



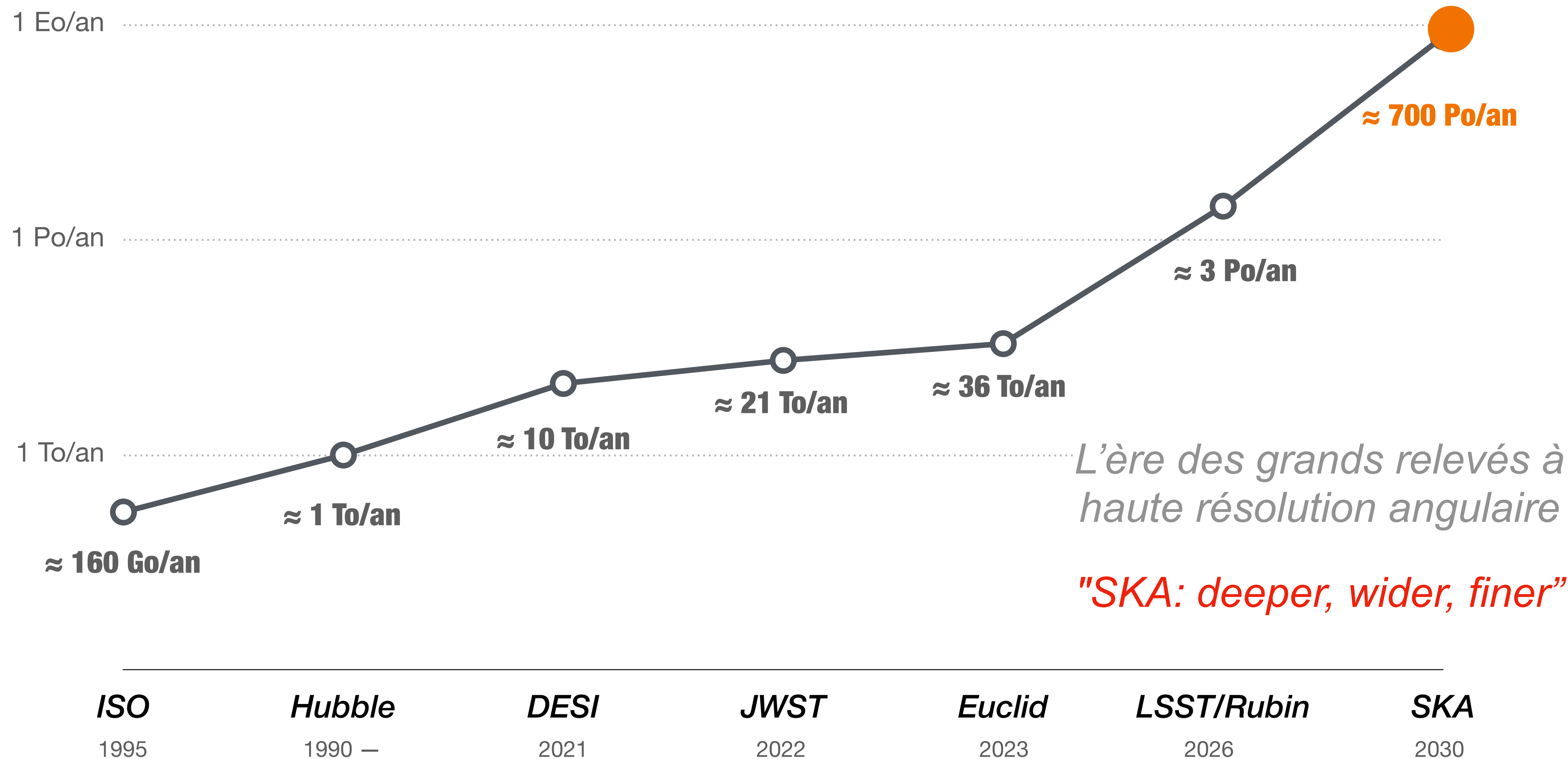
**Les défis astrophysiques
du HPDA :**

**L'exemple du Square
Kilometre Array**

Benjamin Magnelli



Trente Ans d'Évolution des Données Astronomiques



HPDA : Quand le Calcul Intensif Rencontre la Donnée Massive

HPC

Calcul Haute Performance

Supercalculateurs, parallélisme massif, exaflops

+

Big Data

Analytique de données massives

Volume, vitesse, variété : ingestion, stockage et fouille de flux de données à très grande échelle.

+

IA / ML

Apprentissage automatique

E.g., Détection et classification automatisées dans des flux trop grands pour les méthodes classiques.

High Performance Data Analytics = utilisation de super-calculateurs pour produire, analyser et traiter à très grande échelle, des flux de données scientifiques gigantesques — l'analyse doit souvent avoir lieu *pendant* l'acquisition, pas après.

Square Kilometre Array

- **SKA-LOW** (50-350 MHz) : 131072 log périodique antennes, réparties dans 512 stations

Distance maximum entre stations: 74 km

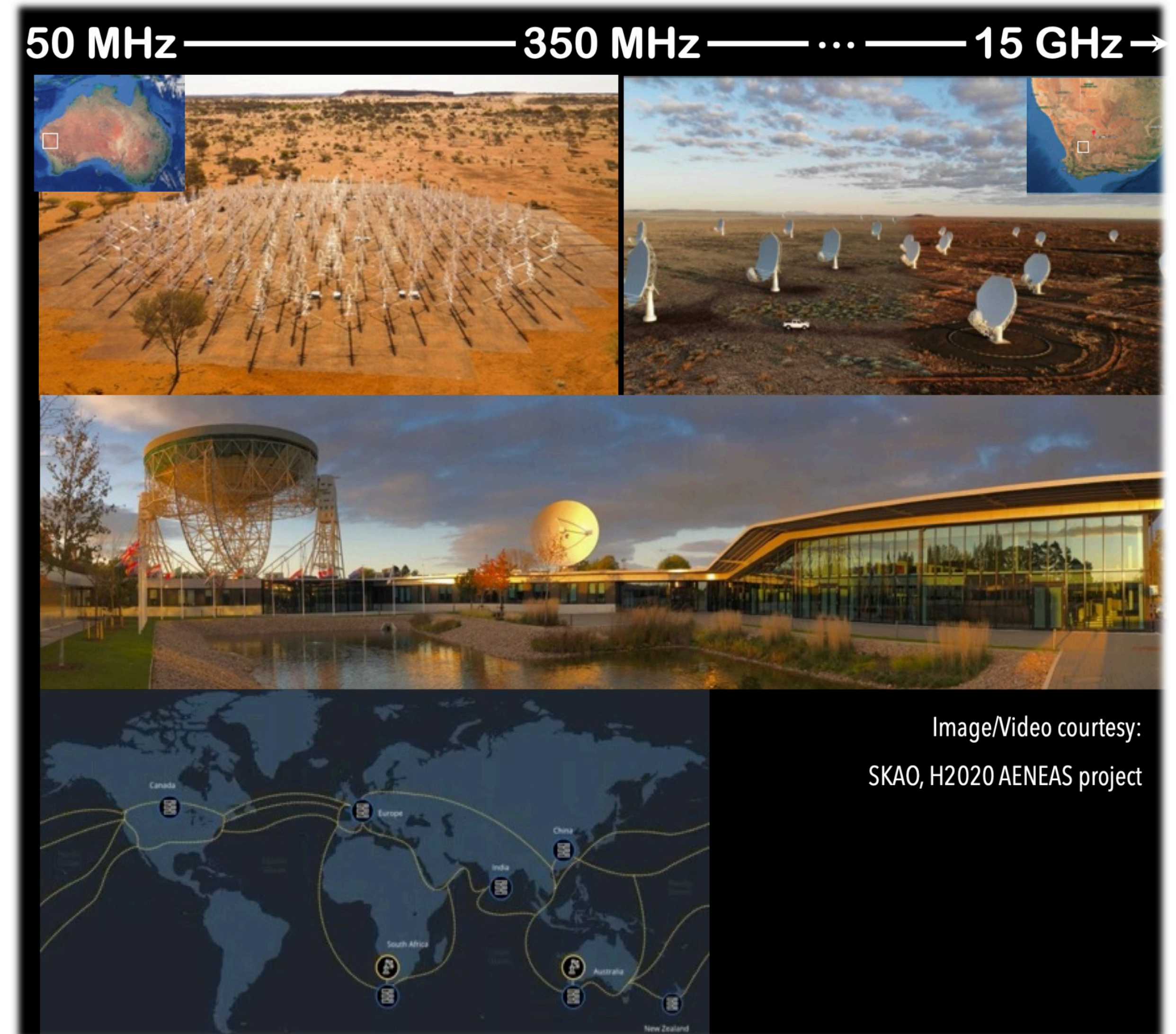
- **SKA-MID** (350 MHz – 15.4 GHz) : 197 antennes entièrement orientables, dont les 64 antennes MeerKAT existantes

Distance maximum entre stations: 150 km

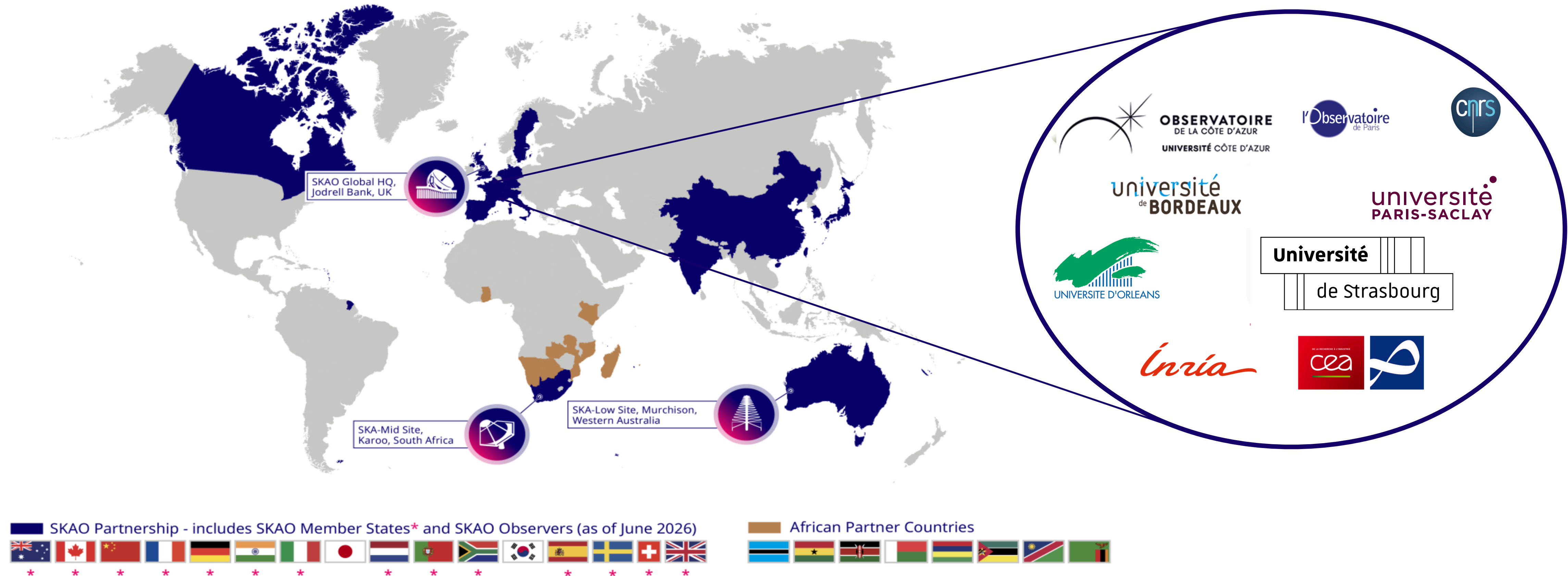
- **SRC-Net** (SKA Regional Center Network) : un réseau mondial de centres de données et de calcul

Par rapport aux observatoires existants :

~ 4 × ~ 8 × ~ 100 ×
Résolution Sensibilité Vitesse de relevé



Square Kilometre Array : Organisation



Une Organisation Intergouvernementale (IGO)

SKAO est une IGO depuis 2021, avec 14 membres
Membres ratifiés: Afrique du Sud, Allemagne, Australie, Canada, Chine, Espagne, **France**, Inde, Italie, Pays-Bas, Portugal, Royaume-Uni, Suisse.

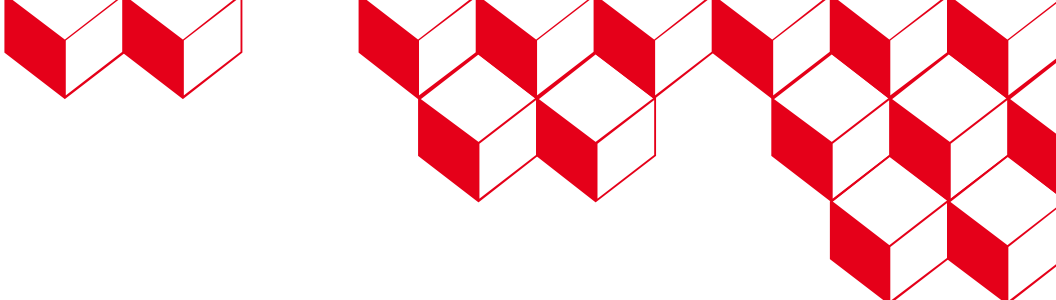
SKA-France

La France est **membre officiel depuis le 4 Juin 2026**.

Directrice : Chiara Ferrari.

Irfu/DEDIP joue un rôle critique dans le déploiement du FR SRC

Square Kilometre Array : Calendrier



- **Objectif** : réaliser le « SKA Baseline Design » (AA4)
- **Financement** : tout n'est pas encore sécurisé, d'où l'application du « Staged Delivery Plan » (AA*)
- **Étape initiale** : développer le plus tôt possible une démonstration fonctionnelle de l'architecture et de la chaîne d'approvisionnement (AA0.5 → July 2026)
- **Étape suivante** : maintenir une installation opérationnelle et en expansion continue, démontrant les performances complètes du design SKA
- **AA2**: le SKAO devient l'observatoire radio le plus puissant sur Terre → Début des “Science Verification”

Milestone Event (earliest)		SKA-Mid (date)	SKA-Low (date)
AA0.5	4 dishes 6 stations	2026 July	2025 July
AA1	8 dishes 18 stations	2028 Mar	2026 June
AA2	64 dishes 64 stations	2029 Mar	2027 June
AA*	144 dishes 307 stations	2030 Jul	2028 May
AA4	197 dishes 512 stations	TBD	TBD

Crédits: SKAO

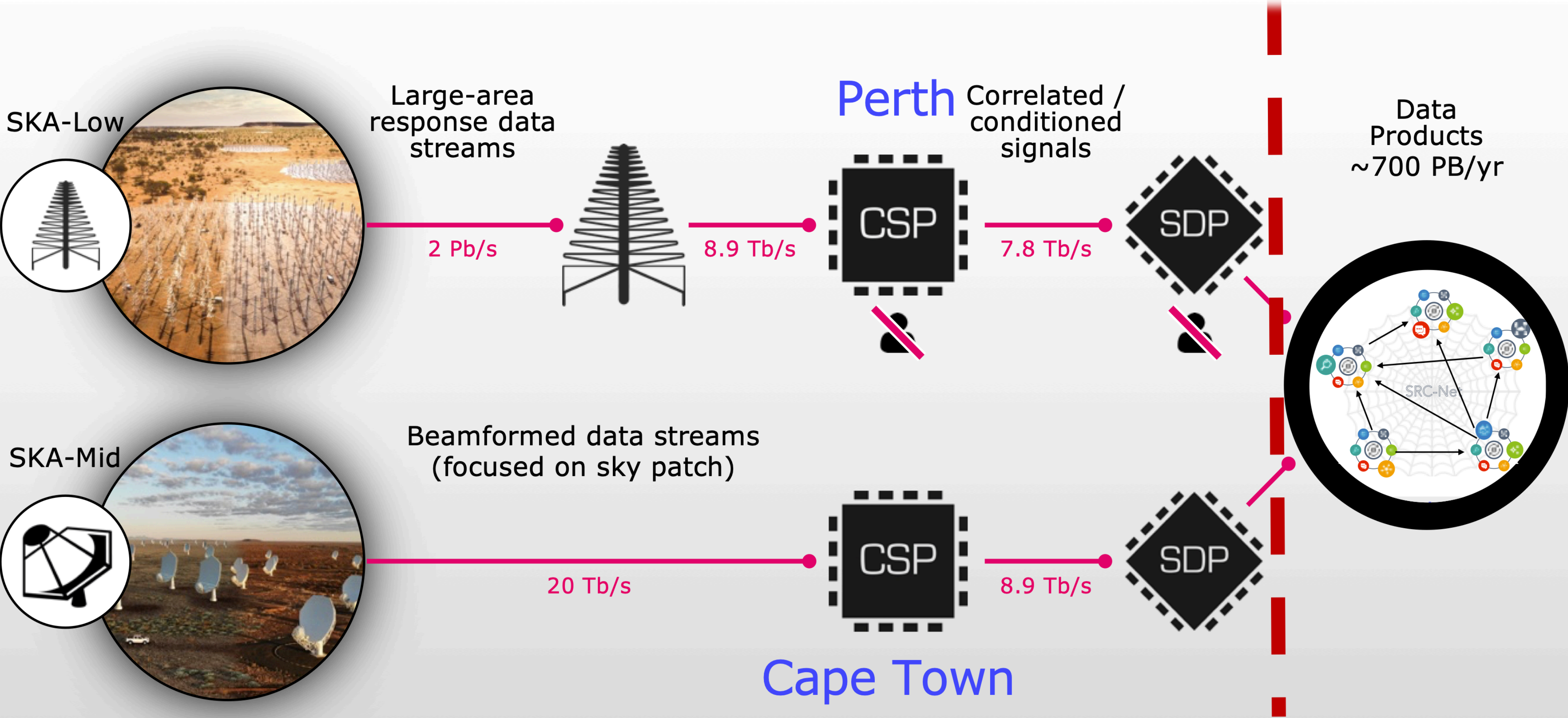


Square Kilometre Array : le Mur des Données !

SKAO

SRC-Net

non financé par SKAO. Collaboration internationale basée sur les contributions des pays membres



Archivage et distribution: Stockage des ~700 PB/an et accès transparent pour la communauté.



Calcul et Analyse: Fournir la puissance de calcul nécessaire à l'analyse scientifique (~35 PFlops)

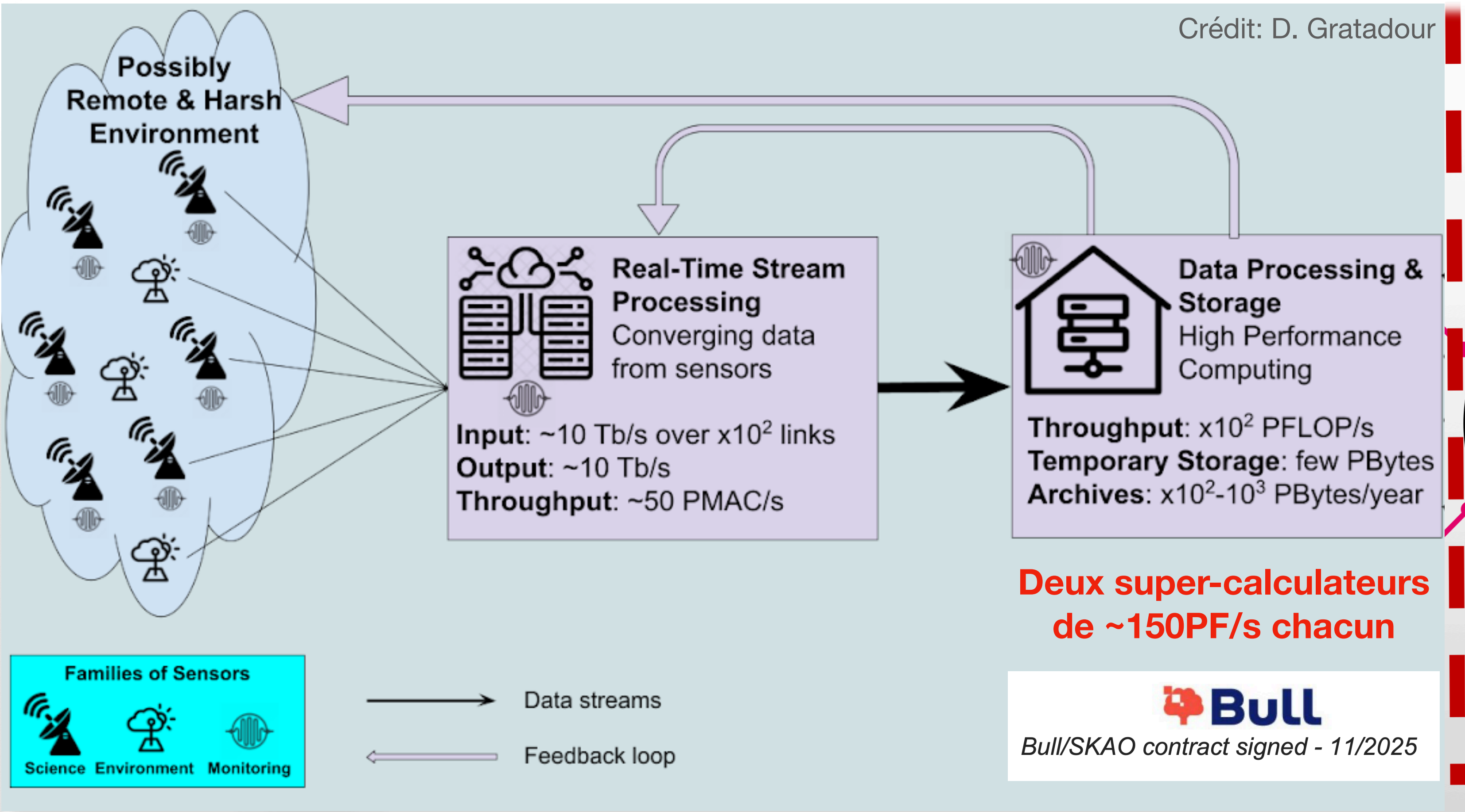


Support et Outils : Offrir un support, des workflows et des outils d'analyse et de visualisation

Slide courtesy: SKAO

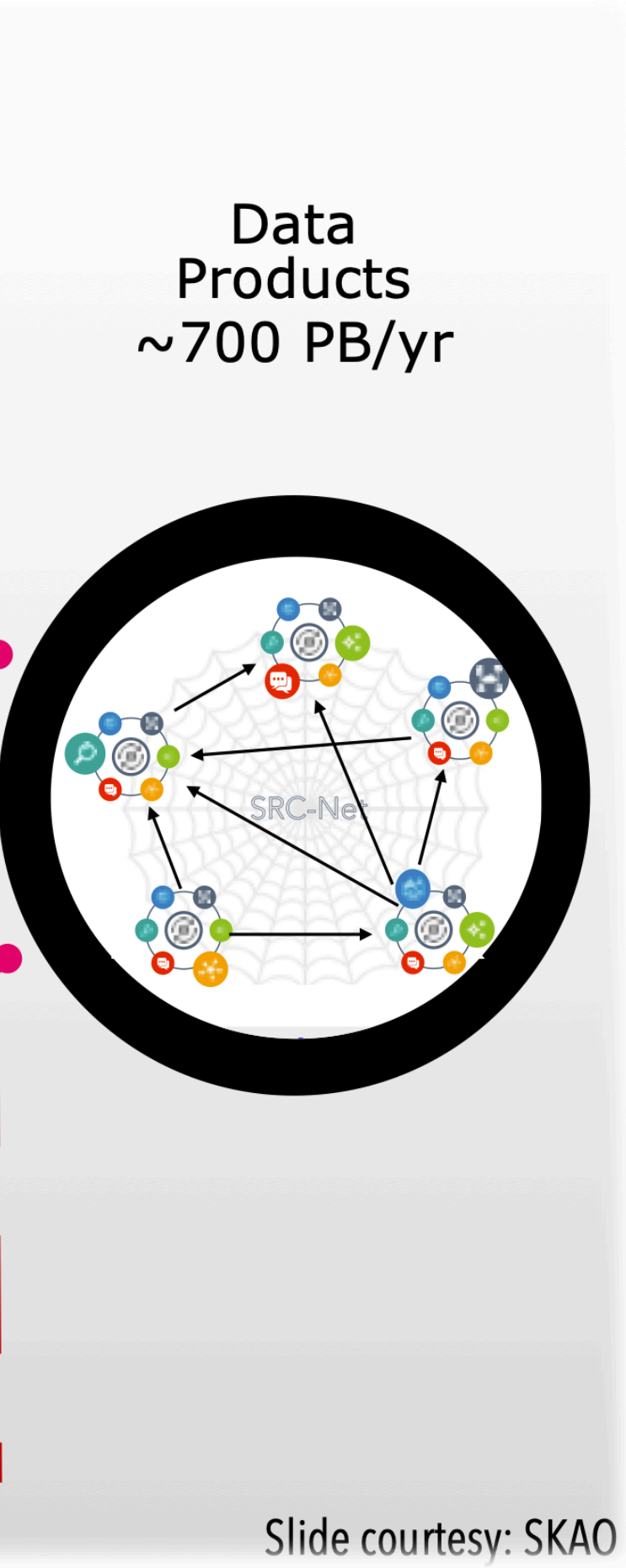
Square Kilometre Array : le Mur des Données !

SKAO



SRC-Net

non financé par SKAO. Collaboration internationale basée sur les contributions des pays membres



- Archivage et distribution:** Stockage des ~ 700 PB/an et accès transparent pour la communauté.
- Calcul et Analyse:** Fournir la puissance de calcul nécessaire à l'analyse scientifique (~ 35 PFlops)
- Support et Outils :** Offrir un support, des workflows et des outils d'analyse et de visualisation

Défi 1 : Volume & temps réel : réduire avant de stocker

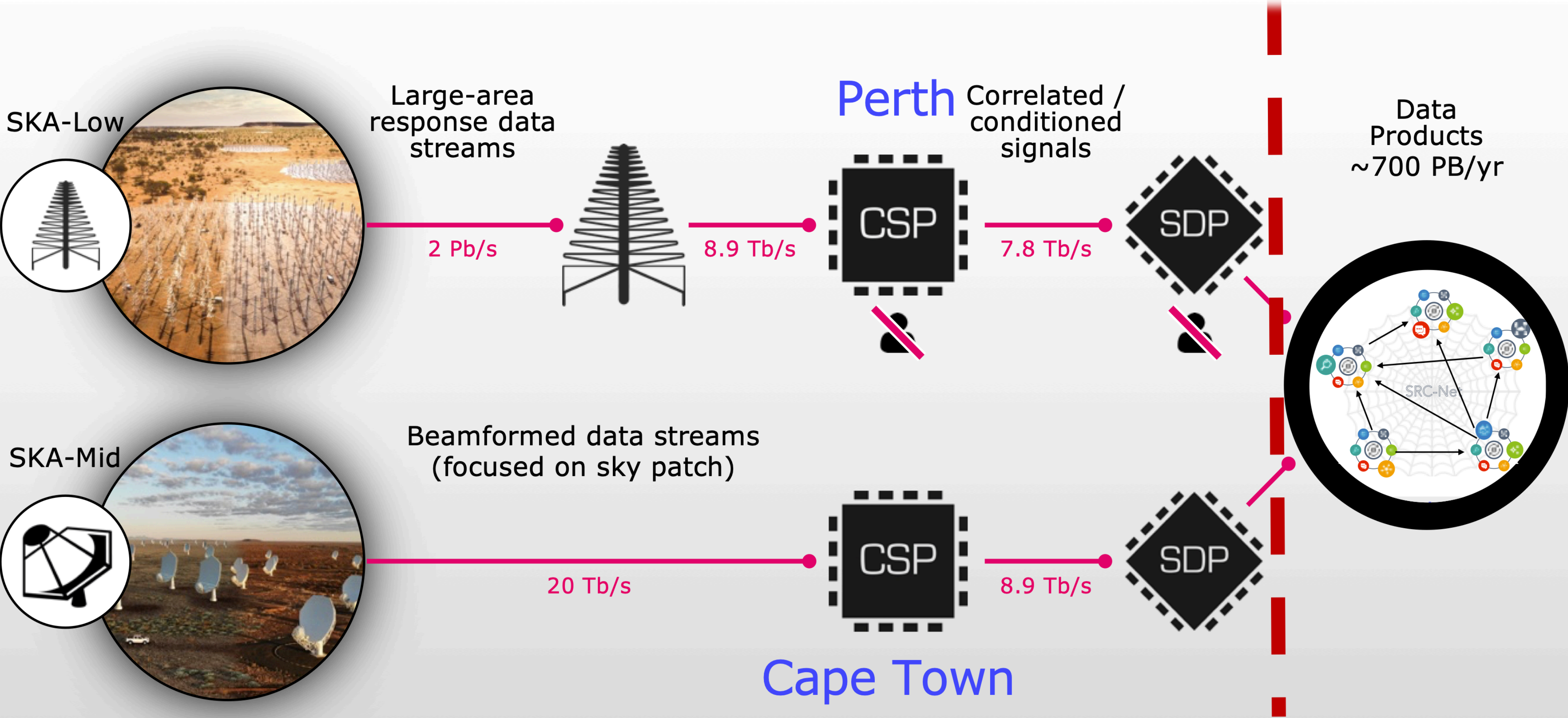
Corrélation, calibration, et reconstruction d'image se font au niveau des CSP et SDP

Square Kilometre Array : le Mur des Données !

SKAO

SRC-Net

non financé par SKAO. Collaboration internationale
basée sur les contributions des pays membres



Archivage et distribution:
Stockage des ~700 PB/an
et accès transparent pour
la communauté.



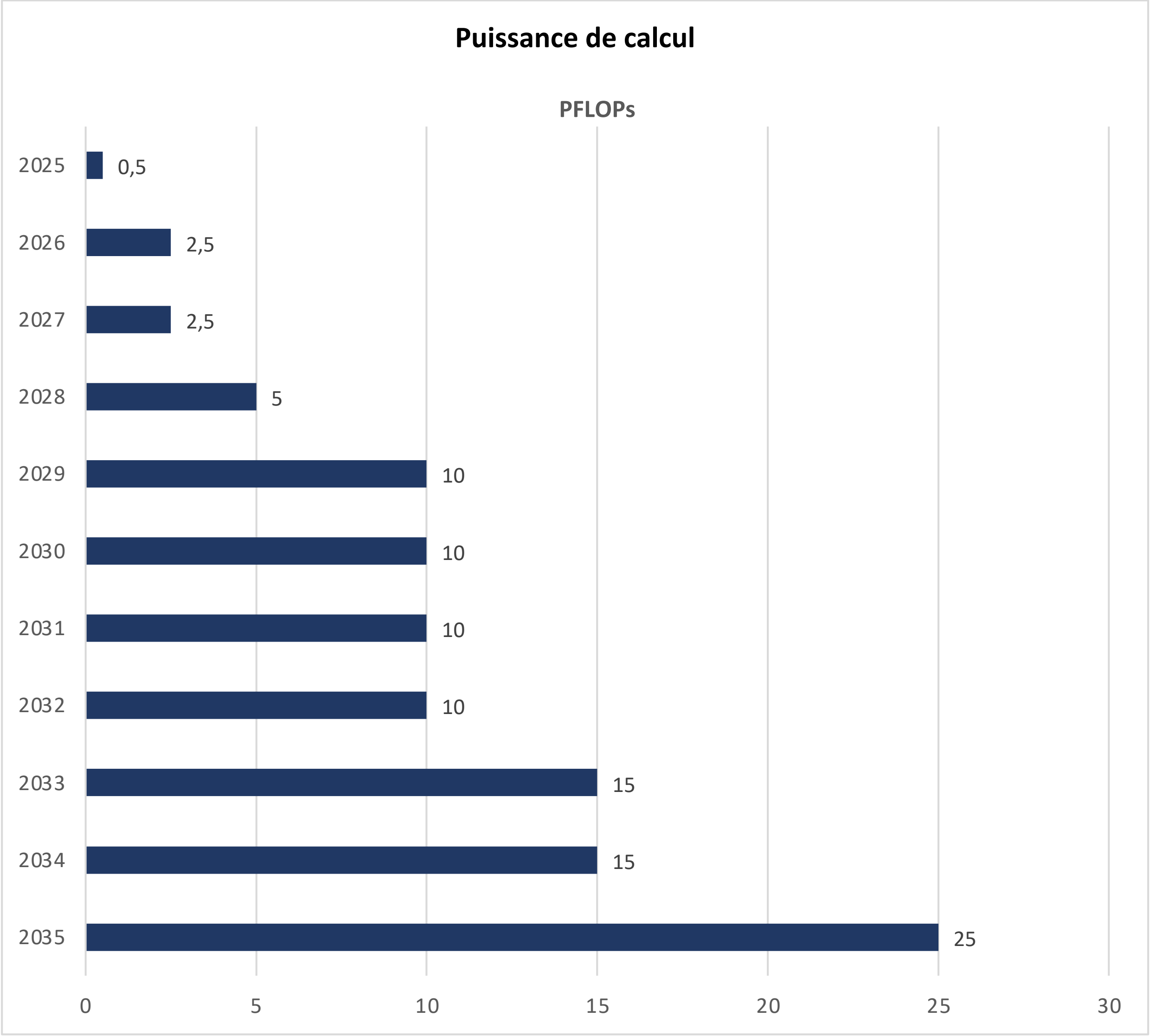
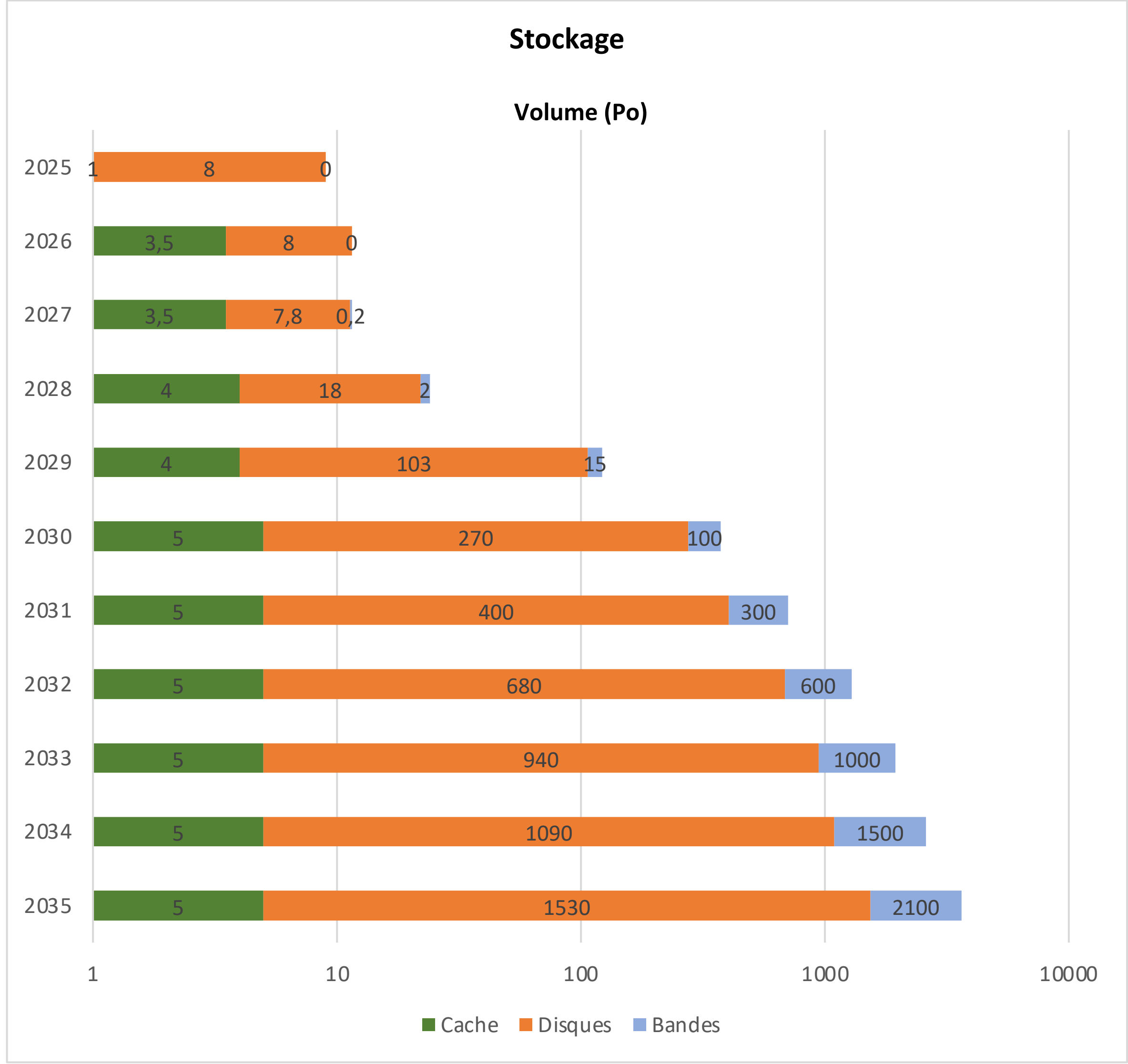
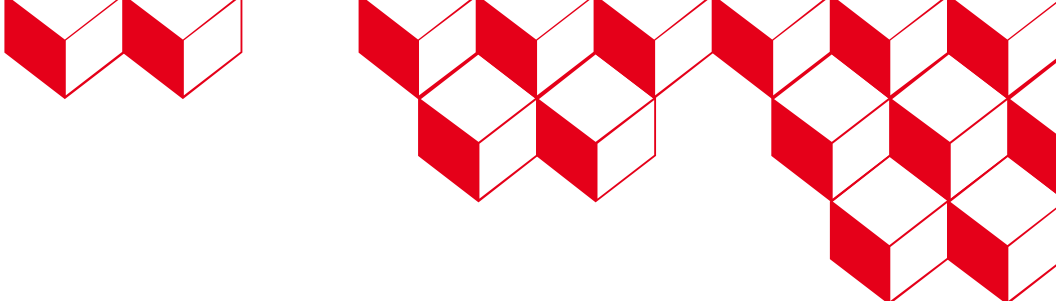
Calcul et Analyse: Fournir
la puissance de calcul
nécessaire à l'analyse
scientifique (~35 PFlops)



Support et Outils : Offrir
un support, des workflows
et des outils d'analyse et
de visualisation

Slide courtesy: SKAO

L'Infrastructure du SRCNet : Volumétrie



♦ Contraintes majeures :

- Les volumes de données à l'échelle exa sont trop importants pour être centralisés.
- Le calcul doit s'effectuer au plus près de la donnée : « déplacer le logiciel vers la donnée » plutôt que l'inverse.
- Les flux scientifiques sont complexes (analyse de données à grande échelle, HPC, IA).

♦ Implications pour l'infrastructure :

- Des **centres hétérogènes et distribués**, disséminés géographiquement sous gouvernance locale : des SRCNets locaux répartis sur 14 pays.
- **Coordonnés** par des standards communs (API) et des services globaux (identité, distribution des données) : le SRCNet global.
- Un **écosystème fédéré**, où les services globaux et les services locaux doivent s'inter-opérer de façon transparente.

Défi 2 : Calcul exascale & architecture mondialement distribuée

Déploiement du FR SRC

Site A : Mésocentre Paris-Saclay



Marco LEONI
Baptiste SOULLARD



David CORRE
Aurélien BAILLY-REYRE
Damien CHAPON
(+ embauche NumPEX)



Michel JOUVIN
Guillaume PHILIPPON



Pile logicielle SRCNet

Ressources allouées (12/2025)

64 vCPUS (128 GB RAM)

10 To stockage (disques)

Lien réseau Univ. Paris-Saclay : 6x10GBPS partagés

Services locaux SRC à déployer

- Virtualisation du stockage : **Rucio Storage element**
- Reverse proxy : **Gatekeeper**
- Visualisation : **SODA**
- Monitoring : **perfSonar**
- Mise sur étagère des données dans un espace de travail utilisateur : **prepareData**
- Plateforme Scientifique : **CANFAR (CADC)**
- Exploitation scientifique interactive : **JupyterHub**

Potentiels sites futurs

Mésocentre Meso-PSL (Meudon)



Collaboration avec Nançay pour traiter des données de Nenufar/LOFAR 2.0

IDRIS (GENCI)

Étude de faisabilité d'utilisation de centre national (en ZRR) pour le FRSRC par Numpex EXA-ATOW (F. Bodin)



01 : Volume & temps réel : réduire avant de stocker

→ Relevé par le CSP et le SDP : la réduction commence dès corrélateurs jusqu'aux super-calculateurs.

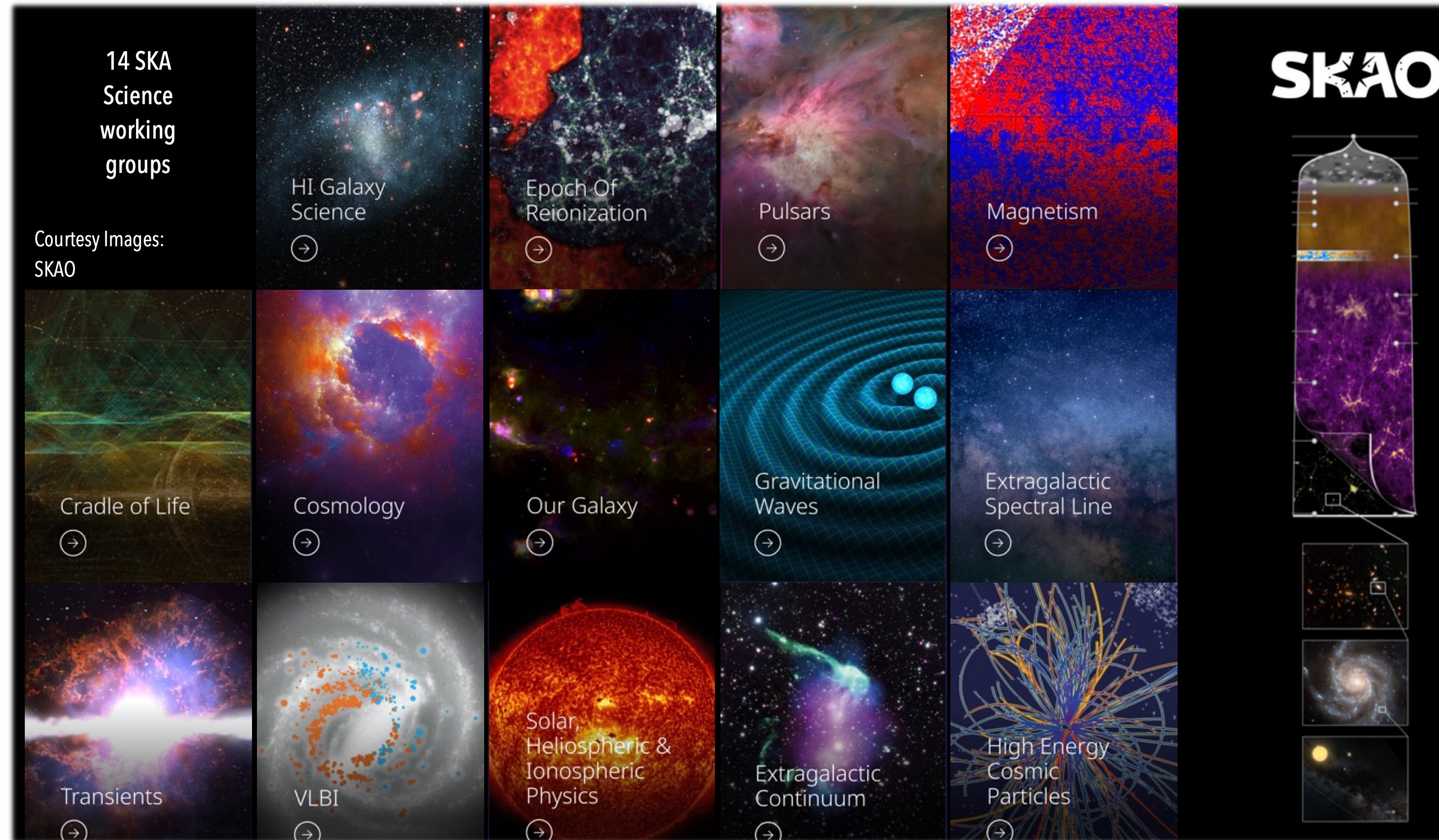
02 : Calcul exascale & architecture mondialement distribuée

→ Relevé par le déploiement complexe du réseau SRCNet, qui fédère la donnée et son analyse.

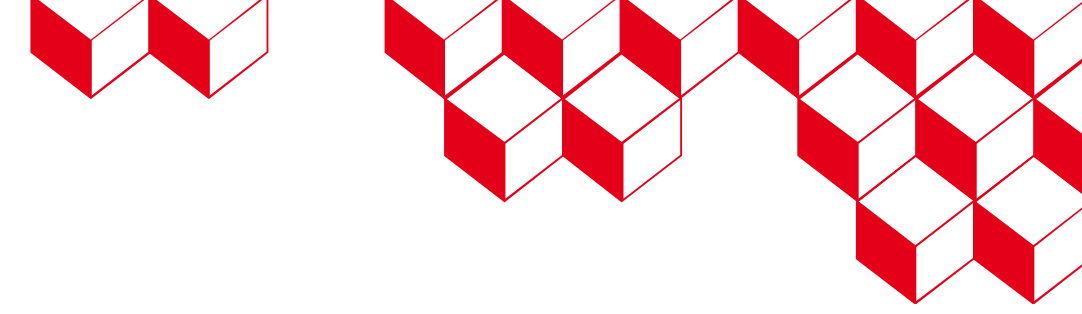
03 : De la donnée à la découverte: multitude de pipelines, IA/ML, et déploiement

→ Relevé par l'Astronome, mais sans professionnels du HPC à ses côtés, il se heurte à un mur logistique

De la Donnée à la Découverte : Multitude de Pipelines



De la Donnée à la Découverte : Multitude de Pipelines



♦ Pulsars

- A. Petiteau et al. — ASTRA project — développement et déploiement des pipelines PTA

♦ Transitoires

- F. Schüssler et al. — Astro-COLIBRI — Développement des stratégies de détection, de communication et de réponse aux transitoires du SKA

♦ Continuum extragalactique

- E. Daddi, B. Magnelli et al. — CIGALE-2D — Développement de méthodes d'ajustement de SED 2D

♦ Systèmes Planétaires – Formation et Magnétisme

- P. Hennebelle, H. Le Breuilly et al. — ECOGAL ; Galactica — Développement de simulation sur la formation d'étoiles et des disques protoplanétaires ; réalisation d'"observation SKA"
- A. Strugarek — ExoMagnet — Développement de simulation sur les magnétosphères exoplanétaires et leur détectabilité avec SKA

♦ Epoque de la réionisation

- S. Farrens, F. Lanusse, J.-L. Starck et al. — TITAN project — Développement de méthodes de séparation de composantes, nettoyage des avant-plans et inférence

♦ Cosmologie

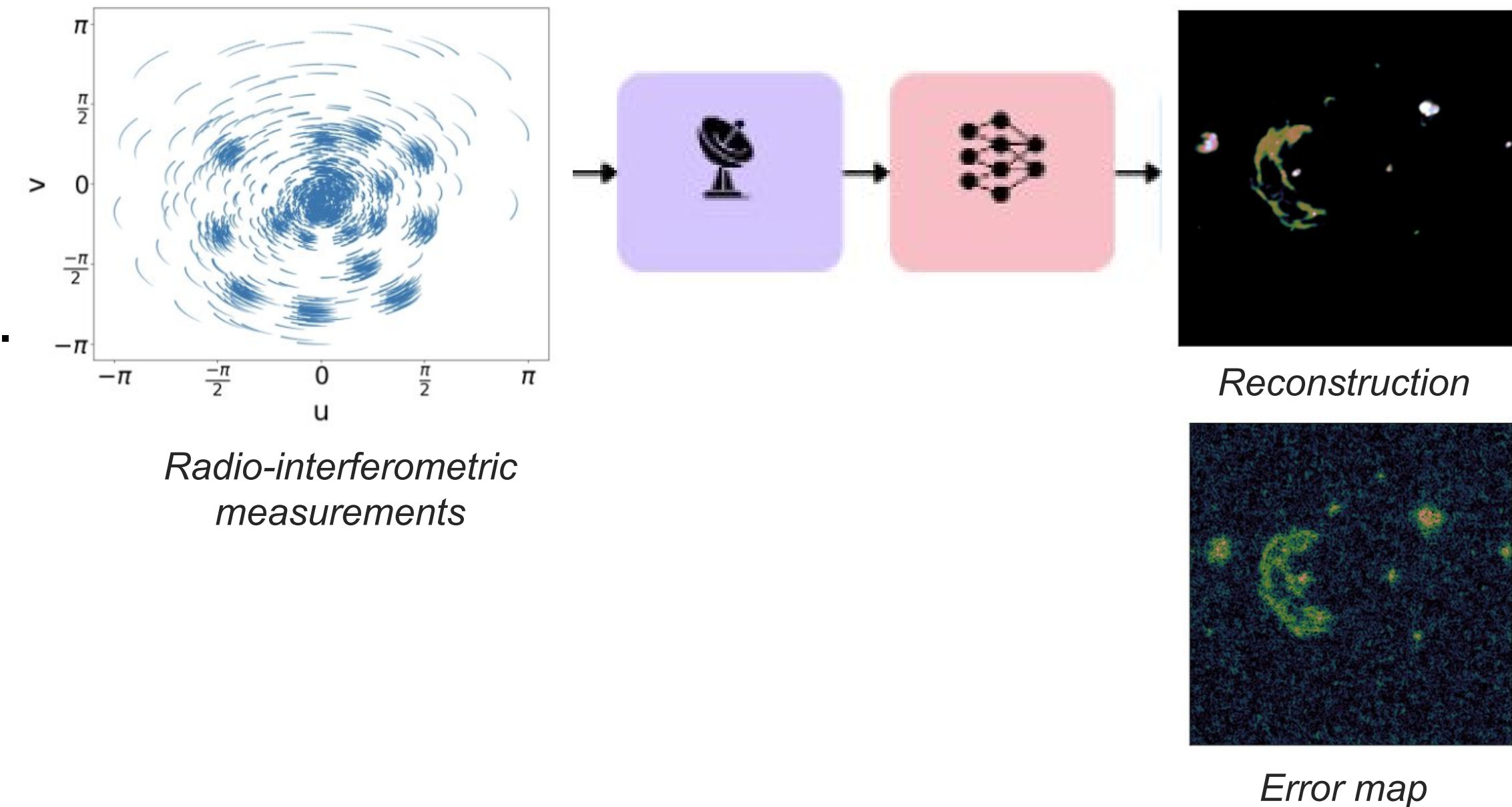
- R. Ansari, S. Farrens, F. Lanusse, J.-L. Starck et al. — TOSCA project — Développement de méthodes génératives d'inférence du shear, et reconstruction d'image ; cartographie d'intensité de HI

♦ ...

Rendre les pipelines IA prêt pour la science...

- ML pour une reconstruction d'image précise
- « physics-informed » : modèle instrumental, inférence, etc.
- Quantification validée des incertitudes

—> Transverses aux domaines scientifiques de l'Irfu



... à l'échelle exascale

- Forte implication dans les travaux de NumPEX sur les workflows couplant HPC et IA
- Développement d'analyses fondées sur le ML, avec un focus sur la reconstruction d'image, au coeur de la problématique SKA, avec passage à l'échelle exascale
- Calcul distribué (multi-GPU) avec des débruiteurs fondés sur le ML (ex. modèles de diffusion, DRUNet, etc.), construits sur DeepInverse



Implication de Tobías Liaudat

De la Donnée à la Découverte : Déploiement

Des scientifiques qui ne sont pas des experts en HPC ou en cloud computing se heurtent à un « mur logistique » qui freine la découverte scientifique, malgré la disponibilité des ressources de calcul

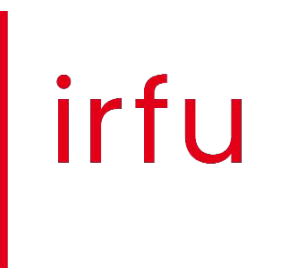
ASTRA project (2026-2030)



LPENS
LABORATOIRE DE PHYSIQUE
DE L'ÉCOLE NORMALE SUPÉRIEURE

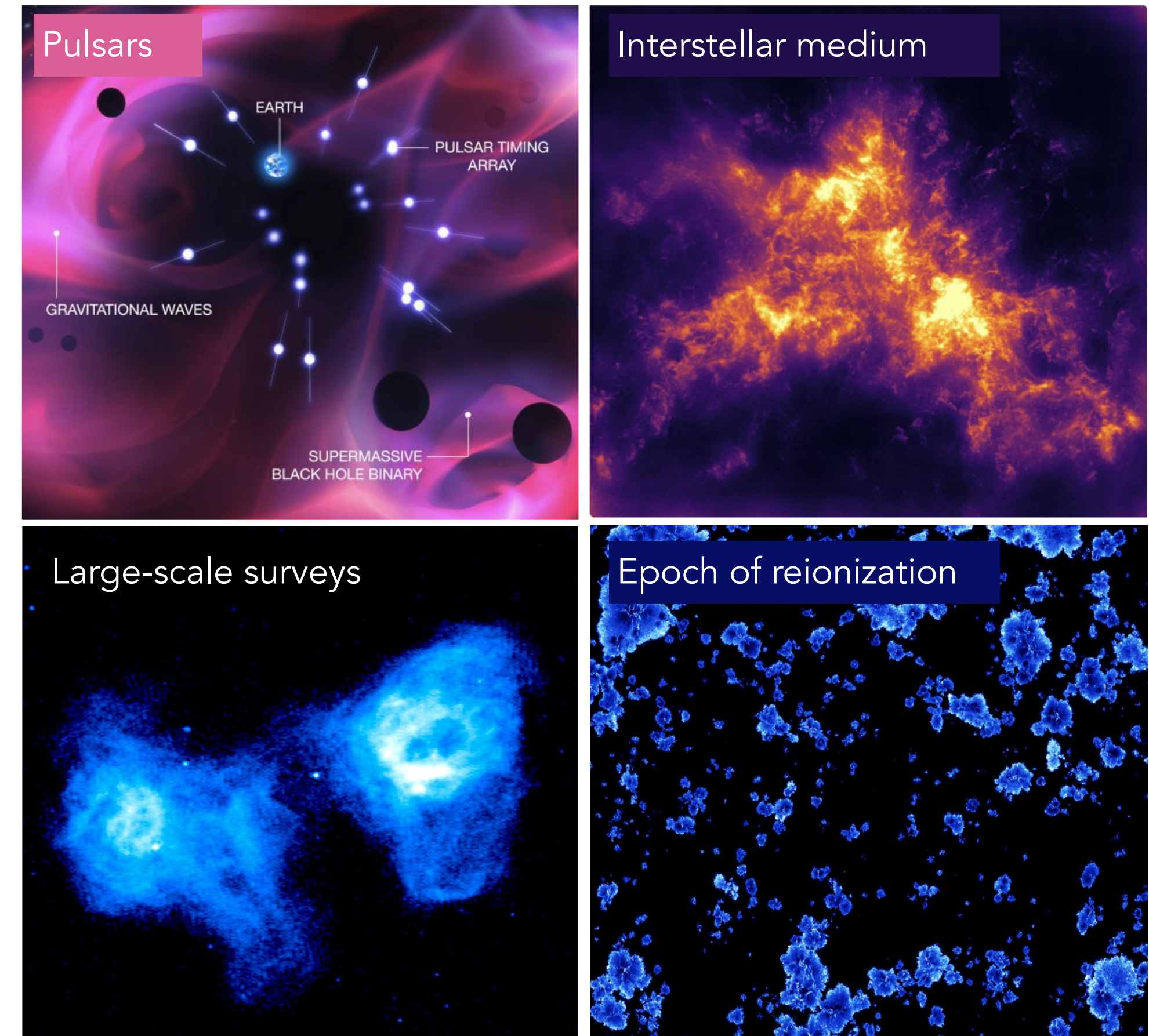


IRISA

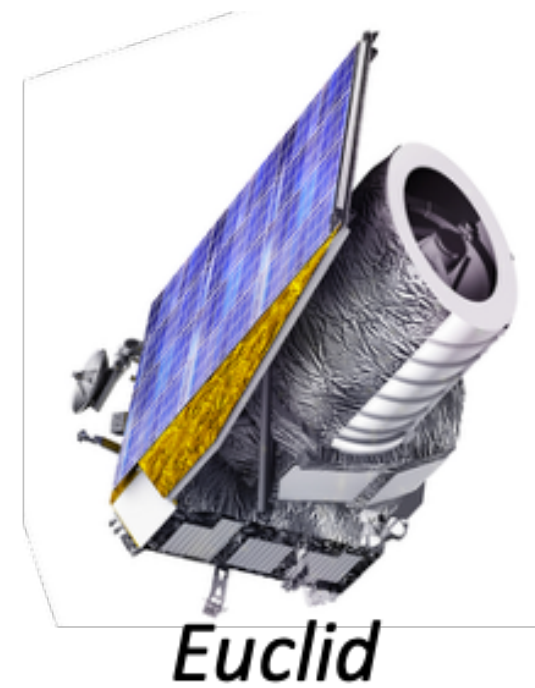
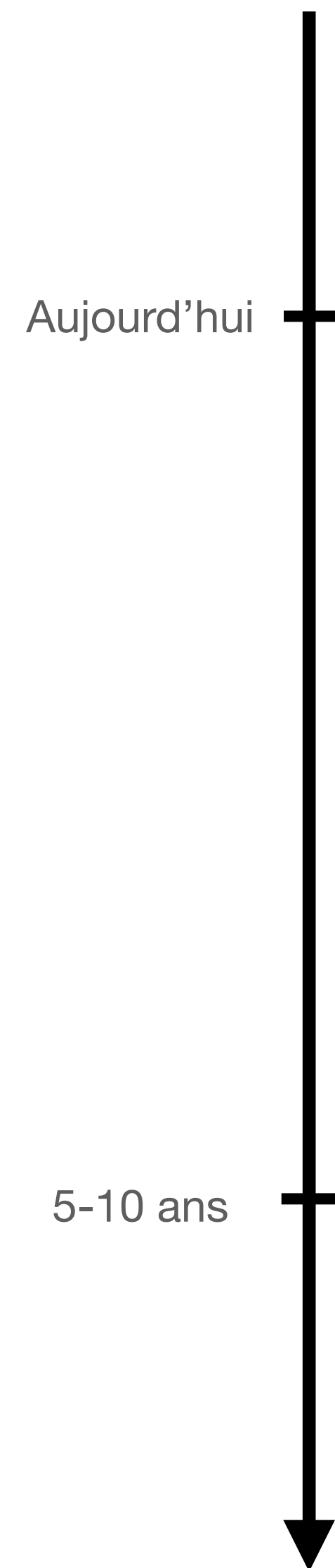


- **Objectifs** : un travail centré sur 4 pipelines matures de radioastronomie, visant à fournir des workflows déployables sur des architectures distribuées et hétérogènes, similaires à celles qui seront mises en œuvre pour le SKA.
- **Méthode** : fournir un cadre générique pour le développement de workflows, intégrant la fédération des services de stockage, de calcul et d'exécution des données, ainsi que l'accès aux ressources (authentification, logistique des données, etc.).
- **Démonstrateur de déploiement sur le premier nœud du FR-SRC.**

Implication de Antoine Petiteau, Damien Chapon



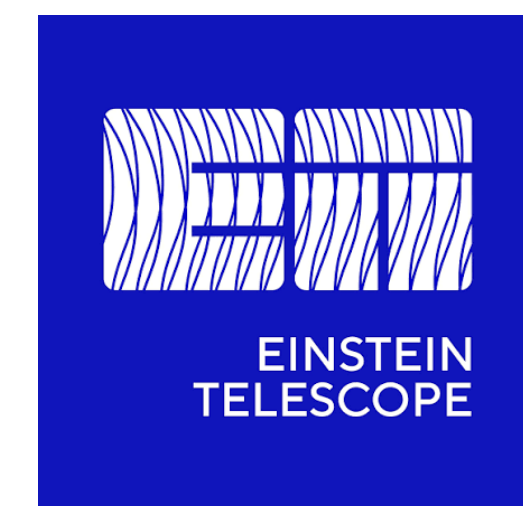
La Prochaine Vague d'Observatoires HPDA



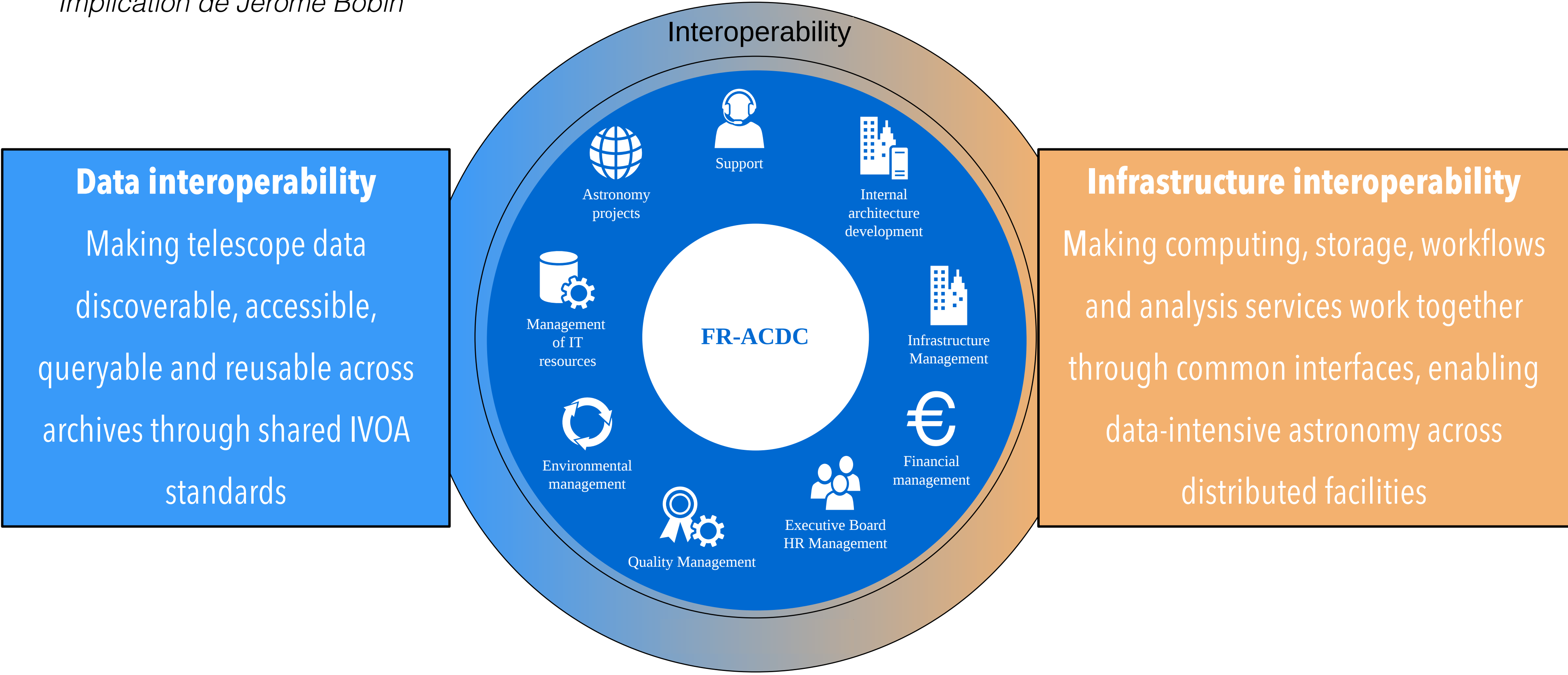
Euclid



WST — ALMA2040 — AtLAST



Implication de Jérôme Bobin



Crédit: C. Ferrari

Besoin d'un accès à des serveurs locaux performants de niveau intermédiaire (« middle tier »)

- Cartes profondes, interférométrie, polarimétrie, spectroscopie, imagerie haut-contraste (JWST, ALMA, VLT, NOEMA, PRIMA, ELT) : des volumes de données limités, qui ne relèvent pas d'un véritable HPDA, mais qui nécessitent des serveurs de stockage (plusieurs centaines de To) et de calcul intermédiaire, en local.
- Développer et valider, sur données de test, nos pipelines scientifiques avant leur déploiement à l'échelle du HPDA.
- La généralisation des méthodes d'apprentissage statistique et de traitement des données, pour l'analyse standard des données, entraîne une généralisation des besoins en GPU.
- Un accès local et réactif, sans les contraintes des grands centres nationaux (e.g., projet de courte durée, pas besoin de transférer des données lourdes...).
- La question du stockage froid local des données se pose également.

La mise en place et la maintenance de ces serveurs locaux restent cruciales



♦ 1- Le vrai HPDA : professionnalisation et industrialisation de nos pipelines d'analyse

- **Infrastructure** : Calcul et stockage sur des moyens partagés et distribués, au niveau national (ex. ACDC) et international (ex. SRCNet) — Pipelines conteneurisés et bâtis sur des standards ouverts
- **Moyen Humain** :
 - Participation active au développement de ces infrastructures partagées, qui mobilisent de nombreux profils aux expertises variées (HPC, cloud, données)
 - Aide à la professionnalisation des pipelines d'analyse, au passage à l'échelle des méthodes d'IA et au déploiement optimal des chaînes d'analyse scientifique sur ces infrastructures partagées

♦ 2- Couche de calcul intermédiaire

- **Infrastructure** : Moyens de calcul et de stockage intermédiaires, pour développer et valider nos pipelines d'analyse, et pour les analyses de fin de chaîne ou les projets observationnels plus modestes
- **Moyen Humain** :
 - Mise en place et maintenance de cette couche intermédiaire

♦ 3- Assistant IA ?

- Pour rester compétitifs dans nos développements et nos analyses, un accès à des moyens LLM performants semble nécessaire — Moyens locaux (MAIA ?) performants, au niveau CEA, DRF ou Irfu ?

Conclusions

- ♦ **Le SKA va produire une quantité massive de données, avec des défis de calcul considérables.**
 - > La réduction des données commence dès le capteur : l'edge computing n'est plus optionnel.
- ♦ **Le calcul et le stockage des données des très grands observatoires sera distribué : bâtir cette infrastructure exige un large éventail d'expertises (CI/CD, HPC/IA distribuées, clusters Kubernetes, orchestration de services, fédération d'identités...).**
 - > Cultiver l'expertise technique nécessaire pour maintenir, faire évoluer et exploiter ces infrastructures de calcul et de stockage distribuées.
- ♦ **Dans ces conditions, le retour scientifique de l'analyse des données à l'échelle exa repose sur la maîtrise de toute la chaîne de calcul, de la conception des workflows jusqu'à leur déploiement :**
 - ♦ Adopter des processus et des outils de développement professionnels, fondés sur des standards communautaires ouverts, garants de la reproductibilité et de la pérennité des pipelines ;
 - ♦ Accélérer le développement et le passage à l'échelle de l'IA
 - > Cultiver une relation étroite entre les équipes scientifiques, les développeurs professionnels et les centres de calcul (mésocentres, GENCI).



Merci !

CEA SACLAY
91191 Gif-sur-Yvette Cedex
France
Benjamin.Magnelli@cea.fr
+ 33 1 69 08 68 25