

Calcul et analyse à l'IN2P3

- Rôle de l'IN2P3 dans le calcul LSST et DESC
- Proposition d'organisation du groupe calcul



Réunion de collaboration LSST-France

APC

18-19 décembre 2013

Dominique Boutigny
CNRS-CC-IN2P3 / SLAC



Vers une contribution de l'IN2P3 au Data Release Processing

Lettre de Jacques Martino au management de LSST – 7/12/2011

*"...While the details of our participation remain to be negotiated, we are baselining a plan to participate in the annual data release processing. In this scenario, we will negotiate with the French NREN Renater in order to provide the required communications bandwidth between the US and Europe. We will also contribute processing hardware and labor necessary to process **approximately 50% of the data** at our CC-IN2P3 computing facility in Lyon, France. The entire data set will be transferred from the LSST Archive Center at NCSA to CC-IN2P3, where 50% will be processed using the Data Management System Data Release Production software, and the processed data will then be transferred back to NCSA. The other 50% of the processed data will be transferred from NCSA to CC-IN2P3, so that both centers can merge the parts and **end up with a full data release**. We have made an estimate of the required hardware contribution based on the Data Management Infrastructure Sizing Model in the above scenario and can commit to provide the necessary resources. We understand that the infrastructure, software, and processing parameters must be maintained consistent with the overall Data Management System design..."*

Jeu complet de données au CC-IN2P3

- Accès facilité aux données
- Impact fort sur notre capacité à exploiter scientifiquement les données



Split DRP

Pas mal d'étapes à passer pour concrétiser cet engagement

- ✓ Convaincre LSST de la capacité de l'IN2P3 à remplir cet engagement
 - Le Data Challenge de cet été était un test crucial → 😊
 - Capacité à s'organiser
 - Mobilisation des ressources nécessaires
 - Mise en place d'un environnement de production à partir de 0
 - Capacité d'adaptation du CC à diverses contraintes apparues en cours de route
 - ...
- ✓ Faire en sorte que LSST confirme sa volonté de faire aboutir le modèle du Split DRP
 - Accord pour ne pas présenter ce modèle lors de la Final Design Review (pas de rôle officiel de l'IN2P3 dans la phase construction du DM) → Le modèle est construit comme si le NCSA assurait l'ensemble du DRP.
 - Inquiétudes levées suite à une discussion avec Steve Kahn qui charge officiellement Jeff Kantor de mettre en place un MOA (board LSSTC d'avril) suivi d'un MOU spécifiant le rôle de l'IN2P3 dans le DRP → 😊 (même si du coup ça va un peu vite !)



Split DRP

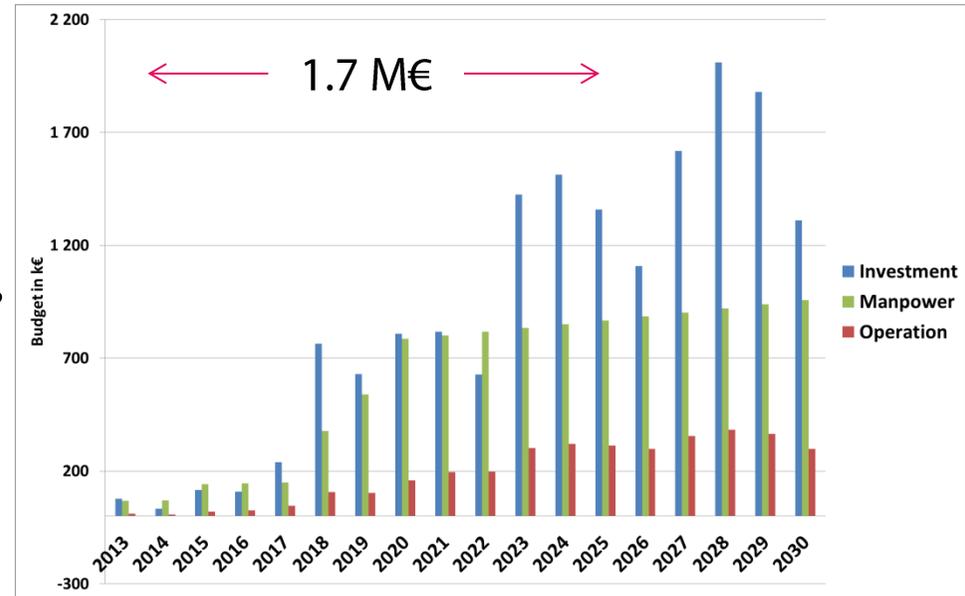
- ✓ Le plus difficile est maintenant d'obtenir une confirmation de l'engagement de l'IN2P3 pour le financement de l'opération

Contribution de ~30 M€ (total) / 16 M€ (hardware) entre maintenant et 2030

Le CC a identifié une ligne budgétaire ~150 k€ (si le budget est confirmé) pour LSST dès 2014 😊

La contribution à la caméra nous assure déjà l'accès aux données

- Il faut montrer que la contribution au calcul donne à l'IN2P3 un avantage décisif au niveau scientifique



➔ Besoin d'un plan cohérent pour l'ensemble du calcul LSST-France (pas seulement le Split DRP)



Remarques

La "construction" du Data Management est intégralement financée par la NSF

- Cadre extrêmement contraint
- Si une tâche de construction est remplie par un contributeur non NSF, LSST doit rembourser la somme correspondante !
- Modèle de calcul rigide et très "conservateur" / prudent
 - Il a fallut démontrer que l'on serait capable aujourd'hui de traiter les données
 - Très peu de place pour l'innovation et la R&D

L'IN2P3 ne peut pas apparaître explicitement dans les tâches de construction
Mais cela peut être vu comme un avantage ...

- Pas/peu de contrainte par rapport projet (pas d'engagement de construction)
- Capacité à expérimenter et proposer des alternatives
- Possibilité de concentrer l'effort pour maximiser le retour scientifique

Mais c'est dangereux : Nécessite une organisation cohérente et focalisée pour éviter la dispersion des efforts



Remarques (suite)

La montée en charge de l'infrastructure de calcul au CC doit se faire progressivement et en minimisant les à coups

Implications :

- Participation régulière aux data challenges
- Contribution à des plateformes de test comme celle pour Qserv
- Articulation du calcul avec la préparation de la science
- Contribution à des tâches de fond comme la production MC
 - Peu de manpower
 - "Exerce" l'infrastructure
 - Contribution visible



Financement

Le budget nécessaire est important

Il faudra certainement aller chercher de l'argent en contribuant à des projets débordant du cadre strict de LSST

- Direction de l'IN2P3 – Chargé de mission pour le calcul – Directeur du CC-IN2P3 – Chargé de mission au CNRS (Michel Daydé) - Chargé de mission au Ministère pour le Calcul Intensif (Mark Asch) ...

Europe :

- Horizon 2020

CPER

ANR

Appel d'offre du G8

- Réunion en marge de Supercomputing pour la mise en place d'une collaboration France / US

autres appels à projets ...

Va certainement nécessiter de mettre en œuvre des collaborations avec la recherche en informatique (à la Petasky)

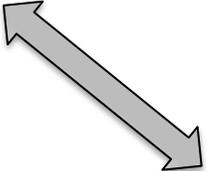
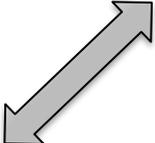
Question récurrente du mariage des données LSST + Euclid

- à manipuler avec précaution !

Le développement d'une infrastructure de traitement des données Astro en France profite à tout le monde...



SLAC



Bases de données

- Qserv
- XLDB
- Collaboration IT

DESC "Science Center"

- Environnement calcul / données pour DESC
- Logiciels / outils spécifiques
- Centre de compétence

Calcul LSST-France

Environnement logiciel

- Code
- Documentation
- Qualité
- Training

Precursor datasets (SDSS – CFHT – DES...)

- Data challenge
- Exploitation scientifique en liaison avec DESC

Split DRP

- Construction de l'infrastructure
- Tests / validations
- R&D

LSST Europe



Concrètement ... pour 2014

Plateforme Qserv

(faisabilité technique et financière en cours d'évaluation au CC)

- 50 machines – 8 à 16 To / machine
- Contribution cruciale et visible dans le projet

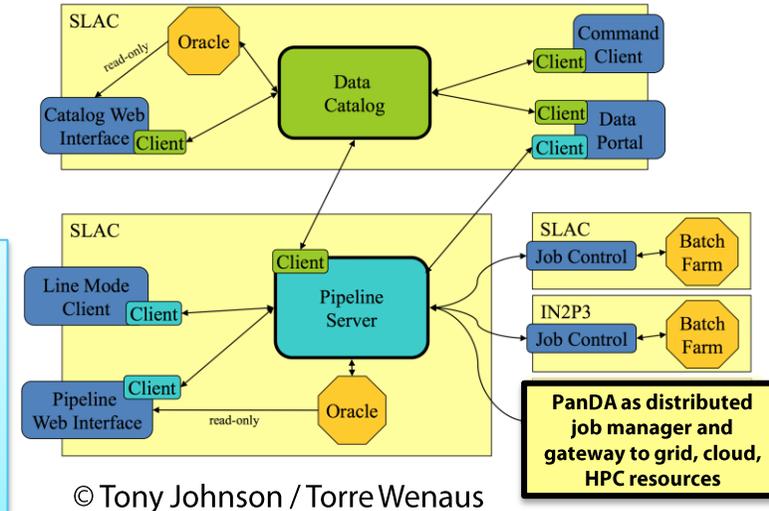
Data Challenge

- Proposition de LSST : reprocessing complet de SDSS/DR10
- Alternative : retraitement des données CFHTLS
 - Plus proche de LSST
 - Plus intéressant scientifiquement (?)
 - Expertise locale

Il faut pouvoir exploiter scientifiquement ces données

- Validation par rapport au traitement officiel CFHT
- Utilisation ciblée en rapport avec la préparation de la science : photo-z – effet chromatique sur le weak lensing ...

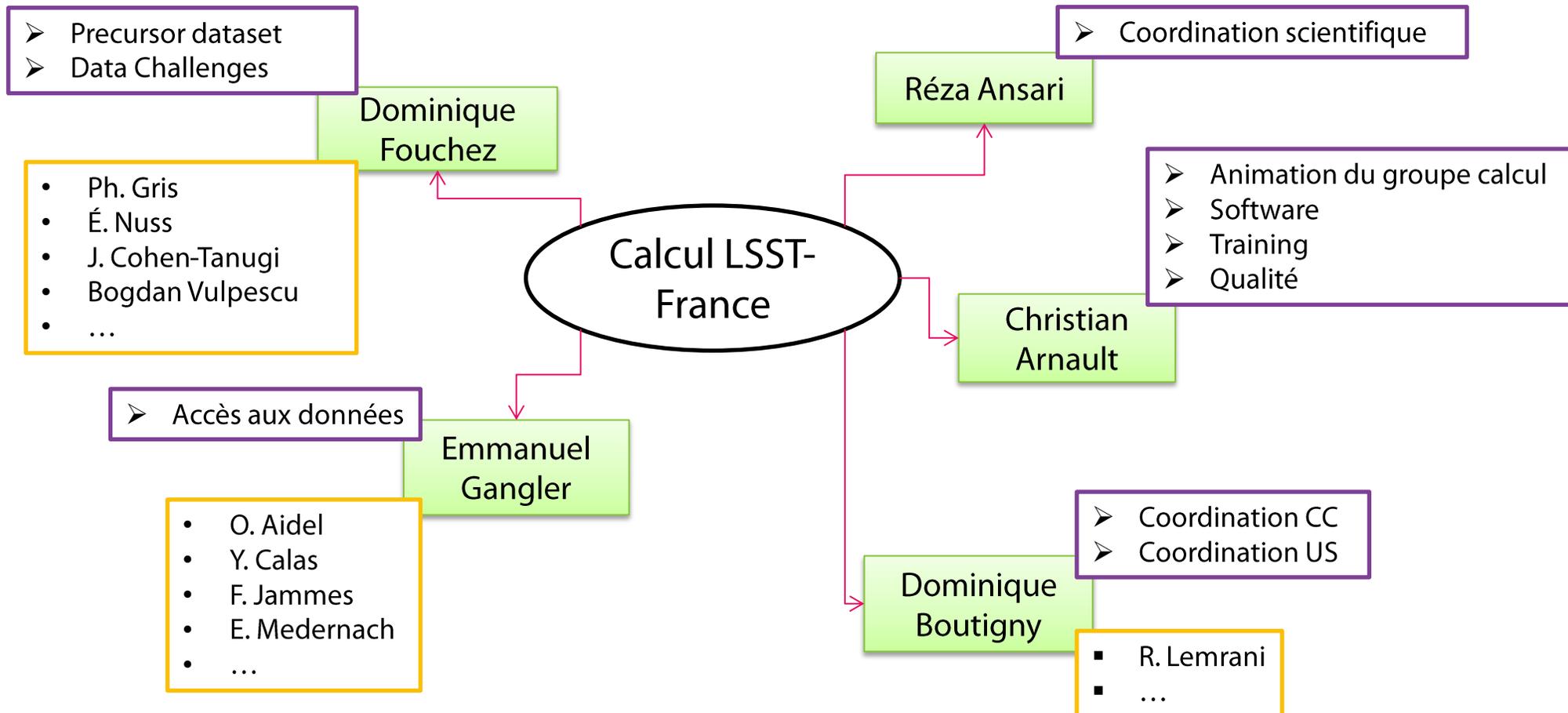
Fourniture de cycles CPU pour ImSim



Piste de collaboration avec SLAC



Proposition d'organisation



Pourrait aussi être complété par une activité de développement dans ImSim ...