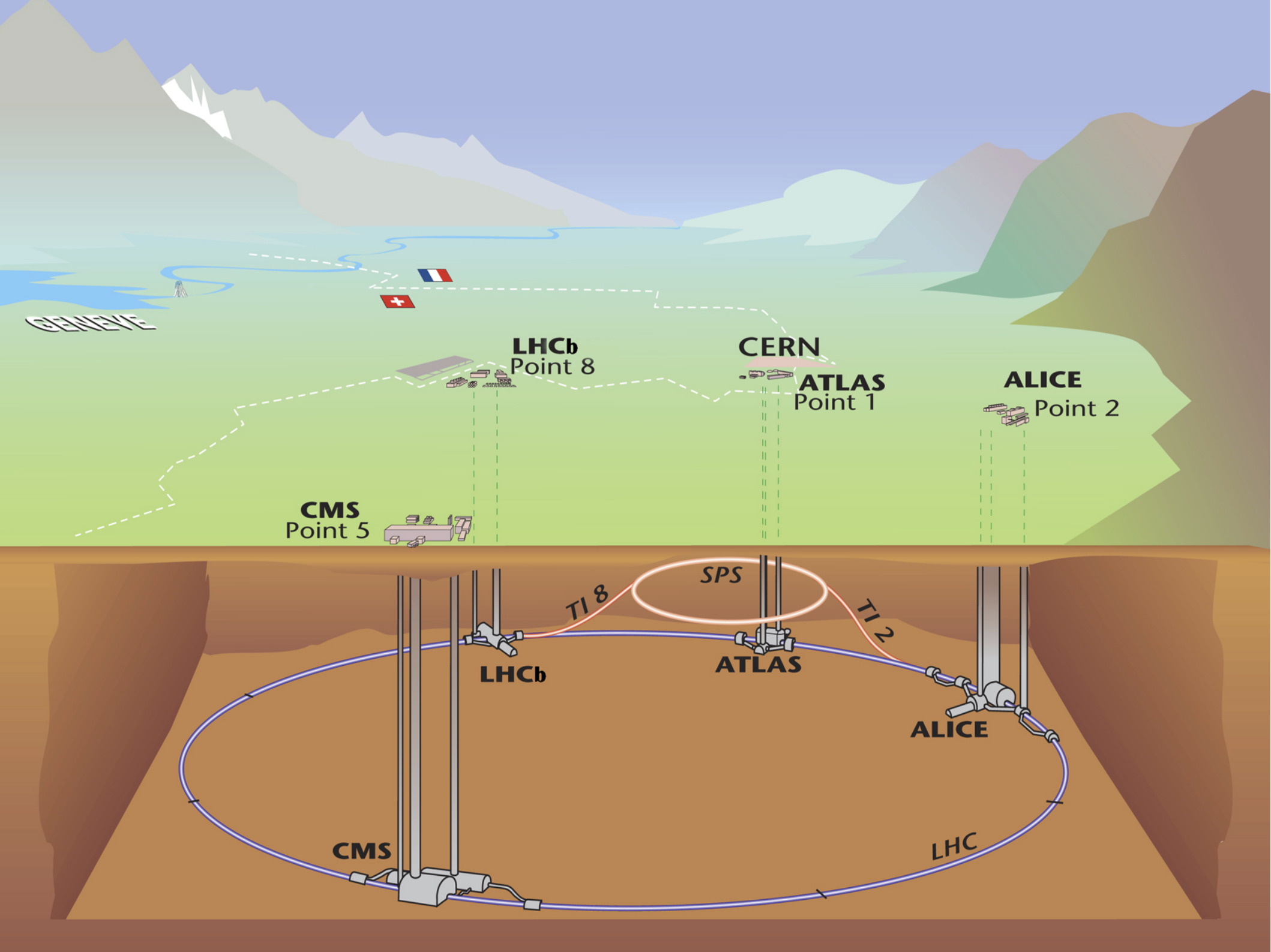
Le modèle de calcul LHC et ses évolutions

Journées Plateforme @ Clermont

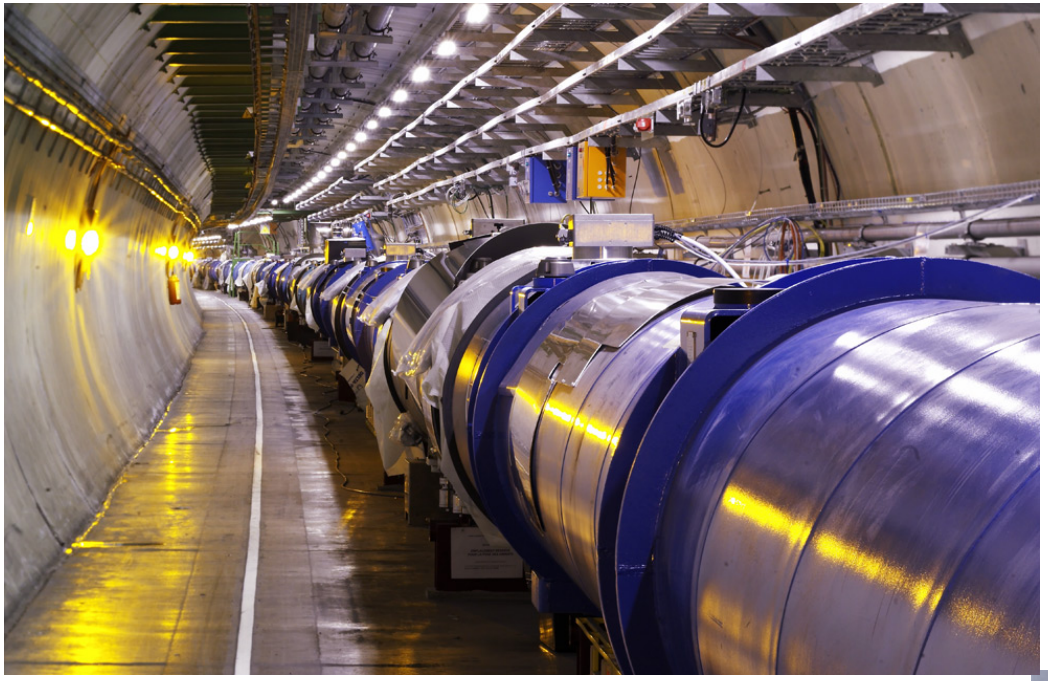
Renaud Vernet

Le LHC


GENEVE
LHCb
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
SPS
TI 8
TI 2
LHCb
ATLAS
ALICE
CMS
LHC

Enjeux du LHC

- Objectif premier
  - HIGGS BOSON → brique manquante du Modèle Standard
    - … jusqu'en 2014
- Processus rares (i.e. intéressants)
  - → collisions : milliards
  - → détecteurs : très précis
  - → Une GRANDE quantité de données à traiter
- Début des opérations en 2010
- 2016
  - Collisions : 1 GHz
  - 25 PB données

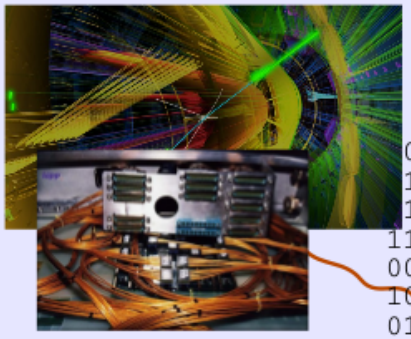


Accélerateur très « lumineux »
→ taux d'interaction très élevé
→ ~ 1 milliard de collisions / s

Détecteurs très complexes
→ enregistrements O(1 GB/s)

Modèle de calcul

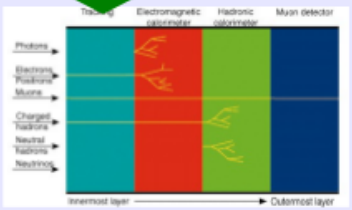
Opérations centralisées



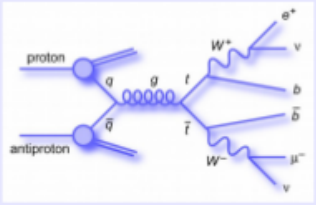
Reconstruction

Analyse

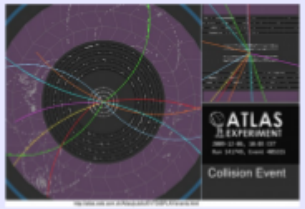
01100 010001
10111 001100
11100 100110
110101 110011
001010 001010
100101 000011
010111 010100

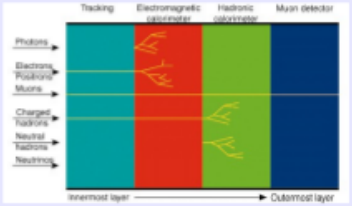


Simulation



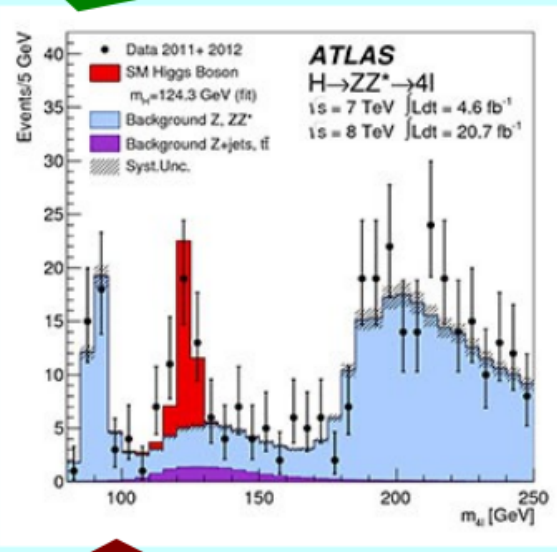
101100 010001
110111 001100
111100 100110
110101 110011
001010 001010
100101 000011
010111 010100





Individuel (chaotique)

Final analysis (n times / day)



Organisation

Naissance de la « grille de calcul »

- Enjeux politiques, et techniques
  - Enormes besoins de CPU et stockage
  - Agences de financement veulent les ressources chez elles

- Eclater les ressources informatiques (~années 2000)
  - Création de la grille de calcul LHC
  - En collaboration avec grille européenne (pas spécifique LHC)

Vision

- Quelques gros datacenters principaux
  - ex. CERN, CC-IN2P3, CNAF…
  - Conservation des données brutes
  - Très bonne connectivité réseau
- Avec nuage de moyens-petits sites
  - Laboratoires, universités
- Et de bonnes interconnexions entre tous les sites
  - NRENs, GEANT, LHCOPN, LHCONE

→ Un ensemble de services informatiques distribués sur la planète et interconnectés par les réseaux ~académiques.

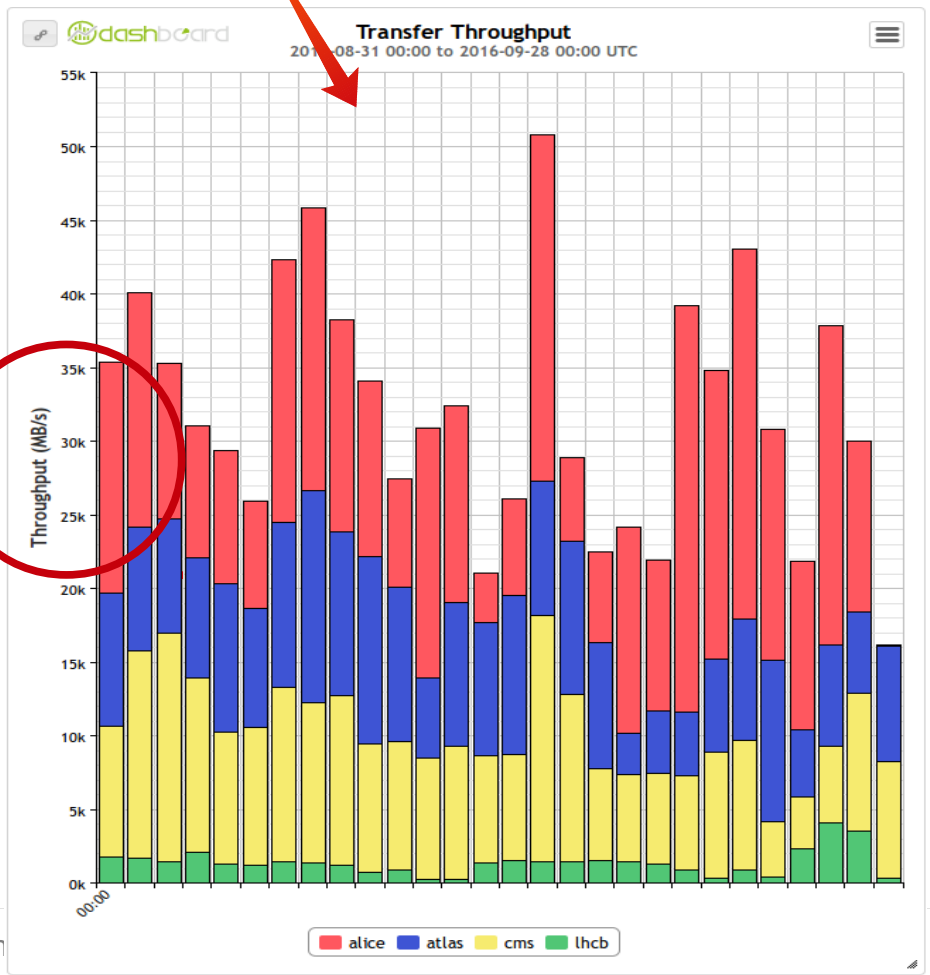
Organisation des sites WLCG

- Tier-0
  - Site central, gros, manpower, disponible, réactif
  - Proche des expériences, bandwidth → CERN
  - Très grosse capacité de CPU et stockage (disque & bandes)
- Tier-1
  - Gros site avec manpower, CPU, disques, bandes
  - Bonne connectivité, disponibilité, réactivité
  - 15 centres dans le monde
- Tier-2, Tier-3
  - Petits sites (exceptions)
  - Fournissent CPU et stockage disque
  - Moins d'exigence sur disponibilité et reactivité

La grille vue du ciel

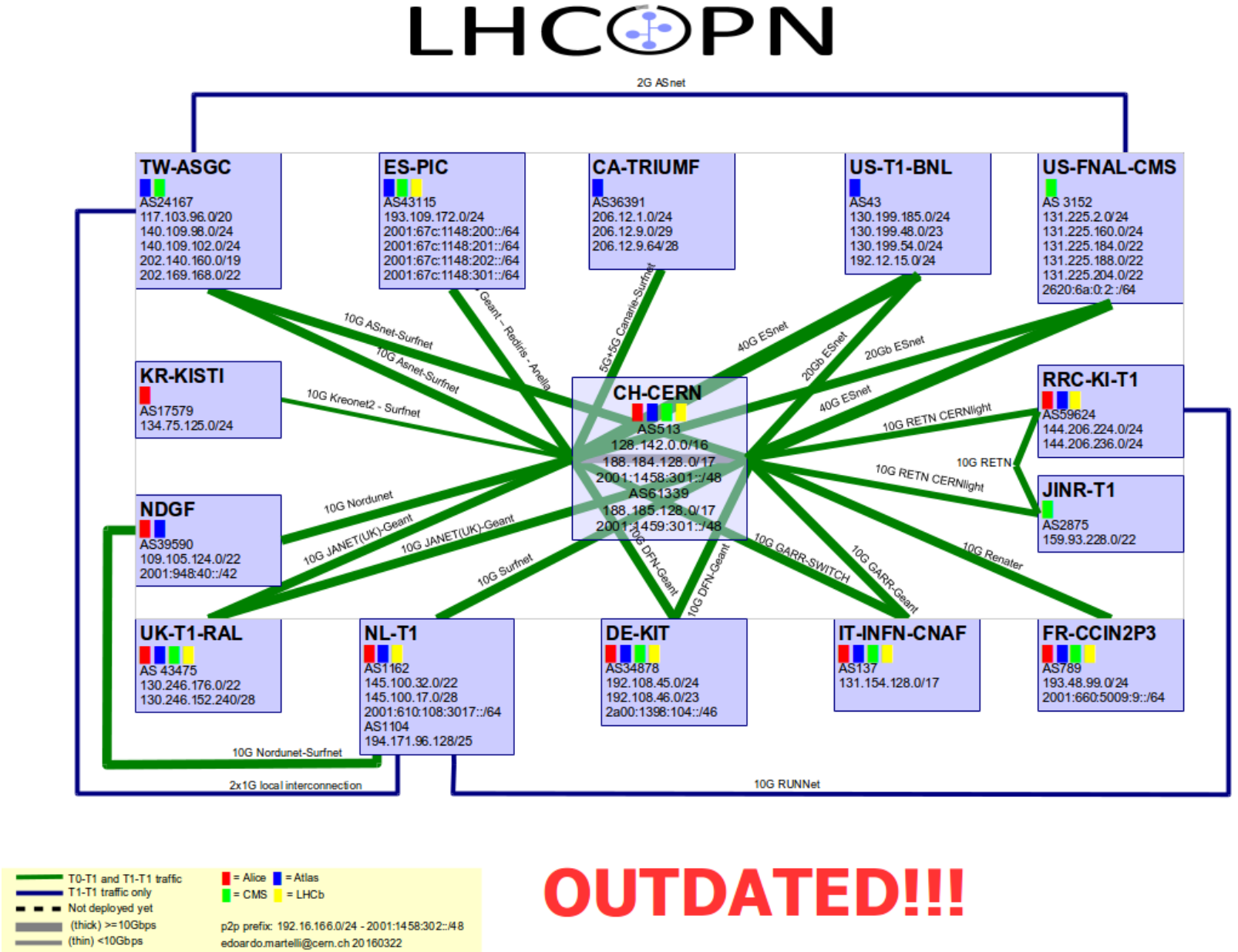
WLCG en quelques chiffres

Transferts reseau	CPUs	Disk storage	Tape storage
30 GByte/s	350k	300 PB	380 PB



- Ressources réparties
  - O(100) sites
- ~350k jobs simultanés

LHC OPN



OUTDATED!!!

SLA et cie

- MoUs
  - WLCG ↔ Agences de financement & sites
- Niveau de service
  - Disponibilité, fiabilité
    - Computing, Réseau, Stockage
  - Réactivité aux incidents

Tout cela est mesuré et public.

- Et beaucoup d'aspects collaboratifs non exprimés en SLA
  - Déploiement nouveaux services, mises à jour OS
  - Migrations (IPv6)
  - Décisions à prendre → consensus

« Contraintes » sur l'infrastructure

- Infrastructure partagée et organisée
  - Un job doit pouvoir tourner partout sur la grille
  - Limiter les efforts d'adaptation aux OS
- Worker nodes
  - EL6 (Scientific Linux 6)
  - Architectures x86_64
- Stockage
  - Moins de contraintes système pour le site
  - Le système de stockage doit juste 'marcher'

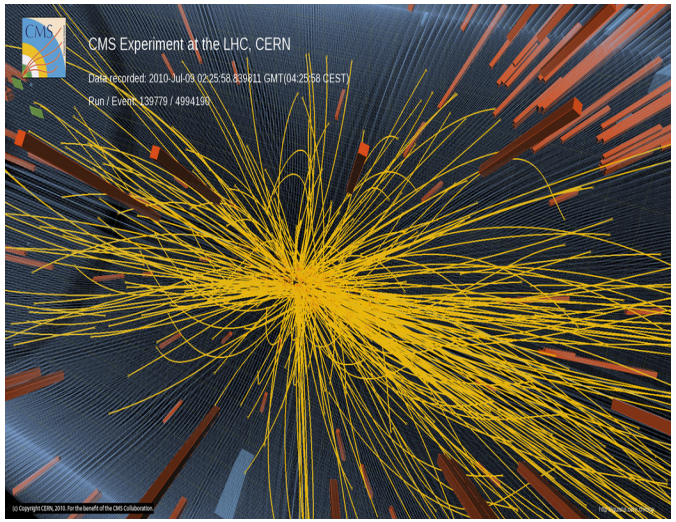
Authentification, autorisation

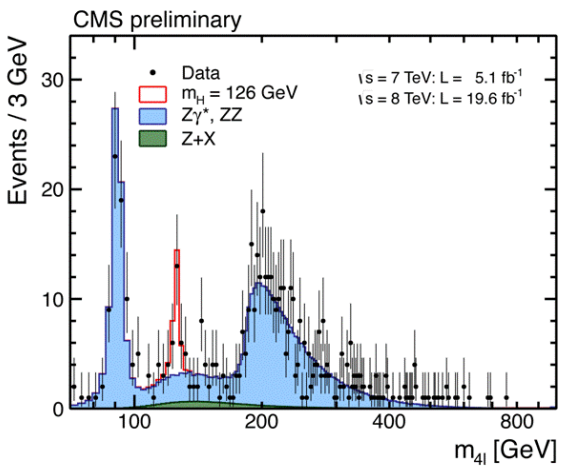
- Tout par certificat
  - Norme X509
  - → grosse organisation et mécanique logicielle
  - → pas facile pour un débutant (!)
- Affecte utilisateurs et (certains) services

- Rôles
  - Personnes → qui peut faire quoi
  - Jobs → permet d'orienter les jobs vers les bonnes queues du batch
    - Du moins au ccin2p3

Modèle de calcul

Traitement des données

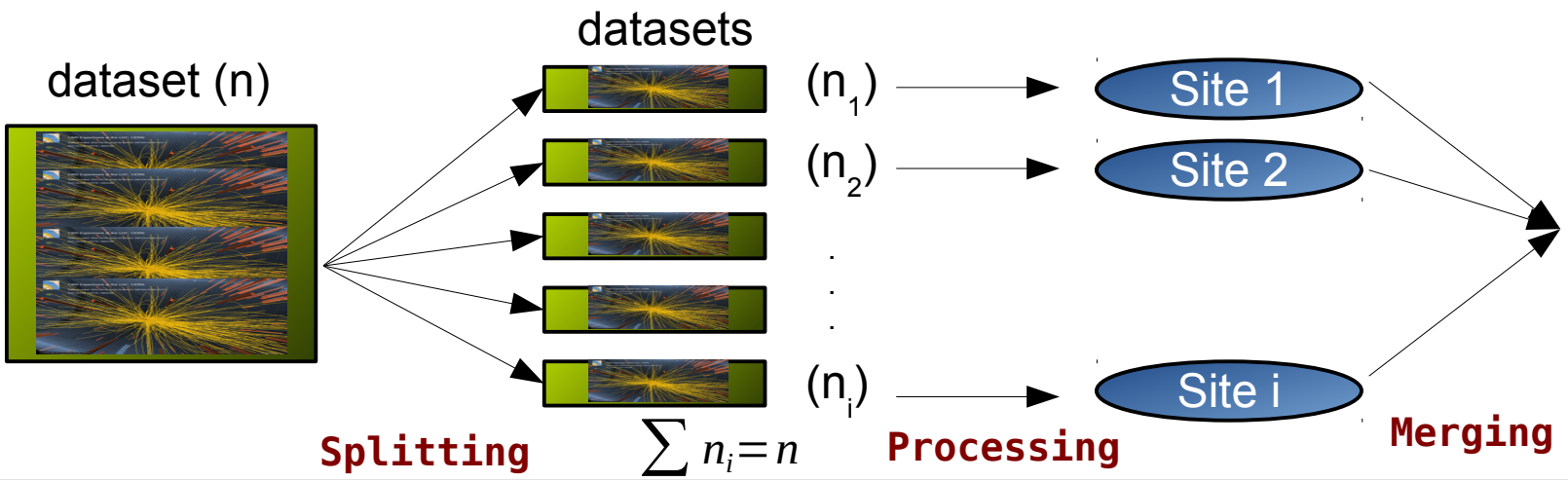


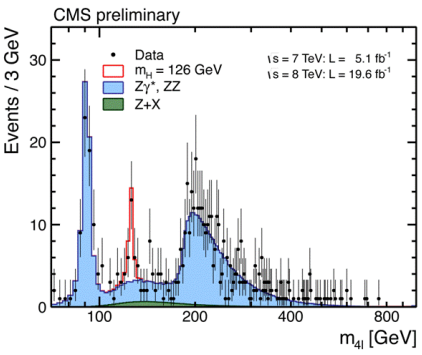


- Processus statistiques
- Evénements indépendants entre eux
- Outil du physicien = histogramme

Plus en détail

- Canal de physique → dataset
  - Liste d'informations pour chaque événement (collision)
  - Evénements indépendants
- Traitement du dataset
  - Lecture séquentielle de tous les fichiers du dataset
- Parallélisation triviale
  - Découpage en sous-tâches



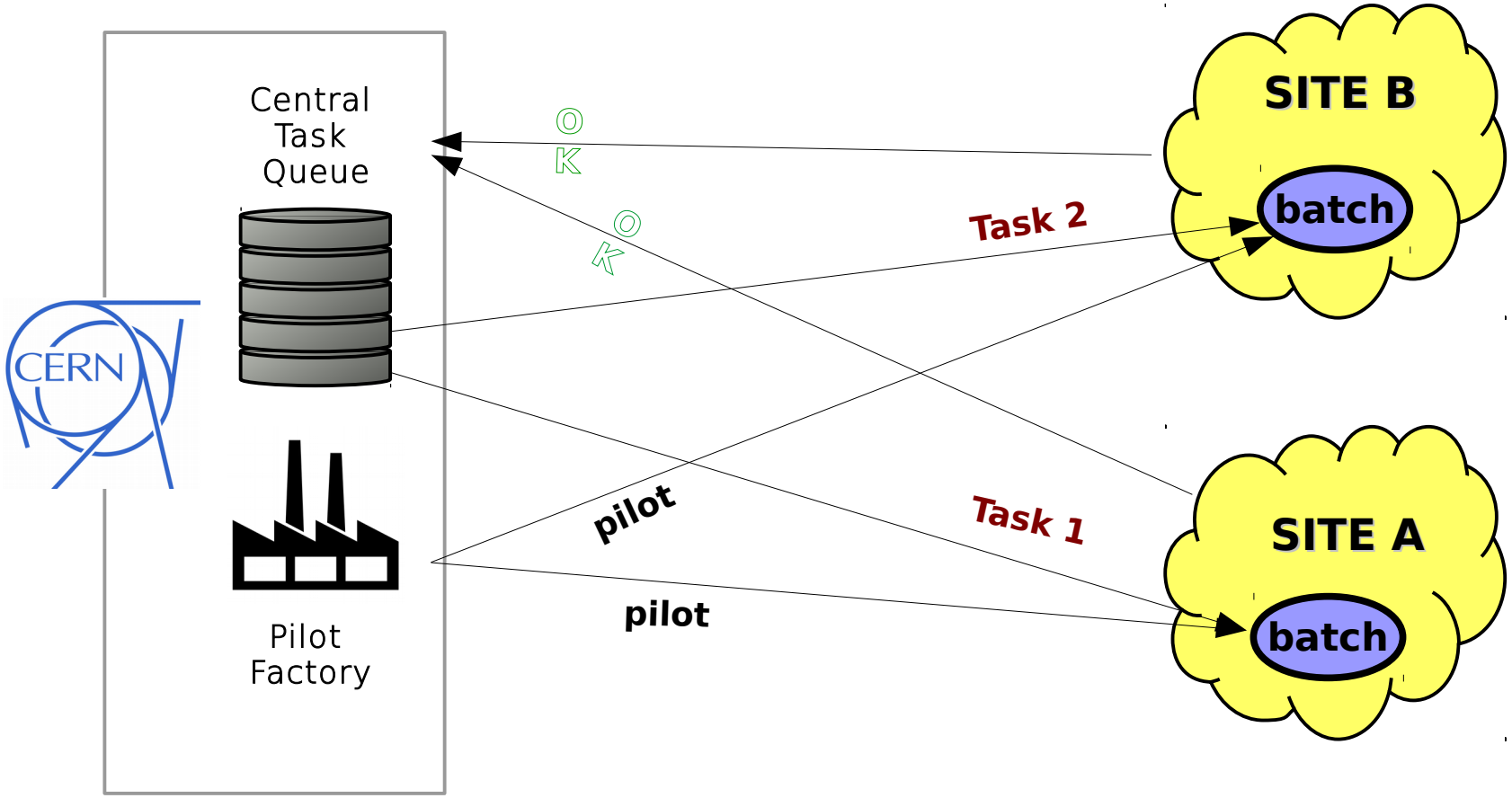


Conséquences

- Sous-traitements indépendants
- « Embarrassingly Parallel Problem »
- Pas de besoin (crucial) de supercomputers
  - car pas de communication entre workers

Challenge :
*** découper le dataset intelligemment**
*** dispatcher sur différents sites**

Job pilotes



~~PUSH~~ → PULL !

Job I/O

- Input
  - Software
    - CMVFS → mécanisme de déploiement de software sur les sites
    - Cache ← squid
  - Conditions database
    - Distributée ← squid
  - Données de physique
    - Stockage local (parfois distant < 10 %)
- Output
  - Résultats
  - Stockage local + répliques distantes

→ Environnement complexe
→ La grille s'appuie sur beaucoup de services
→ Redondance des services nécessaires
→ L'organisation est un pilier de la grille

Modèle de stockage

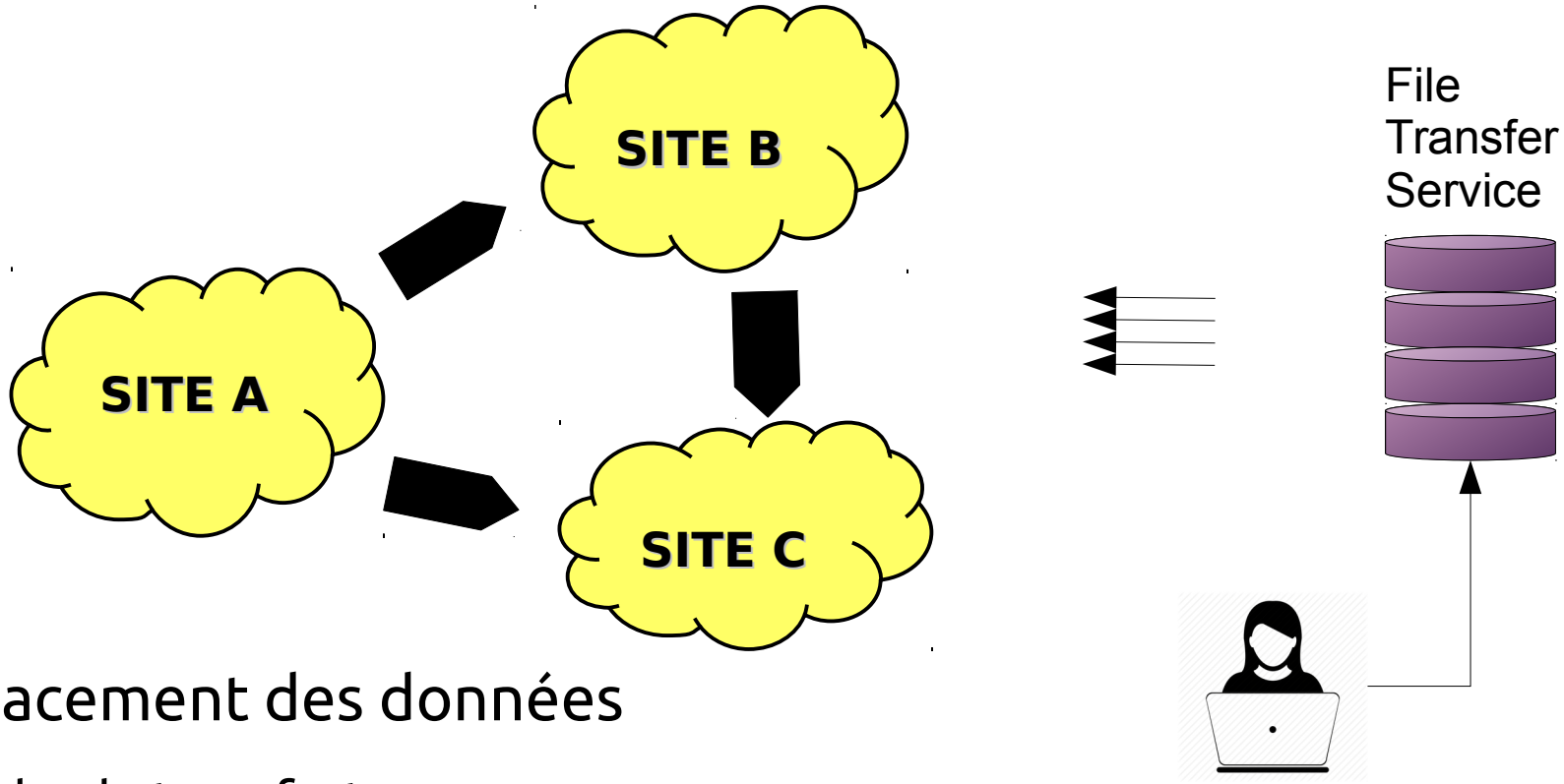
Catalogues de données

- Mapping fichiers logiques (LFN) ↔ physiques (PFN)
  - Ex :
    - LFN : /<namespace>/data/2011/…/fichier
    - PFN : root://host:port//<...>/<guid>
- Implémentations
  - LCG File Catalog (LFC)
  - Dirac
  - Alien...
- Fonctionnalités importantes
  - Gestion des répliques de données
  - Vue globale et detaillée du stockage grille

Utilisation bandes magnétiques

- En gros 50 % du stockage total
- Objectif premier : ‘archivage’ des données brutes
  - Faible risque de pertes
  - Acces 'organisé' et peu fréquent
- Et aussi
  - Simulation (certains cas)
    - Ecriture, (lecture), destruction
- Idéal
  - Gros fichiers (O(GB))
  - Communication site ↔ expérience
    - Ex : Campagne de reconstruction

Transferts de données



- Déplacement des données
  - Jobs de transferts
  - Queues
- → gestion par service ± centralisé
  - FTS

Technologies de stockage

- Facteur crucial
  - Performant en I/O
- Supportent nombreux accès simultanés
- Technos
  - dDcache , DPM, Xrootd, EOS
  - Et bandes magnétiques
    - Ex : TSM, HPSS, Castor

Et le stockage objet ?

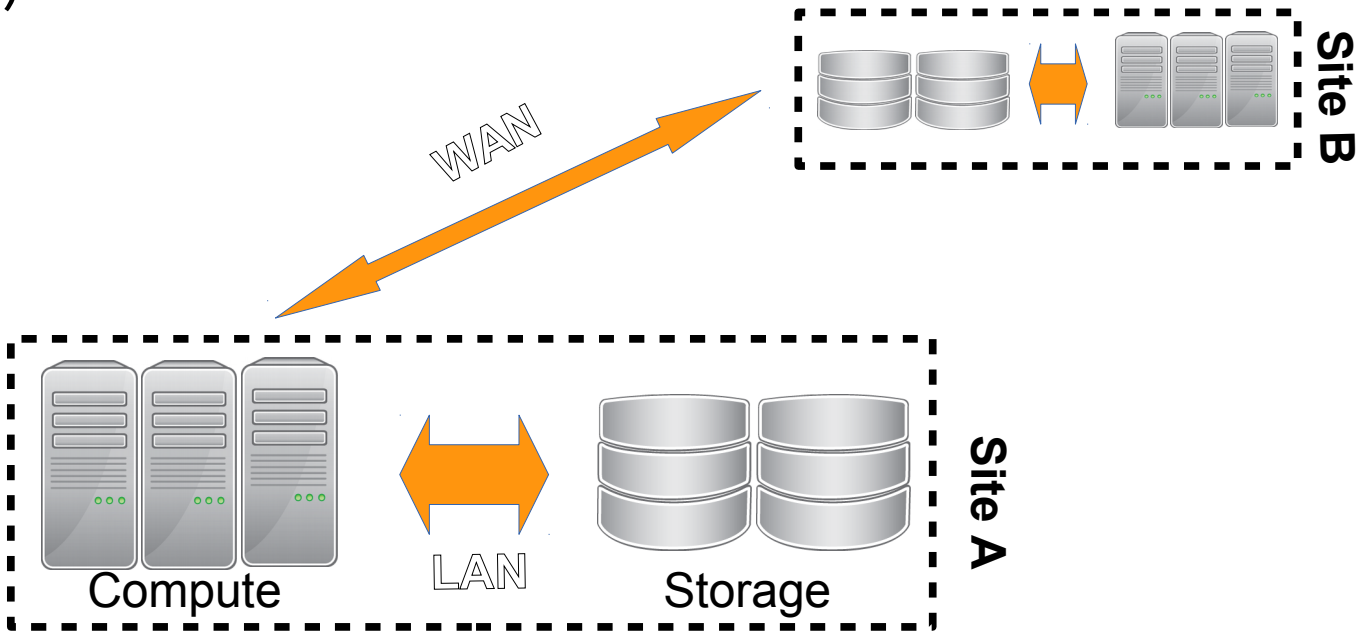
*** alternatives intéressantes**
*** compatibilité (API standards)**
*** performances ?**

Evolutions

- récentes et à venir -

Rétention des données

- LHC fonctionne très bien
  - Beaucoup de données à traiter (et stocker)
  - Moyens financiers limités
- Politique de rétention des données de + en + « agressive »
  - Réplication des datasets « populaires »
  - Effacement des datasets « impopulaires » ou obsolètes
  - Accès WAN (xrootd)

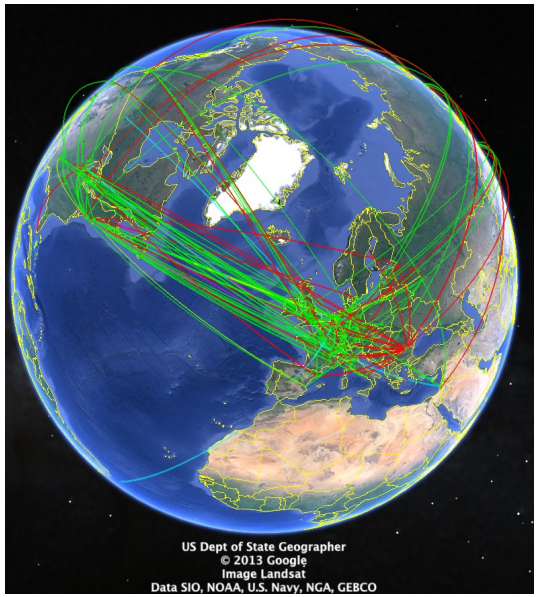


Optimisations diverses

- Jobs Multicore
  - Diminution de l'empreinte mémoire
- Plus forte utilisation du réseau
  - Permet de diminuer le besoin de stockage
- GPU
  - En cours
  - Adaptation du software nécessaire
  - Effort important !

Cloud

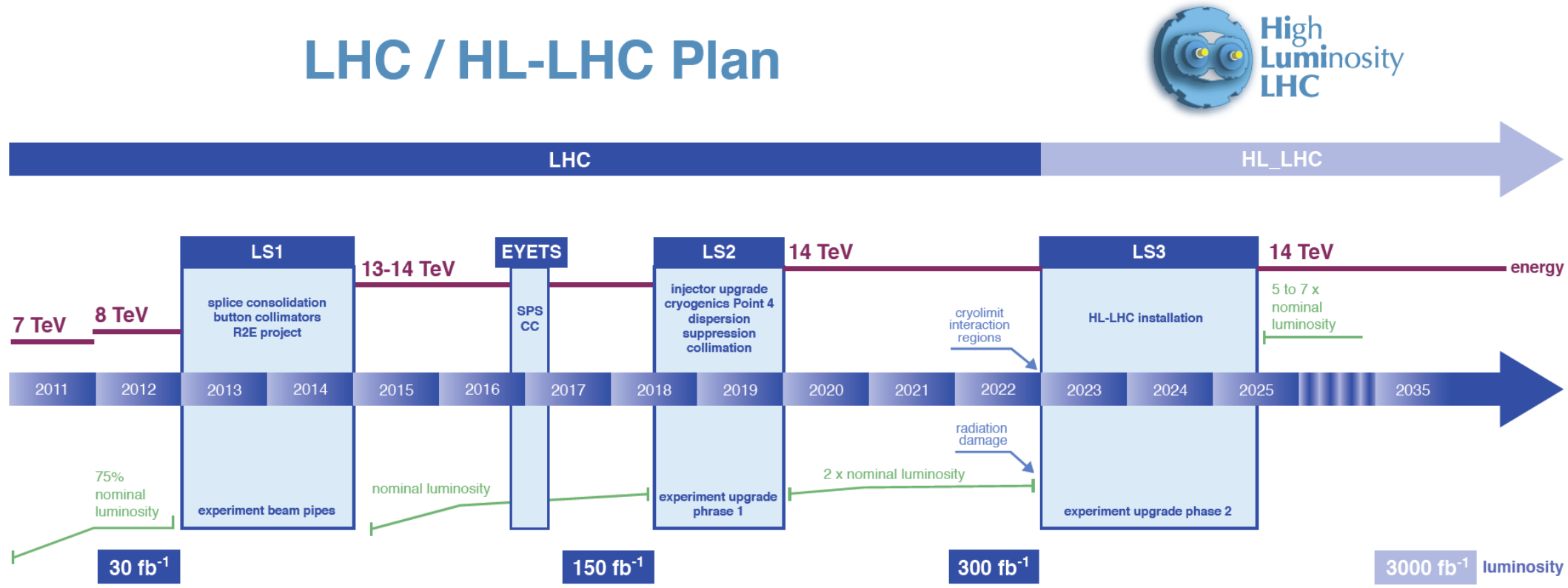
- Provision à la demande ↔ élasticité
  - Gestion des pics de charge
  - Gestion VM par expérience
    - Système, jobs
- Intérêt : sur grosses infrastructures
  - « public cloud »
    - (mon avis)
- Performance accès données
  - Simu ☺, analyse 😐



HPC

- Centres HPC (« supercomputer »)
  - Grosses capacités CPU
  - Complément aux centres traditionnels
  - Bénéfice des cycles perdus autrement
- Jobs de simulation
  - CPU >> I/O
  - 60-70 % des besoins CPU du LHC
- Obstacle principal
  - Connectivité extérieure
    - CVMFS, CondDB ...
  - Mise en place ‘gateway’
    - Contact local nécessaire
- Une réalité aujourd’hui

LHC roadmap


LHC / HL-LHC Plan
High Luminosity LHC
LHC
HL_LHC
LS1
splice consolidation
button collimators
R2E project
EYETS
SPS
CC
LS2
injector upgrade
cryogenics Point 4
dispersion
suppression
collimation
LS3
HL-LHC installation
7 TeV
8 TeV
13-14 TeV
14 TeV
14 TeV
energy
cryolimit
interaction
regions
5 to 7 x
nominal
luminosity
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2035
radiation
damage
75%
nominal
luminosity
nominal luminosity
2 x nominal luminosity
experiment beam pipes
experiment upgrade
phrase 1
experiment upgrade phase 2
30 fb-1
150 fb-1
300 fb-1
3000 fb-1
luminosity

Bilan

- WLCG : une infrastructure de production qui fonctionne très bien
  - Éprouvée
  - Sécurisée
- Atouts
  - Pragmatisme, agilité
  - Collaboration qui fonctionne (>100 sites)

- Avenir
  - Réseau de + en + sollicité
  - Bénéfice de nouvelles architectures de calcul
  - Bénéfice d'une interopérabilité avec environnements + standards
    - (mon point de vue)