PENNACCHIO Elisabetta, PUGNÈRE Denis
CNRS / IN2P3 / IPNL

Acquisition de données à haut débit défis sur le stockage online des données : Exemple pour ProtoDUNE Dual Phase

11èmes journées informatique de l'IN2P3-CEA/IRFU,
Port Bail, 1-4/09/2018

Plan de la présentation

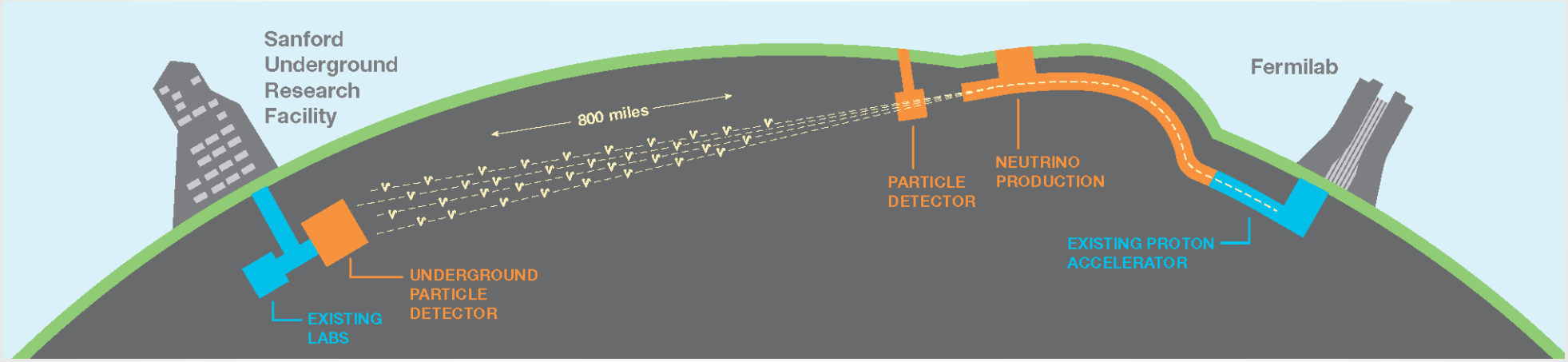
- Les DAQ
- L’expérience DUNE
- La DAQ de ProtoDUNE Dual-Phase
- Les besoins de stockage de l’expérience
- Design du backend de la DAQ
- Choix du système de stockage
- Data challenges
- Conclusion

Les DAQ

- DAQ : partie vitale d'une expérience pour récupérer les données des (sous-)détecteurs
- Est souvent placée dans des environnements contraints (espace, consommation électrique, refroidissement)
- Le traitement « Online » doit être efficace pour sélectionner, traiter, stocker les données, puis les envoyer vers des datacenters « offline » :
  - Réduction du flux (sélection, compression, zero-suppression)
  - Utilisation du matériel à sa pleine capacité, utilisation de matériels spécialisés (FPGA, GPU…)

DUNE http://www.dunescience.org/

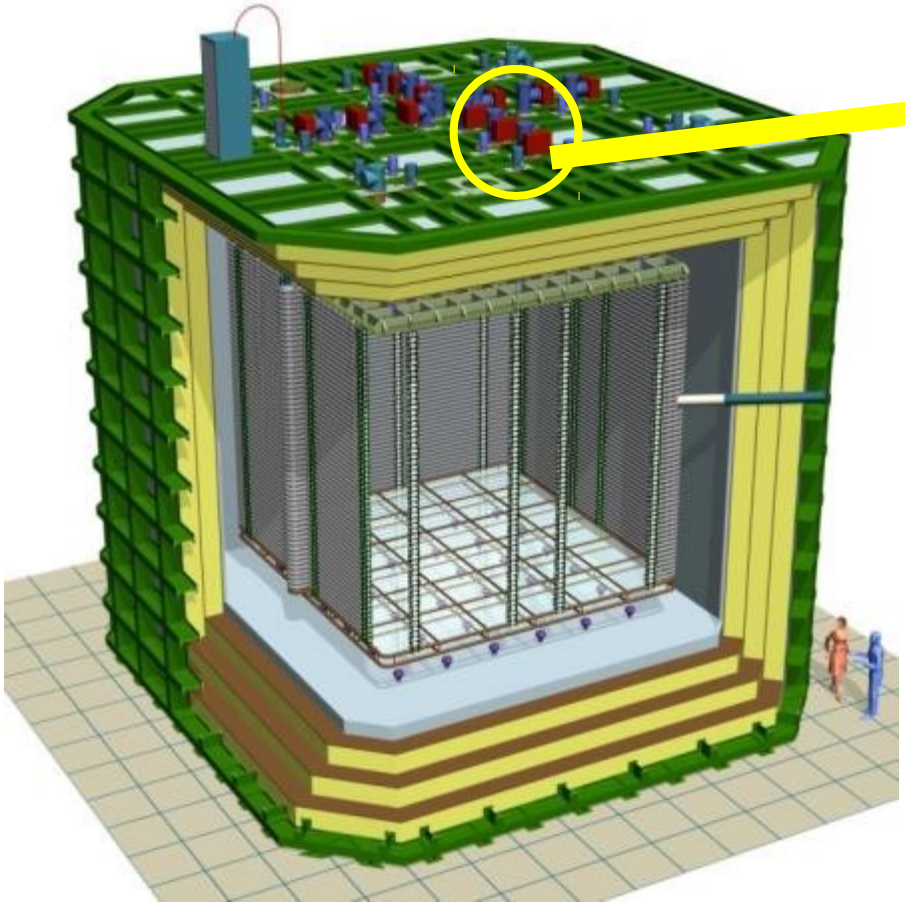
- DUNE : Deep Underground Neutrino Experiment (~2025)
- Expérience d'oscillations de neutrinos de nouvelle génération (mesure de la hiérarchie de masses, étude de la violation de CP)



- Faisceau + détecteur proche @ Fermilab
- Détecteur lointain (4 modules TPC argon liquide 10kT) @ Sanford (souterrain)
- 2 technologies : single-phase / dual-phase (dual-phase => amplification du signal)
- 2 prototypes actuels ProtoDUNE en surface au CERN (fin 2018) :
  NP02/WA105 (dual phase) et NP04 (single phase)

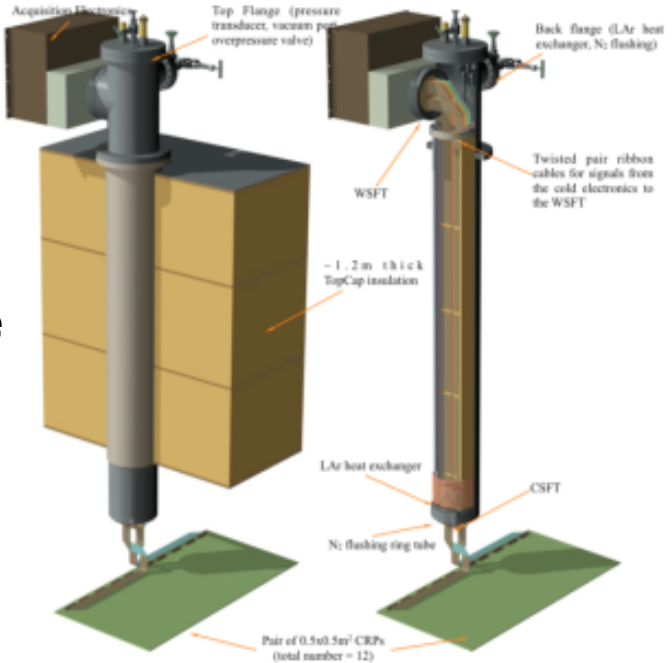
Dual-phase protoDUNE DAQ :

6x6x6m3 active volume (300T liquid argon = 1/20 10kton LBNO)



µTCA crate

10Gb/s



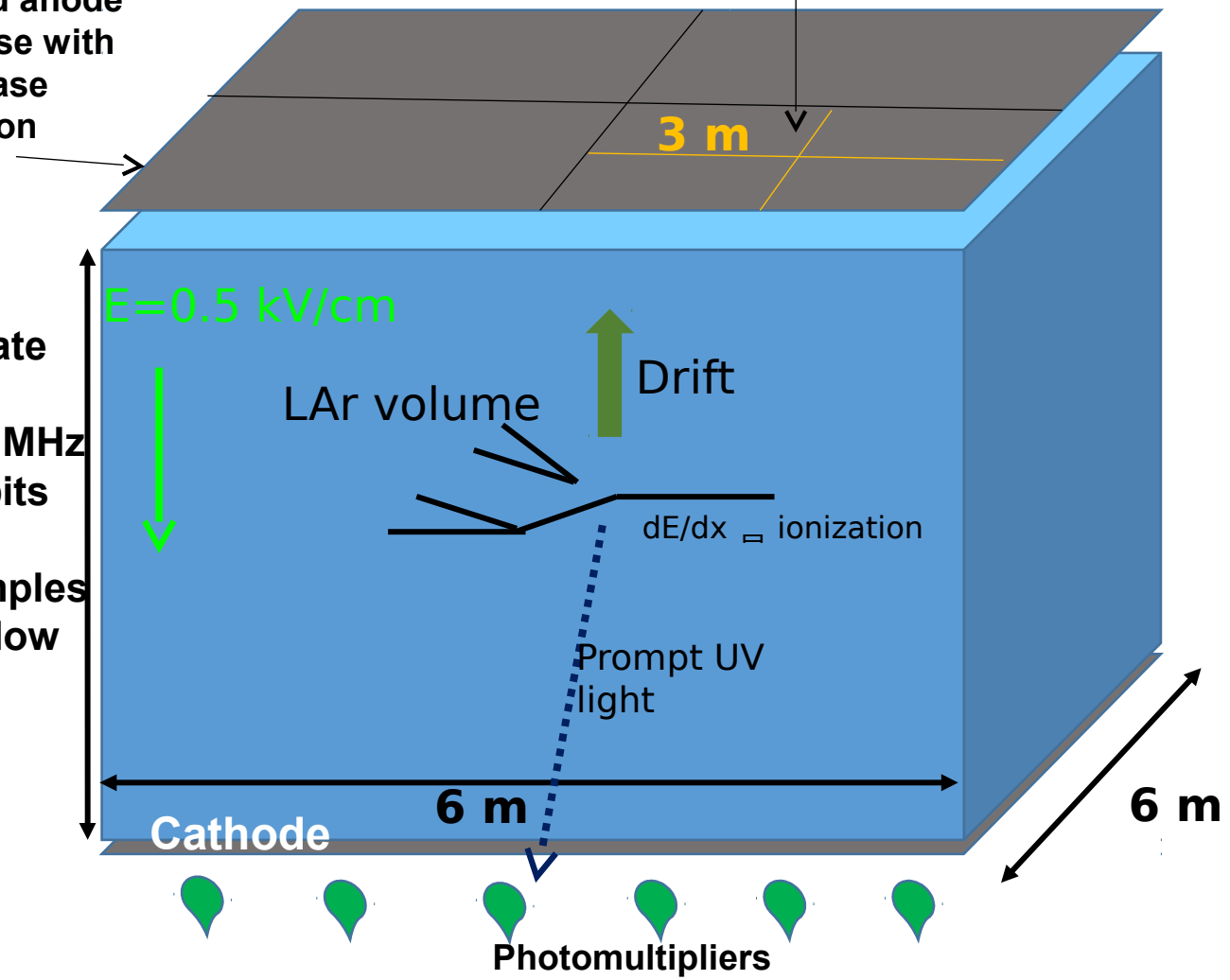
Charge: 12 uTCA crates, 10 AMC cards/crate, 64 ch/card => 7680 channels

Double phase liquid argon TPC
6x6x6 m3 active volume

→ **Event size: drift window of**
7680 channels x 10000 samples = 146.8 MB

X and Y charge collection strips
3.125 mm pitch, 3 m long
→ **7680 readout channels**

Segmented anode
In gas phase with
double-phase
amplification



Drift coordinate
6 m = 4 ms
sampling 2.5 MHz
(400 ns), 12 bits

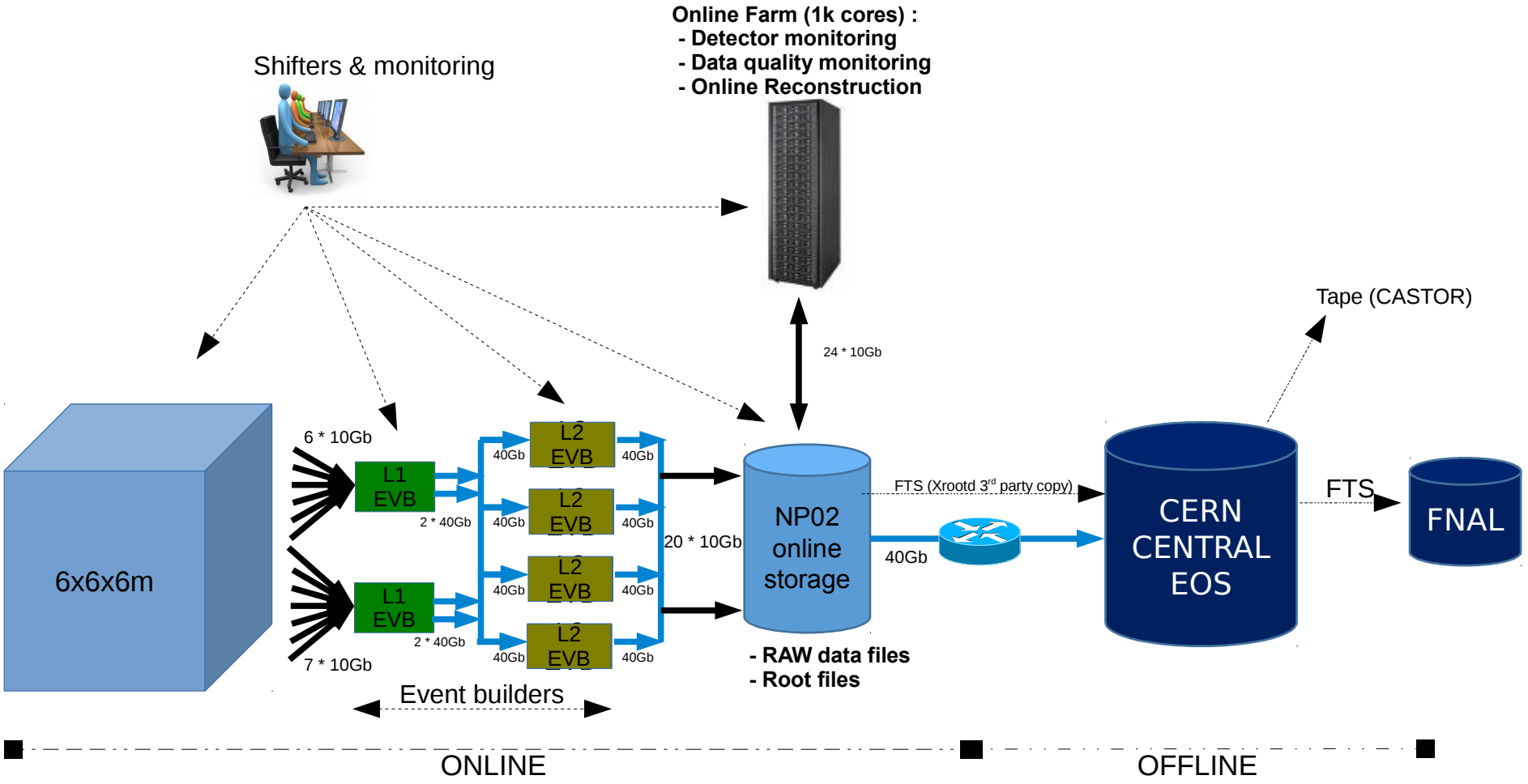
→ **10000 samples**
per drift window

Besoins de stockage de l’expérience

- Débits débits d’acquisition des deux expériences :
  - ProtoDUNE single-phase : event size 230 MB, trigger rate 25Hz
    => data rate 46Gb/s => 11.25 Gb/s assuming a compression factor = 4
  - ProtoDUNE dual-phase : 146.8MB / event, trigger rate 100Hz
    7680 channels, 10 000 samples, 12 bits (2.5Mhz : drift window 4ms) :
    => data rate 120Gb/s => ~10Gb/s assuming a compression factor = 10

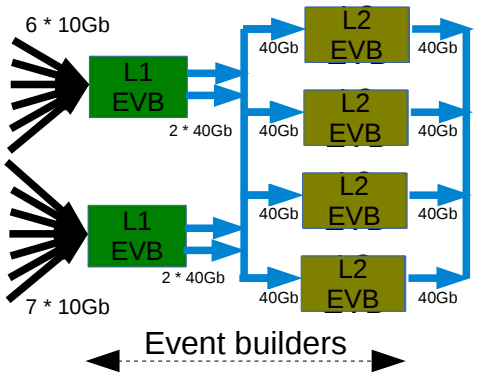
- La solution de stockage pour ProtoDUNE dual-phase :
  - Taille 1 Po (buffer local de 1 à 3 jours de prise de données)
  - Est conçue pour pouvoir absorber 130Gb/s (60Gb/s + 70Gb/s) en écriture
  - Compression sans perte Huffman appliquée, réduction de la taille des événements par un facteur 10
  - Permettre la lecture des données pendant la prise de données afin de déterminer la qualité des données
  - La recopie des données vers le T0 du CERN (lien 40Gb/s)

Design du back-end de la DAQ de ProtoDUNE Dual-Phase



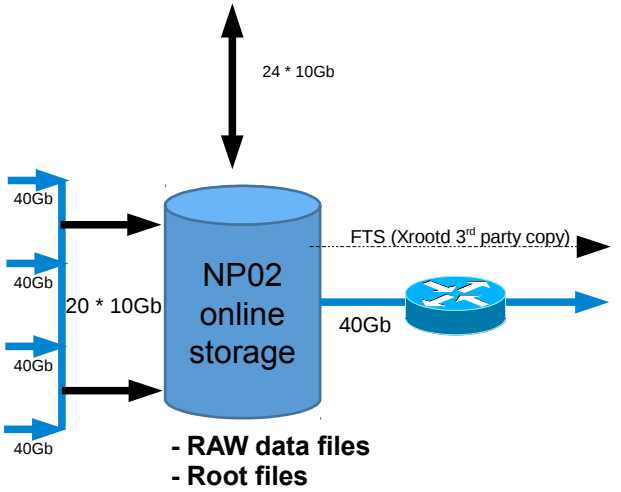
ProtoDUNE Dual-Phase Event-builders

- 2 * L1 Event Builders
  - Dell R740, 2 * CPU Gold 5122, 384 Go RAM
  - 2 * Intel X710 4*10Gb/s
  - 2 Mellanox Connect-X3 Pro 2 * 40Gb/s
  - Collectent les données : chaque EVB = 1/2 du détecteur
- 4 * L2 Event Builders
  - Dell R740, 2 * CPU Gold 5122, 192 Go RAM)
  - 2 Mellanox Connect-X3 Pro 2 * 40Gb/s
  - Assemblage de l'evennement (chaque moitié des EVB-L1 = 1/2 du détecteur), écriture des fichiers sur le stockage online
- Communication RDMA/RoCE entre les Event-Builders



ProtoDUNE Dual-Phase « online » storage system

- 20 * Data storage servers (= 20 EOS FST)
  - Dell R510, 2 * CPU E5620, 32 Go RAM) : 12 * 3TB SAS HDD
  - Dell MD1200 : 12 * 3TB SAS HDD
  - 1 * 10Gb/s
  - 4 * RAID 6 on 6 HDD
- 2 * EOS Metadata servers (MGM)
  - Dell R610, 2 * CPU E5540, 48 Go RAM)



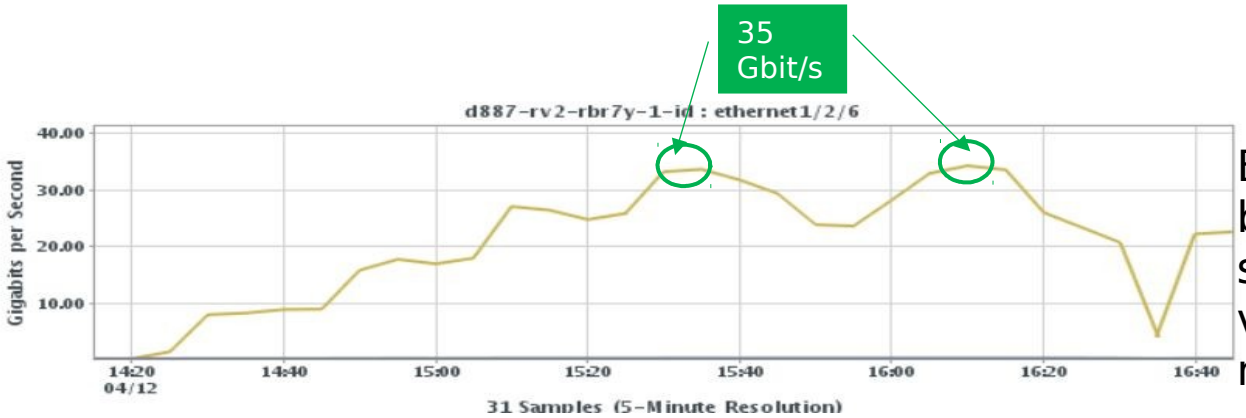
Choix du système de stockage

- EOS choisi (après les tests réalisés en 2016) :
  - Stockage low-latency,
  - Contrôle d’accès : POSIX, Kerberos, GSI,
  - Protocole d’accès XrootD, POSIX,
  - Support du 3rd party-copy,
  - Support intégré des checksums,
  - Redondance :
    - Serveurs de Metadonnées
    - Données (replicas, striping RAIN)
  - Gestion du cycle de vie des data servers (draining, entrée/sortie de production)

Data challenges

- Internes à l’expérience :
  - Taux d’écriture sur le stockage « online »
  - Comportement du stockage en fonction des profils d'I/O
- Stockage online => stockage offline (EOS T0 CERN)
  - Plusieurs data-challenges de transfert online => offline : chaque expérience indépendamment, ou data-challenges conjoints
  - Plusieurs objectifs :
    - Qualification du lien réseau 40Gb/s (EHN1, bâtiment 887) => (T0 CERN, bâtiment 513)
    - Étudier le comportement du système de stockage online lors du transfert
    - Étudier le comportement du système de stockage offline (eospublic.cern.ch : instance partagée par plusieurs expériences) lors des transferts
    - Tester plusieurs systèmes de transfert de fichiers : FTS, scripts maison…
- Mais aussi transferts CERN => FNAL

Exemples de data challenges

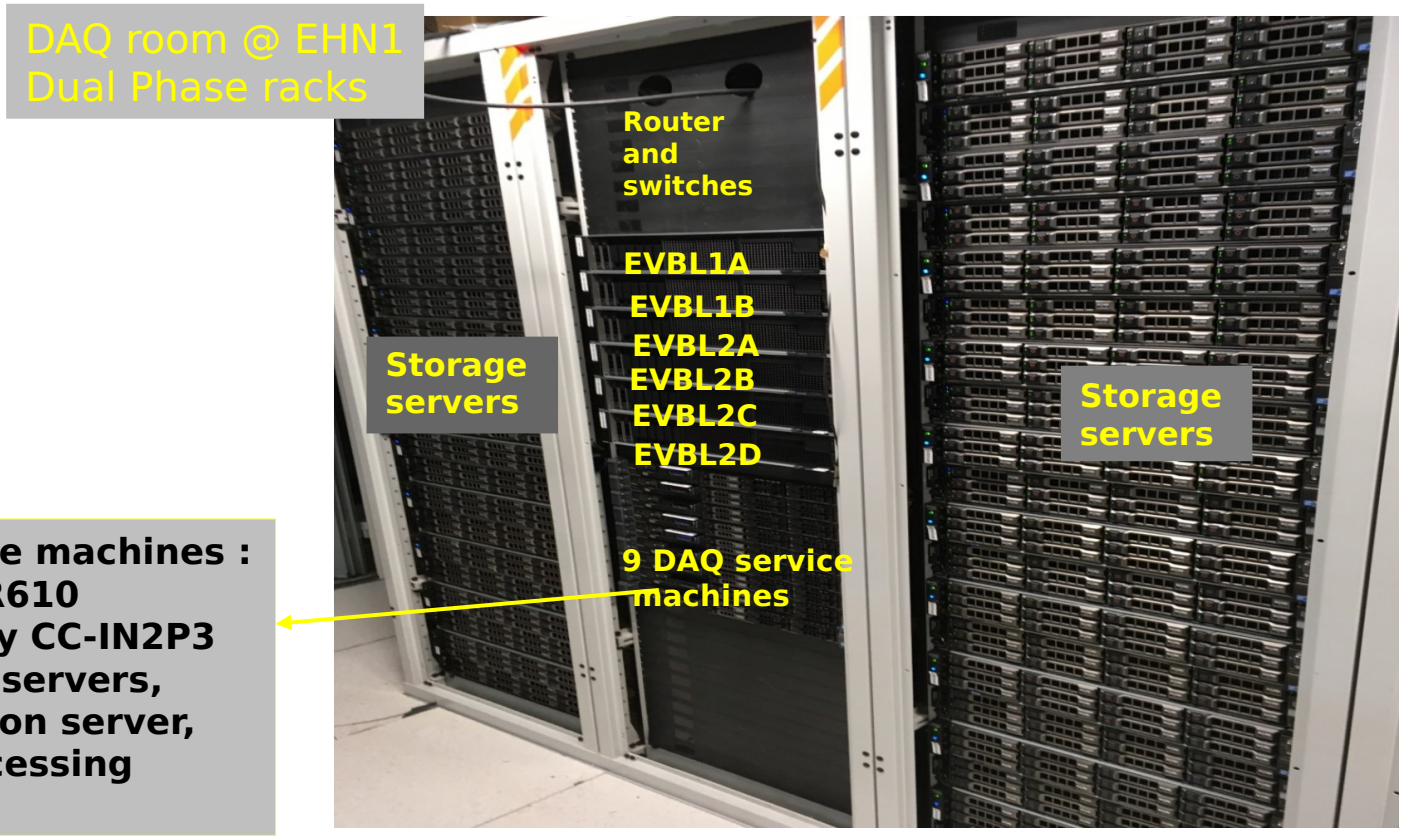


Extrait d'un data challenge de transfert buffer online <=> buffer offline : scripts « maison », variation du nombre de data servers mis en œuvre (1, 2, 3, 4, 5...)

Extrait d'un data challenge transfert FTS : buffer online <=> buffer offline, 130 jobs en //, 1 stream / job :

Conclusion

Pour l’expérience ProtoDUNE Dual-Phase, nous avons conçu un backend DAQ et implémenté un système de stockage distribué, testé par différents scénarios de data challenges, répondant aux besoins de l’expérience

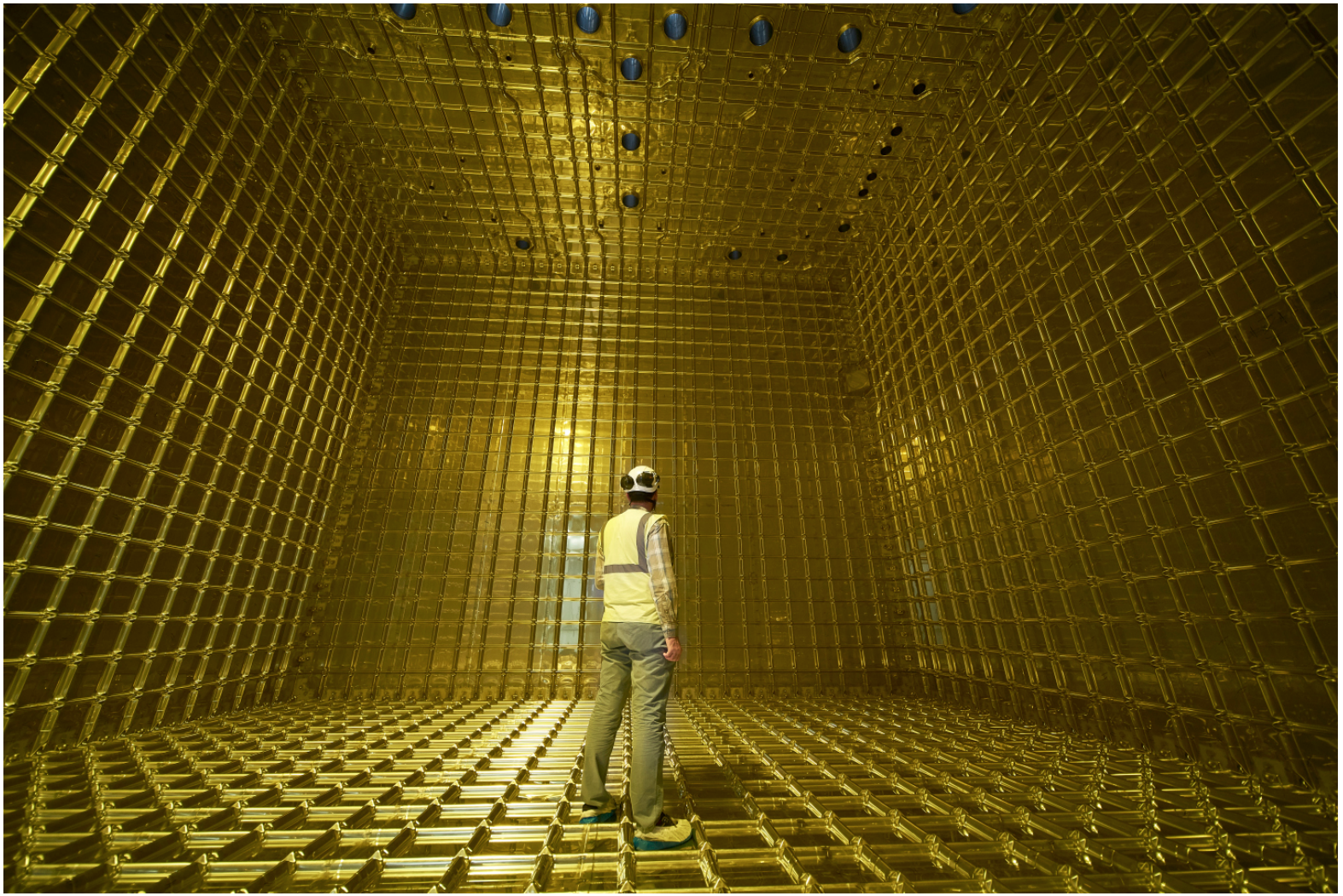


Remerciements

- CC-IN2P3 : serveurs de stockage, CPU farm, services nodes
- La division IT du CERN : en particulier Giovanna Lehmann Miotto, Marc Collignon, Xavier Espinal
- KEK (Japon)
- La Neutrino Platform du CERN

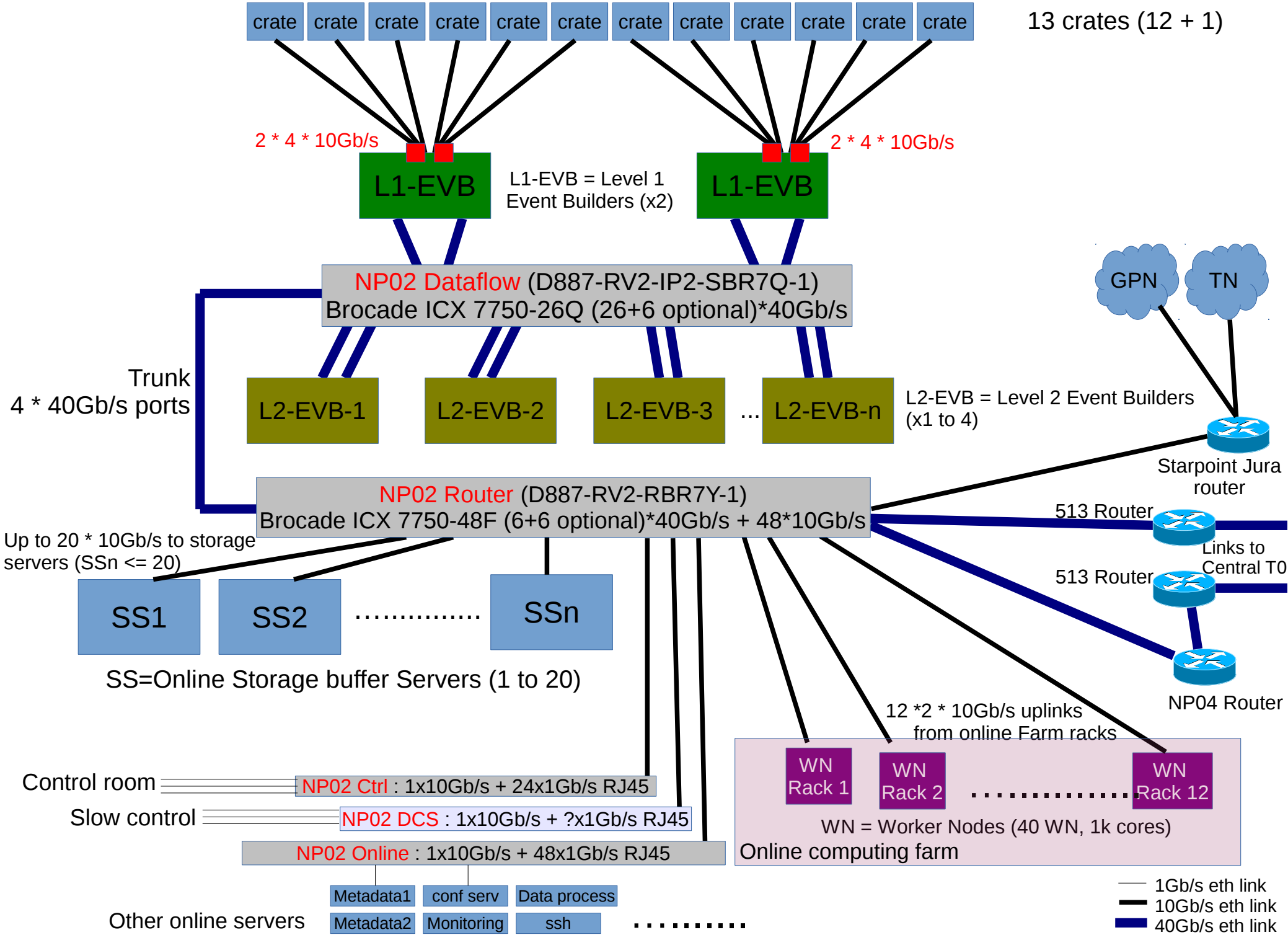
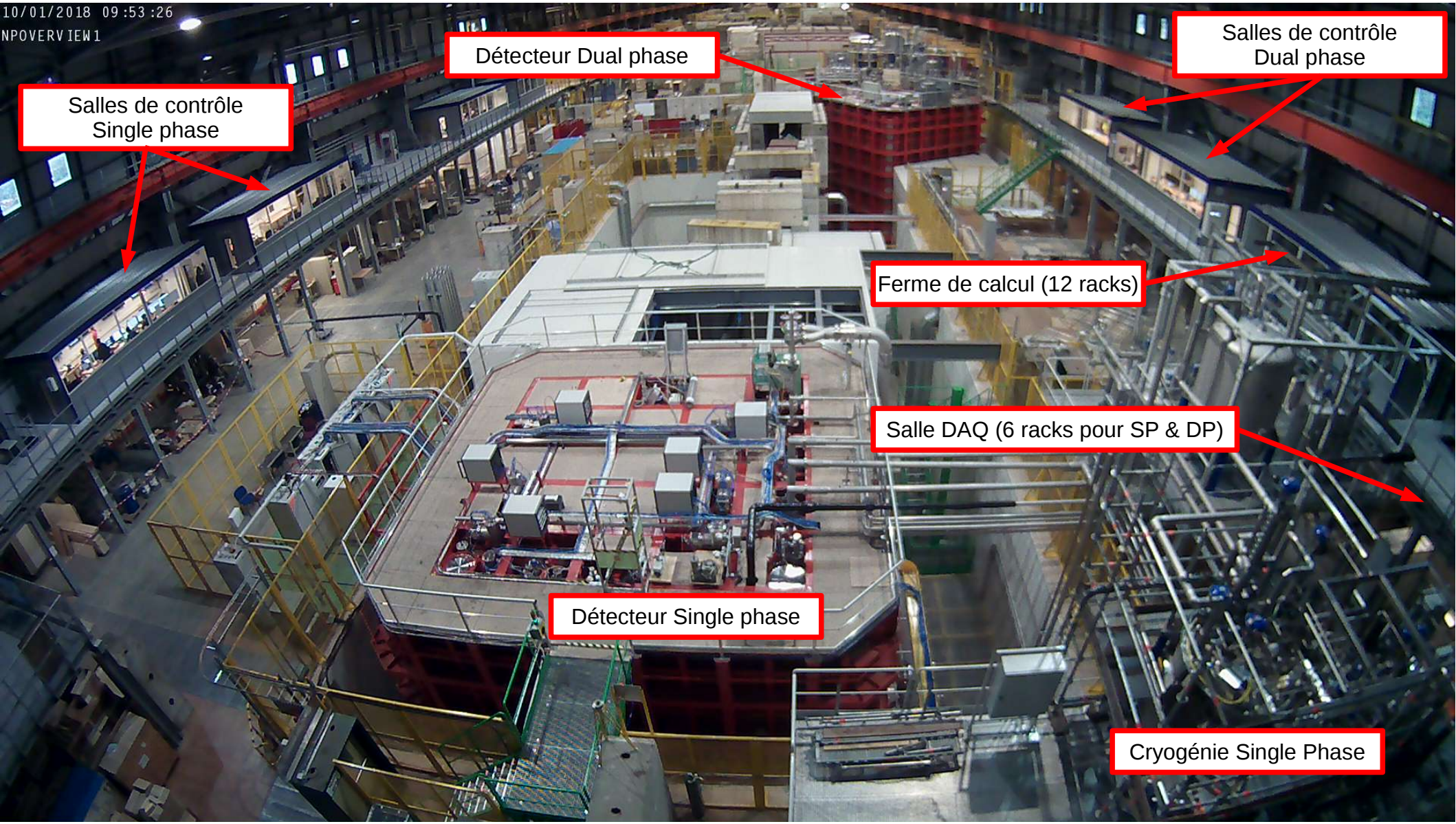
Détails du diagramme physique détaillé du backend DAQ et du bâtiment de l’expérience dans les backup slides suivantes...

Backup slides



Inside ProtoDUNE : http://cds.cern.ch/images/CERN-PHOTO-201710-248-2

ProtoDUNE-DP network physical diagram



EHN1 (CERN North area) au 01/10/2018