ET ORIGINALITE DE L'INSTRUMENT SCIENTIFIQUE

De nouveaux besoins et un changement de paradigme. Dans la plupart des domaines de recherche en sciences de la Terre, des Planètes et de l'Univers, on assiste aujourd'hui à une augmentation très importante du volume de données, issues des observations et des simulations, mais également à une augmentation de leur complexité et de leur diversité. De nombreux enjeux scientifiques requièrent aujourd'hui d'analyser et de modéliser des données non seulement issues d'expériences ou de simulations du même type, mais également au travers de domaines, de types de données et d'échelles spatiales et temporelles différents.

La conception et le suivi de missions spatiales, de grands instruments, de systèmes d'observation et d'observatoires, dont la taille et la complexité exigent aujourd'hui des simulations et des chaînes de traitement et d'analyse complexes pour caractériser leur réponse dans leur environnement (calibration), pour la capture et la réduction des données, ainsi que pour la quantification probabiliste des incertitudes et des niveaux bruit associés.

Le contrôle et la certification de la qualité et l'intégrité des données, de leurs annotations et de leur provenance sont des tâches importantes pour les centres d'archivages associés à ces systèmes d'observation. Elles dépendent aujourd'hui de plateformes de calcul et d'analyse avec des interfaces permettant d'accéder, de transférer, et de stocker temporairement de grands jeux de données, ainsi que des technologies de Virtualisation facilitant la préservation des logiciels d'analyse et de leur environnement, des expertises et de la connaissance de ces systèmes.

Exploiter les volumes et la diversité de ces données dans différents contextes requiert aujourd'hui des méthodes innovantes de traitement et d'analyse statistique, de simulation multi-échelle et multi-physique et de modélisation (inversion, imagerie, assimilation, inférence probabiliste) qui transforment la nature de la recherche dans les sciences de la Terre, des Planètes et de l'Univers. Ces méthodes dépendent de nouvelles ressources et d'architectures de calcul et d'analyse de données, ainsi que de services et d'environnements « virtualisés » qui en facilitent l'utilisation et s'adaptent de manière agile et flexible aux différentes chaînes d'utilisation de données.

Un nouvel instrument scientifique multidisciplinaire. L'architecture et les ressources existantes aujourd'hui à l'IPGP et à l'APC ne suffisent plus à répondre aux besoins de ces applications. Il est devenu nécessaire de fédérer et *in fine* intégrer ces ressources au sein d'une plateforme multidisciplinaire de calcul et d'analyse de données, d'en faire évoluer les capacités, l'architecture et les services, en synergie avec un environnement de type « Concurent Design Facility » (CDF) pour la conception et le suivi de missions spatiales, et des systèmes d'observation. Ce nouvel instrument scientifique doit être interfacé avec les centres d'archivage (nationaux et internationaux) des grands systèmes dans lesquels l'APC et l'IPGP jouent un rôle stratégique.

Il doit offrir des ressources dimensionnées pour les nouveaux besoins des applications multi- et interdisciplinaires en sciences de la Terre, des Planètes et de l'Univers, et un environnement de développement et de production adaptés à :

- la conception collaborative et au suivi de missions spatiales, et des grands systèmes d'observation ;
- de nouvelles méthodes de traitement et d'analyse statistique adaptées au volume et à la diversité des données, associés à des méthodes de « machine learning » et de nouvelles approches multi-messagers ;

- de nouvelles méthodes de modélisation numérique multi-physique et multi-échelle, d'assimilation de données, et d'inférence dans un cadre probabiliste ;
- l'orchestration des phases de calcul et d'analyse des différentes chaînes d'utilisation des données ainsi que l'optimisation des flux de données entre ces phases ;
- l'exploitation des nouvelles architectures de calcul et de données, et des nouvelles technologies de Virtualisation.

Une architecture adaptée et des services. L'architecture et les infrastructures « hardware » doivent permettre :

- d'assurer un flux de traitement et d'analyse aussi proche que possible de la vitesse d'accès aux données ;
- d'offrir une puissance de calcul adaptées aux nouvelles simulations multi-échelles et multi-physique ;
- d'exploiter de nouveaux modèles de programmation parallèle, d'exécution des tâches et des flux de données.
- de disposer de capacités et d'architectures de stockage adaptées aux différentes phases des chaînes d'utilisation des données.

Le projet DANTE fournira des services agiles et flexibles, qui ont fait le succès de FACe et de S-CAPAD, autorisant un déploiement rapide des applications dont la durée d'exécution peut aller de plusieurs heures à plusieurs mois, et un environnement pour le développement et l'adaptation des ces applications aux nouvelles architectures de calcul intensif.

Dimensionnement « research-driven » de l'instrument. Le nouvel instrument scientifique doit répondre aux besoins des applications à l'APC et l'IPGP :

- analyse et modélisation de masses de données d'un volume de 200 To et +
- simulations et modélisations exploitant une puissance de calcul de ~200 Tflops
- optimisation des flux de données tout au long des chaînes d'utilisation de données orchestrant phases de calcul et d'analyse
- exploitation des nouveaux paradigmes de programmation parallèle et des technologies de Virtualisation
- environnement de développement collaboratif autorisant une organisation flexible des ressources et des services (Cloud) et l'ordonnancement de plusieurs applications simultanées sur des durées de quelques heures à plusieurs semaines.

Cet instrument fédérera et intégrera au sein d'un cœur de réseau 10Gb/s, et d'un réseau d'interconnexion Infiniband :

- des ressources de stockage (~750 To), pour des données primaires et secondaires, *persistant et sécurisés* sur la durée des cycles d'utilisation des données (quelques mois à plusieurs années).
- des ressources pour le calcul haut débit (HTC), intégrant une forte capacité (~2 Po) de stockage locale (systèmes de fichiers parallèles et distribués, bases de données) et de mémoire (256 Go/nœud), et garantissant un flux de traitement proche du débit des données.
- des ressources de calcul parallèle (multi-cœurs, MIC), garantissant une puissance de calcul (~200 Tflops), pour les phases d'analyse statistique complexe, de simulation numérique, et de modélisation (assimilation, inférence probabiliste), incluant des méthodes de « machine learning ».
- une architecture fédérant ces ressources et optimisant le flux de données entre les différentes phases des chaînes complètes d'utilisation des données, et au travers des différentes ressources de calcul et de stockage de la plateforme.

- différents modèles de programmation parallèle (tâches et flux de données, MapReduce), de communication et d'exécution (OpenMP, MPI, SPARK), de langages (Fortran, C/C++, Python) adaptés aux différentes phases de la chaîne complète d'utilisation de données.
- des technologies de virtualisation (docker, OpenStack, Kubernetes) associés à de nouveaux environnements Big Data (SPARK, STORM).
- une infrastructure d'hébergement garantissant la maitrise et l'optimisation des coûts fluide et électrique.
- un environnement et des services, i.e. « Concurent Design Facility » (CDF), pour la développement collaboratif et le suivi de missions spatiale, de grands instruments et systèmes d'observation, et d'observatoires.

Une « Concurrent Design Facility ». La CDF a été conçue à la demande du CNES, et disposera d'une salle commune équipée de postes de travail, associés à un système de diffusion et réception multimédia (matrice HDMI, écran interactif, ~10 écrans escamotables, un équipement de visioconférence) ainsi qu'un ensemble d'outils logiciels, adaptés aux différents champs d'expertise. Cet équipement permet une interaction rapide et efficace entre les différentes disciplines, et les équipes internationales, engagées dans ces projets, via des méthodes d'ingénierie concourante. De cette manière, un ensemble cohérent de paramètres de conception peut être défini et échangé, et tout changement susceptible d'avoir un impact sur d'autres disciplines ou composantes du projet être identifié et évalué collectivement, garantissant un résultat cohérent et de grande qualité dans le minimum de temps.

La CDF n'offre pas seulement une aide à la conception collaborative de missions spatiales, de grands instruments, et observatoires. Elle permet aussi, associée aux ressources de calcul et d'analyse de données de la plateforme, d'aboutir à une définition précise de détecteurs, systèmes, sous-systèmes, ou tout autre développement instrumental devant être analysé et simulé au niveau système. Divers projets peuvent bénéficier de cet outil qui réduit les coûts liés aux déplacements d'experts extérieurs et permet l'organisation de téléréunions. Cette salle de développement collaboratif a aussi une vocation pédagogique et de formation, avec une diffusion potentiellement internationale.

Un « laboratoire numérique ». La diversité des données, de leur analyse et de leur modélisation en sciences de la Terre, des Planètes et de l'Univers, fait du projet DANTE un formidable « test bed » pour les nouvelles technologies et méthodes de calcul et d'analyse, aux niveaux matériel, système d'exploitation, logiciels d'analyse et de modélisation, et schémas de stockage et de base de données. Cet laboratoire numérique fourni par DANTE jouera un rôle important en tant que pôle de connaissances et d'expertise multi- et interdisciplinaires favorisant et accélérant les projets numériques (« machine shop »), le partage de méthodes et de logiciel, d'outils de visualisation, des interfaces, ainsi que la conception de systèmes d'observation, et d'outils de diffusion des connaissances.

Il fournira un environnement pour des « sprints » entre experts des domaines applicatifs, des sciences des données, de la recherche informatique et les « data providers » afin d'accélérer de bout et bout les diverses chaînes d'utilisation des données. Il constituera enfin un laboratoire numérique pour l'émergence des nouvelles thématiques et méthodes interdisciplinaires d'inférence statistique afin d'exploiter les informations contenues les volumes et la grande diversité de données issues des systèmes d'observation et des simulations numériques.

Il favorisera également l'animation et la formation d'une communauté de jeunes chercheurs et d'étudiants autour de l'utilisation de ce nouvel instrument, des méthodes et des technologies, devenant ainsi un point de rassemblement et d'attractivité pour les communautés scientifiques de l'USPC, intégré aux formations universitaires de l'USPC.

LOCALISATION

L'IPGP dispose dans ses bâtiments de l'Îlot Cuvier de 3 salles informatiques, partiellement utilisées et conçues pour pouvoir absorber les extensions dans les prochaines années. Afin d'assurer une sécurité maximale pour les ressources primordiales (cœur de réseau, serveurs centraux, etc.), les services généraux sont localisés dans une salle avec une politique d'accès spécifique. Les ressources de la plateforme de calcul et d'analyse de données seront installées dans deux salles machines du bâtiment de l'IPGP sur l'Îlot Cuvier :

- Surface au sol : 2 X 16 baies (2 X 38,2 m^2), actuellement partiellement utilisées
- Climatisation : 2 X 70 kW (soufflage inversé)
- Puissance électrique : 2 X 112 kW (sécurisé dont une partie ondulée)
- Caractéristiques : accès contrôlé (badges personnalisés)

Ces salles disposent de l'alimentation électrique nécessaire pour la nouvelle plateforme. La température et la consommation électrique sont suivies en temps réel et à distance, au moyen de capteurs disposés dans ces salles. Ces sondes consultables à distance envoient des alertes en cas de dépassement des seuils. Par ailleurs ces salles sont également surveillées par une équipe de Sécurité Incendie au cours de rondes régulières, qui prévient les correspondants et le service de maintenance en cas événement inhabituel.

En ce qui concerne l'efficacité énergétique actuelle de ces salles, la création d'allées chaudes/froides, la pose de dalles percées dans les allées froides, le cloisonnement des équipements et l'augmentation de la température de consigne a pour objectif de garantir un PUE bien inférieur à 1,8. L'installation des compteurs dans les salles, et la connexion des nouveaux équipements à des PDU pour la surveillance et l'administration, permettront d'améliorer encore cette efficacité.

Une nouvelle liaison 10 Gb/s entre l'IPGP et l'extérieur (RENATER) est par ailleurs planifiée.

La salle du « Concurrent Design Facility » restera hébergée dans les locaux de l'APC sur le campus Paris Rive Gauche. Elle sera équipée de postes de travail, associés à un système de diffusion et réception multimédia, ainsi qu'un ensemble d'outils logiciels spécifiques aux différents champs d'expertise.

MANAGEMENT ET MODALITES D'ACCES DES EQUIPEMENTS

La plateforme DANTE est un instrument scientifique ambitieux, avec un budget d'équipement de l'ordre de 1,3 M€. La gouvernance du projet est construite autour d'un responsable scientifique, d'un responsable technique avec un comité de pilotage composé de représentants des utilisateurs, d'un représentant de la direction de l'IPGP et de l'APC, d'un représentant de l'USPC (en lien avec la plateforme CIRRUS et l'EUR) et du CNRS, et d'un représentant de GENCI et des méso-centres (FR-T2).

La stratégie de DANTE est de construire de nouvelles synergies avec d'autres communautés scientifiques, au delà de la communauté en formation Terre-Planète-Univers, par exemple la biologie, la médecine et la santé, l'informatique, les sciences économiques et sociales, en particulier dans le cadre de l'USPC, qui partagent des pratiques de recherche et des expertises sur le calcul et l'analyse de données. La gouvernance de la plateforme sera agile et flexible afin de prendre en compte cette extension vers d'autres disciplines et intégrer de nouveaux groupes de recherche.

Le modèle économique de DANTE s'inspire des recommandations de la feuille de route du MENESR (2017) sur la modernisation des infrastructures et services numériques des établissements de

l'enseignement supérieur et de la recherche. Le coût induit de fonctionnement de la plateforme de calcul et d'analyse est ainsi estimée comme suit :

- La plateforme DANTE comprendra 4 Baies 42U remplies à 80%.
- Les frais induit de fonctionnement (maintenance bâtiment, amortissement baie vide, fluides) sont estimés à 12 k€/an pour 1 baie 42U remplie à 80% : soit 48 k€/an pour les 4 Baies.
- La plateforme disposera de 3000 cœurs de calcul et d'analyse, soit une capacité de 263 kheures/an.

Deux modèles économiques sont aujourd'hui considérés à titre d'exemple :

- *Sans amortissement des ressources de calcul et d'analyse*. Si 10% de ces heures de calcul et d'analyse sont facturés à 2 centimes/cœur/heure, cela permet de dégager ~52 k€/an (240 k€ sur 5 ans) pour couvrir les frais induits de fonctionnement (> 48 k€/an).
- *Avec amortissement des ressources de calcul et d'analyse*. Si 50% de ces heures de calcul et d'analyse sont facturées toujours à 2 centimes/cœur/heure, cela permet de dégager ~263 k€/an (~1 300 k€ sur 5 ans).

Les utilisateurs de DANTE jouent aujourd'hui un rôle moteur dans de nombreux projets nationaux et internationaux, et ont déjà rencontré un succès considérable auprès des grandes agences de financement européennes et internationales dans le cadre de missions spatiales et de grands équipements (TGIR, IR) ainsi que de projets ERC, H2020, TRN, et ANR. Ils se sont déjà engagés à inclure dans le cadre de ces projets les coûts induits d'utilisation de la plateforme, ce qui doit garantir la viabilité du modèle économique considéré.

Le projet DANTE entend également encourager le calcul et l'analyse des données dans les pratiques de recherche au sein de l'IPGP et de l'APC, et de l'USPC. A ce titre, l'accès de ces ressources sera ouvert à des applications non encore supportées par des projets nationaux ou internationaux, en particulier afin de permettre de construire de tels projets. Un partie des coûts induits par l'utilisation de la plateforme dans ce contexte sera pris en charge par l'IPGP, l'APC.

Par ailleurs, les utilisateurs pourront demander dans leurs projets des compléments d'équipements spécifiques qui seront intégrés et administrés dans le cadre de la plateforme DANTE. Cela permettra également d'avoir la flexibilité nécessaire pour intégrer les besoins de nouvelles communautés d'utilisateurs au sein de l'USPC.

L'accès à la plateforme DANTE sera naturellement ouvert à tous les chercheurs de l'IPGP et de l'APC et à leurs collaborateurs extérieurs (nationaux et internationaux), sur simple dépôt d'une fiche concise explicitant l'application et le projet scientifique. En tant que nœud de la plateforme numérique partagée CIRRUS de l'USPC, pour la recherche et la formation, la plateforme DANTE sera également ouverte à l'ensemble des chercheurs et étudiants de l'USPC sur simple demande. La procédure restera volontairement légère afin de préserver une réactivité et une flexibilité au fil de l'eau.

Durant le projet, un système de certificat (AAI et IdP) sera mis en place en collaboration avec RENATER, afin de faciliter l'accès et l'utilisation de ces ressources et de ces services par les utilisateurs. La stratégie de DANTE est d'offrir les services nécessaires pour respecter les diverses politiques d'accès, de propriété et de confidentialité des données hébergées sur la plateforme lors de leur analyse. Elle entend également promouvoir et faciliter dans les pratiques de recherche l'Open Data et l'Open science (e.g. FAIR Data).

Une charte d'utilisation de DANTE imposera que toutes publications, issues de l'utilisation des ressources et des services de DANTE, fassent explicitement référence à DANTE et au soutien de la

région Île-de-France, et que celles-ci soit transmises (eg. Doi) afin d'être référencées et visibles sur le site du projet DANTE.

MOYENS HUMAINS AFFECTES A L'INSTRUMENT

La responsabilité scientifique du projet sera conjointement assuré par Jean-Pierre Vilotte (Physicien des observatoires, IPGP) et par Cécile Calvet (IR 2 CNRS, APC), en particulier pour les missions spatiales.

La responsabilité technique et opérationnelle du projet pour l'aspect Cluster de calcul sera assurée par Geneviève Moguilny (IR1 CNRS, IPGP). L'équipe est complétée par 2 ingénieurs informaticiens : Martin Souchal (IE2 P7, APC) pour les ressources cluster, et David Weissenbach (IE CNRS, IPGP), en particulier pour les aspects Virtualisation et services de type Cloud.

La responsabilité technique et opérationnelle du projet pour l'aspect CDF sera assurée par Michèle Detournay (IRHC CNRS, APC). L'équipe est complétée par Sébastien Zappino (AI P7, APC) et par Hubert Halloin (MC P7, APC) en qualité de référent scientifique.

Le projet DANTE bénéficiera du soutien des services informatiques de l'IPGP et de l'APC. Les ressources de la plateforme hébergées au sein des locaux de l'IPGP, seront une composante de l'unité mixte de service (UMS) de l'IPGP.

Le projet DANTE bénéficiera par ailleurs d'un fort soutien dans le cadre des projets et instruments nationaux et internationaux de l'IPGP et de l'APC, ainsi des projets de missions spatiales dans lesquels ils sont engagés. Dans ce cadre, des CDD chercheurs et ingénieurs financés par ces projets contribuent aux développements, à la validation et à l'intégration de logiciels sous la forme en particulier de machines virtuelles, et au pôle de connaissance associé à cette plateforme.

Dans le cadre de sa contribution au projet CIRRUS, le projet DANTE est appuyé par la ComUE USPC, et continuera à bénéficier des soutiens de l'USPC à CIRRUS, ce qui permettra de bénéficier de personnel contractuel en appui à l'accès et à l'utilisation de la plateforme par d'autres communautés scientifiques de l'USPC.

Dans le cadre de son nouveau projet quinquennal, l'IPGP a identifié de nouveaux besoins avec en particulier l'affichage d'un ingénieur de recherche pour le développement et l'intégration logiciel, en soutien au calcul et à l'analyse des données dans les nouvelles pratiques de recherche, ainsi qu'un profil de chercheur ou d'ingénieur de recherche en sciences des données.

IMPACT POTENTIEL SCIENTIFIQUE ET TECHNOLOGIQUE FRANCILIEN

De part les volumes et la diversité des données et des chaînes d'utilisation de ces données, issus de missions spatiales, des systèmes d'observation et de surveillance, et de simulations numériques en sciences de la Terre, des Planètes et de l'Univers, la plateforme de calcul et d'analyse de données DANTE constituera un instrument scientifique unique en Île-de-France, et un formidable laboratoire numérique qui renforcera l'écosystème du calcul en Île-de-France, en favorisant le développement : de nouvelles méthodes, de technologies et d'architectures de calcul et de données ; de technologies de Virtualisation (Cloud) ; et un modèle organisationnel (technique, humain) pour l'exploitation et le « stewardship » de cet instrument.

Au niveau de la région, la stratégie « research-driven » et fédérative de DANTE est de construire une synergie et des collaborations fructueuses avec :

- les centres nationaux de GENCI : IDRIS, TGCC, CINES, et du CNRS : CC-IN2P3, en liaison avec le CNRS, le CEA, le CNES, afin d'étudier les conditions et les modèles d'organisation pour l'hébergement dans les futurs grands « Data Centres » nationaux labélisés, d'une telle plateforme multidisciplinaire de calcul et d'analyse de données adaptée aux grandes masses et à la diversité des données issues des systèmes d'observation en sciences de la Terre, des Planètes

et de l'Univers, et qui sont archivées de manière distribuée à l'échelle européenne et internationale.

- la maison de la simulation[27] (CNRS, CEA, INRIA, université d'Orsay et université de Versailles – St Quentin), localisée sur le plateau d'Orsay, pour le développement et l'enseignement de nouvelles méthodes de calcul, d'analyse et de visualisation de gros volumes de données.
- l'effort de structuration, encouragé aujourd'hui par le MENESR dans le cadre de la feuille de route nationale du numérique, des méso-centres nationaux (FR-T2) et du du GIS France Grilles[28], au sein duquel l'APC et l'IPGP ont contribué à travers de nœuds spécifiques, et de l'Institut des Grilles et du Cloud (CNRS).
- l'initiative du European Open Science Cloud[29] (EOSC), impulsée par la commission européenne, auquel les équipes de l'APC et de l'IPGP contribuent déjà en liaison avec le CNRS, le CEA et le CNES, en particulier pour la suite du projet EOSCpilot[30].
- la communauté de recherche en informatique et en mathématiques, en particulier au travers de l'Institut de Mathématiques de Jussieu Paris Rive Gauche[31] et de l'Institut du calcul et de la simulation[32] (UPMC) pour le développement de méthodes innovantes en calcul et analyse de gros volumes de données, incluant les nouvelles approches multi-messagers et les avancées dans les domaines de l'inférence probabiliste en grande dimension et du « machine learning ».
- La communauté de recherche du Climat, en particulier au travers de l'infrastructure national CLIMERI[33] et de son nœud CICLAD[34] à l'IPSL (UPMC) ;
- les grands observatoires des sciences de l'Univers franciliens (e.g. Observatoire de Paris-Meudon, Observatoire Ecce Terra) ;
- la recherche industrielle et les agences spatiales, en particulier au travers des partenariats déjà existants à l'IPGP et à l'APC dans les domaines de l'exploration géophysique (Chaire ANR Industrielle) et du spatial (CNES, ESA, NASA).

Dans le cadre de l'USPC, le projet DANTE est articulé avec : le Campus Spatial PRG, en collaboration avec le CNES ; le projet d'Institut des Sciences des Données, au sein de la ComUE ; et au niveau de l'enseignement avec le projet d'Ecole Universitaire de Recherche Terre-Planètes-Univers, porté par l'APC et l'IPGP.

Dans le paysage complexe et en pleine mutation de la recherche et de l'enseignement de la région Île-de-France, DANTE renforcera le potentiel scientifique et technologique en répondant aux besoins

---

27 http://www.maisondelasimulation.fr/

28 http://www.france-grilles.fr/accueil/

29 https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

30 http://libereurope.eu/blog/2017/02/23/eoscpilot-european-open-science-cloud-research-pilot/

31 https://www.imj-prg.fr

32 http://www.upmc.fr/fr/recherche/modelisation_ingenierie/les_structures_federatives/institut_du_calcul_et_de_la_simulation.html

33 http://www.enseignementsup-recherche.gouv.fr/cid99402/infrastructure-nationale-de-modelisation-du-systeme-climatique-de-la-terre-climeri-fr.html

34 http://ciclad-web.ipsl.jussieu.fr

de la recherche « data-intensive » dans les domaines académique et finalisé issus des sciences de la Terre, des Planètes et de l'Univers, ainsi que des agences spatiales (CNES, ESA, NASA).

Les chercheurs de l'IPGP et de l'APC pilotent et participent à un grand nombre de : projets européens ERC; missions spatiales en liaison avec le CNES, l'ESA, la NASA et les agences japonaise et chinoise ; grands instruments et systèmes d'observation internationaux, inclus dans les feuilles de route nationales (SNRI) et européennes (ESFRI). Ces projets renforceront la visibilité nationale et internationale du projet DANTE et de la région Île-de-France en tant que pôle de calcul et d'analyse de données en sciences de la Terre, des Planètes et de l'Univers.

Les données archivées dans les infrastructures distribuées, et associées au grands systèmes d'observation et observatoires nationaux et internationaux, ne sont une source de connaissance utile que si leur qualité est certifiée et qu'elles peuvent être réutilisables (méta données, annotations, provenance, logiciel) dans d'autres contextes que ceux des expériences qui les ont acquises. Les technologies de Virtualisation et les archives de machines virtuelles, associées au pôle de connaissance et d'expertise multidisciplinaire du projet DANTE seront d'une grande valeur pour la communauté afin de préserver l'expertise et la connaissance ds systèmes d'observation et des méthodes de traitement et d'analyse.

IMPACT SOCIO-ECONOMIQUE POTENTIEL

La science des données est un des grands défis cognitifs, économiques et sociétaux de nos temps. Elle concerne la capture, l'analyse, la visualisation, le stockage persistant, la sécurité et le partage des données, souvent non structurées, épars ou inhomogènes. Elle touche aussi à plusieurs enjeux éthiques et sociales. Elle couvre le domaine du Big Data caractérisé par l'augmentation exponentielle de la vitesse de calcul, la capacité de stockage, la capacité réseau mais aussi par de nouvelles méthodes d'extraction des informations pertinentes exploitant la *cornucopia* et la diversité des données. Elle couvre également l'Internet des Objets caractérisé par la collection des données générées par les réseaux de capteurs distribués et déployés dans des environnements hostiles pour l'étude de la Terre et de l'Univers. La plateforme DANTE dans ses différentes composantes, infrastructure, pôle de compétences, et outil de recherche et de formation, jouera un grand rôle dans le réseau d'informatique francilien.

Deuxièmement, soixante ans après le lancement de Spoutnik, le spatial est présent un peu partout dans nos vies : information, télécommunications, gestion des transports, observation de la planète, météorologie, aménagement du territoire, agronomie, océanographie, gestion des ressources, astronomie, astrophysique, physique fondamentale, cosmologie ... Cette révolution du spatial a bien évidemment des conséquences sur la recherche et l'enseignement, que ce soit en Sciences de la Terre et de l'Environnement, en Géographie, en Planétologie, en Physique, en Sciences de l'Univers, et en aménagement des villes. Le spatial devient ainsi un débouché important pour les étudiants des différentes filières rattachées. Ce sont des dizaines, voire des centaines de milliers d'emplois, dont beaucoup hautement qualifiés, qui vont être créés dans les prochaines années dans ces domaines : positionnement satellitaire, prévention et gestion des risques naturels ou des catastrophes, environnement, développement urbain, gestion et développement des ressources terrestres et maritimes, etc. Tous ces secteurs d'activités et leurs applications requièrent en parallèle des compétences spécifiques en matière de traitement, d'analyse et de modélisation des données, qui nécessitent des nouveaux besoins dans l'enseignement et la recherche qui vont générer de nouveaux emplois à leur tour. Ce sera un deuxième but sociétal de DANTE

D 'autres objectifs, seront développées ultérieurement avec l'inclusion de nouveaux partenaires : coté biologie les grandes bases de données du vivant et leur analyse, coté sciences sociales les grandes bases de données sociales dont l'analyse se développe à des institutions partenaires de l'USPC (Science-PO) et avec lesquelles des programmes interdisciplinaires de l'USPC (« Politique de la Terre », Sciences-Po/IPGP) sont en développement.

INCIDENCE SUR LA FORMATION DES JEUNES CHERCHEURS

Les pratiques de recherche en Sciences de la Terre, des Planètes et de l'Univers font appel de manière de plus en plus intensive au calcul et à l'analyse de gros volumes et d'une grande diversité de données. Cette évolution est partagée par d'autres communautés scientifiques au sein de la ComUE USPC, en particulier bioinformatique, médecine et santé, et sciences humaines et sociales. Il est indispensable aujourd'hui de donner aux générations futures de scientifiques et d'ingénieurs des compétences en sciences des données, en modélisation numérique et en informatique théorique, adaptées aux nouveaux enjeux de la recherche afin de leur permettre de réussir leur carrière dans un environnement de plus en plus compétitif. La plateforme multidisciplinaire de calcul et d'analyse de données DANTE entend constituer un pôle de connaissance et de référence à tous les niveaux de l'enseignement au sein de la ComUE USPC.

Ecole Universitaire de Recherche Terre-Planètes-Univers. Un premier point de focalisation sera l'école universitaire de recherche (EUR) Earth-Planets-Universe, que l'IPGP, APC et AIM proposent de mettre en place au sein de l'USPC. Les étudiants recrutés dans cette EUR (flux prévu : une soixantaine d'étudiants par an, venant du monde entier) auront la possibilité de développer un portefeuille de compétences transverses lors de leur cursus. Ces compétences porteront en particulier sur les nouvelles méthodes et technologies de calcul et d'analyse de données qui doivent être mise en œuvre pour tenter de répondre à des grandes questions (disciplinaires et interdisciplinaires) de la recherche fondamentale ou appliquée, ainsi qu'aux défis des nouveaux systèmes d'observation en sciences de la Terre, des Planètes et de l'Univers.

Des formations spécifiques, sont déjà offertes actuellement (en anglais) dans le cadre de l'école doctorale STEP'UP et proposées à l'ensemble des acteurs de la ComUE. Elles sont alignées sur les nouvelles pratiques de recherche en sciences de la Terre, des Planètes et de l'Univers, et sont conçues en complément de l'offre proposée par les centres nationaux (e.g. l'IDRIS). Dans l'EUR, ces formations seront renforcées et élargies. Elles s'appuieront sur les ressources et le pôle de connaissance et de mutualisation d'expertises multi- et interdisciplinaires que constituera la plateforme DANTE. Plusieurs journées de formation, incluant de nombreux travaux pratiques et des environnements de virtualisation dédiés, seront offertes aux étudiants tout au long de leur parcours dans l'EUR.

L'ensemble constituera une « rampe de connaissance » permettant à ces étudiants et jeunes chercheurs de s'engager progressivement dans ces nouvelles méthodes d'inférence statistique pour l'extraction de nouvelles informations enfouies dans la masse et la diversité des données issues des grands systèmes d'observation et des simulations numériques, et ce en fonction des besoins de leurs applications. Dans ce cadre, DANTE sera représenté au sein du conseil de gouvernance de l'EUR afin de développer la meilleure synergie possible. Ces formations sont et seront conçues de manière agile en coévolution avec les besoins et les pratiques de la recherche, ainsi qu'avec l'évolution rapide des méthodes et des technologies de calcul et d'analyse des données.

Outre les actions de formation, DANTE sera la plateforme sur laquelle les étudiants (niveau master et doctorat) pourront naturellement mener leurs travaux de recherches (stages de master et doctorat).

Chaire de physique de l'intérieur de la Terre du Collège de France. L'évolution rapide des pratiques de recherche, et les nouvelles connaissances sur la dynamique de l'intérieur de la Terre et des planètes, que fournit aujourd'hui l'analyse de la grande masse et de la diversité des données issues

des systèmes d'observation et des simulation numériques, sont au cœur de l'enseignement et de la recherche de la chaire de Physique de l'Intérieur de la Terre au Collège de France[35]. L'équipe de recherche du titulaire de cette chaire, le Professeur Barbara Romanowicz, est implantée à l'IPGP et a été financée jusqu'en 2016 par une ERC (WAVETOMO). Elle bénéficiera de manière cruciale de l'instrument scientifique DANTE, qui contribuera ainsi à la réalisation d'une des principales missions du Collège de France : enseigner « le savoir en train de se faire ». A ce titre, le Collège de France est impliqué dans et soutient le projet DANTE.

Chaire ANR industrielle en géophysique d'exploration. Dans le cadre de la Chaire Industrielle ANR, associant IPGP, Ecole des Mines de Paris, Total et Schlumberger, un programme de recherche et de formation spécifique à l'exploration géophysique a pour but de former des jeunes chercheurs destinés à mener des activités de recherche dans les milieux académique et industriel. Un système de bourse permet de recruter au niveau mondial. L'enseignement porte sur les méthodes de pointe en exploration géophysique et en relation avec des problématiques industrielles, mais également en sciences de la Terre, pour lesquelles les aspects calcul et analyse de gros volumes de données sont devenus critiques.

La plateforme DANTE permettra de familiariser les étudiants avec les environnements et les méthodes de calcul et d'analyse de données massives qui leur permettra de résoudre de nouveaux problèmes dans ces domaines, en particulier : imagerie, tomographie et migration de données sismiques actives. Ils pourront ainsi acquérir une expertise unique pour le développement de leur carrière.

Pôle de connaissance multidisciplinaire en Sciences des Données. DANTE entend progressivement élargir et renforcer ce volet formation en étroite collaboration avec d'autres communautés scientifiques de la ComUE USPC, qui partagent ces besoins en calcul et analyse de données dans leurs pratiques de recherche, en particulier dans les domaines de la médecine et de la santé, de la bioinformatique, et des sciences humaines et sociales. DANTE contribuera ainsi au projet d'un Institut multidisciplinaire en Sciences des Données porté par l'USPC, et dont le but est de créer de nouvelles synergies pour une meilleure mutualisation des expertises existantes au sein de l'USPC dans les domaines de la recherche informatique, des sciences des données, de l'intelligence artificielle, des mathématiques, de la physique des milieux complexes, des sciences des matériaux, de sciences de la Terre, des Planètes et de l'Univers, de l'environnement, de la linguistique, de la médecine et de la santé, et des sciences humaines et sociales.

POTENTIEL POUR LA SENSIBILISATION DU GRAND PUBLIC AUX ENJEUX DE LA RECHERCHE

Depuis leur déménagement dans les nouveaux locaux de l'Îlot Cuvier et de Paris Rive Gauche, l'IPGP et l'APC sont devenus une vitrine attractive du potentiel de recherche scientifique et d'enseignement en Île-de-France. Le succès considérable rencontré par les manifestations organisées ces dernières années dans le cadre de la Fête de la Science, atteste de cette nouvelle dynamique et de cette visibilité accrue.

Le projet est construit en forte synergie avec le Labex UnivEarths. Plusieurs activités du Labex (MOOCs, initiatives de science citoyenne, écoles et ateliers pour les professeurs des lycées et collèges, etc.) bénéficieront et s'appuieront sur DANTE et son pôle de connaissance en calcul et d'analyse de données. Des expériences de réalité virtuelle (présence Martienne ou présence au fonds de la mer) pourront ainsi être développées et soutenues.

[35] http://www.college-de-france.fr/site/barbara-romanowicz/

Les cellules de communication de l'IPGP (Médi@terre) et de l'APC disposent d'équipements innovants permettant la diffusion de la recherche en sciences de la Terre, des Planètes et de l'Univers vers le grand public (mur d'écrans, projections, conférences, expériences en direct).

Le projet DANTE bénéficiera et contribuera à ce potentiel pour la diffusion et la sensibilisation du grand public aux enjeux et aux résultats des méthodes de calcul et de l'analyse de données, tant au niveau de la recherche fondamentale et industrielle qu'au niveau de l'élaboration et du suivi des missions spatiales et des systèmes d'observation, et d'applications sociétales comme la surveillance et l'évaluation des risques naturels associés au séismes et aux volcans. DANTE, au travers d'outils de visualisation et d'imagerie scientifique, augmentera l'impact de ces manifestations et la visibilité des investissements.

Au-delà de cette dynamique événementielle, DANTE entend se doter d'un portail scientifique à destination du grand public et d'un « hub » de connaissance et de mutualisation d'expertise à destination de la communauté scientifique. Ce canaux de diffusion permettront de présenter les activités, les méthodes, les logiciels et leurs étapes d'avancement, et fournira un matériel pédagogique (visualisations, images synthétiques, animations informatiques) à destination du grand public et des acteurs de la vulgarisation scientifique (journaux, télévisions) avec lesquels l'IPGP et l'APC entretiennent des collaborations de longue date.

Dans le cadre de la Chaire de Physique de l'Intérieur de la Terre du Professeur Barbara Romanowicz, DANTE bénéficiera des moyens de diffusion du Collège de France qui soutient le projet. L'excellence de cette institution, reconnue en France et à l'étranger, dans le domaine de la diffusion des savoirs et la dissémination de la recherche de haut niveau auprès du grand public sera un atout important pour le projet et ne pourra qu'augmenter sa visibilité.

Enfin, DANTE s'articule autour de grands projets de missions spatiales, d'instruments et systèmes d'observation nationaux et internationaux. Ces projets incluent eux-mêmes une stratégie de diffusion des savoirs et de dissémination des résultats de la recherche à l'attention du grand public. DANTE bénéficiera de ces canaux et exploitera d'une manière générale toutes les synergies avec les canaux de diffusion apportés par le réseau dense de projets scientifiques nationaux et internationaux qui contribuent à cet instrument scientifique.

PLANNING ET DATES CLEFS

Le planning prévisionnel et les principales dates clefs du projet sont:

- **Janvier 2018** : déménagement des cluster Arago de FACe (APC) dans les locaux de l'Îlot Cuvier de l'IPGP et fédération avec les ressources de S-CAPAD (IPGP)
- **Juin 2018** : équipement de la nouvelle salle CDF dans les locaux de l'APC sur le camus Paris Rive Gauche
- **Mai - Septembre 2018** : définition et préparation de l'appel d'offre pour l'acquisition des ressources de la nouvelle plateforme de calcul et d'analyse de données.
- **Octobre/Novembre 2018** : Lancement de l'appel d'offre.
- **Juin 2019** : Désignation du lauréat et préparation du déploiement des nouvelles ressources de calcul et de stockage.
- **Septembre 2019** : Début de l'installation des nouvelles ressources.
- **Décembre 2019** : Ouverture de la plateforme aux utilisateurs.

RECAPITULATIF DU BUDGET TOTAL DE L'OPERATION

Le budget total de l'opération DANTE peut de décomposer de la manière suivante:

1. **Calcul et analyse de données** : (813 k€)

- 72 serveurs à 28 coeurs et 64 Go de RAM
- 16 serveurs à 28 coeurs et 128 Go de RAM
- 4 serveurs "épais" à 20 coeurs, 256 Go de RAM et 2,4 To de disques SSD
- 2 To bruts de stockage parallèle
- Réseau Infiniband (interconnection)

2. **Environnement de Virtualisation** : (152 k€)
   - 16 noeuds à 36 coeurs et 128 Go de RAM
   - 48 To bruts de stockage parallèle
   - Réseau Infiniband (interconnection)
3. **Stockage persistant** : (180 KE)
   - Baie hautement sécurisée de 591 To utiles
4. **Salle « Concurrent Design Facility »** : (59 k€)
   - 14 moniteurs 24" sur tables escamotables
   - Connectique HDMI
   - Système de visio-conférence
   - 1 écran tactile

Coût total de l'opération : 1 204 k€

Contribution Région Île-de-France demandée : 600 k€

Contribution IPGP et APC sur fonds propre déjà disponibles : 604 k€

PUBLICATIONS ET REFERENCES

[1] C. Cavet, A. Petiteau, M. Le Jeune, E. Plagnol, E. Marin-Martholaz, J-B. Bayle, **A proto-Data Processing Center for LISA**, 11th International LISA Symposium, Accepted (2017)

[2] Babak, S., Gair, J., Sesana, A., Barausse, E., Sopuerta, C. F., Berry, C. P. L., Berti, E., Amaro-Seoane, P., Petiteau, A., Klein, A., **Science with the space-based interferometer LISA. V: Extreme mass-ratio inspirals**, arXiv170309722B (2017)

[3] Amaro-Seoane, P., Audley, H., Babak, S., Baker, J., Barausse, E. et al., **Laser Interferometer Space Antenna**, arXiv170200786A (2017)

[4] Armano, M., Audley, H., Auger, G., Baird, J. T., Binetruy, P. et al., **Charge-induced force-noise on free-falling test masses: results from LISA Pathfinder**, arXiv170204633A (2017)

[5] M. Poncet, T. Faure, C. Cavet, A. Petiteau, P.-M. Brunet, E. Keryell-Even, S. Gadioux, M. Burgaud, **Enabling collaboration between space agencies using private and cloud based clusters**, BiDS'16 (2016)

[6] Savchenko, V., Bazzano, A., Bozzo, E., Brandt, S., Chenevez, **INTEGRAL IBIS, SPI, and JEM-X observations of LVT151012**, arXiv170401633S (2017)

[7] M. Airaj, C. Biscarat, C. Cavet, N. Clémentin, S. Geiger, C. Gondrand, V. Hamar, M. Jouvin, V. Legoll, S. Li, C. Loomis, M. Marquillie, G. Mathieu, J. Pansanel, G. Philippon, J.-M. Pierson, M. Puel, G. Romier, F. Thiebolt, A. Tsaregorodtsev, **FG-Cloud : Cloud communautaire distribué à vocation scientifique**, hal-in2p3-01285123 (2015)

[8] M. Airaj, C. Cavet, V. Hamar, M. Jouvin, C. Loomis, A. Lopez Garcia, G. Mathieu, V. Mendez, J. Pansanel, J.-M. Pierson, M. Puel, F. Thiebolt, A. Tsaregorodtsev, **Vers une fédération de Cloud académique dans France Grilles**, hal-00927506 (2013)

[9] C. Cavet, M. Le Jeune, F. Dodu, M. Detournay, **Utilisation du Cloud StratusLab : tests de performance des clusters virtuels**, hal-00766067 (2012)