Centre de Calcul

de l'Institut National de Physique Nucléaire
et de Physique des Particules

ANF Ceph 2022
Ceph au CC-IN2P3

Loïc Tortay, tortay@cc.in2p3.fr

CC-IN2P3

- IN2P3
- CC
  - stockage et traitement des données pour les expériences & collaborations auxquelles participent les physiciens de l'IN2P3
  - 3500 utilisateurs, 40 à 50% hors IN2P3 (& CEA)
  - activité d'ouverture
  - 70 Pio (utiles) de disques (& Flash), 125 Pio sur bandes
  - ~1300 machines de calcul, ~500 serveurs de stockage
  - 80 personnes

Ceph @CC : brève chronologie

- 2017 : cluster pour OpenStack @CC (RBD)
- 2019 : cluster hébergé pour l'IFB (OpenStack)
- 2020 : cluster pour LSST (CephFS)
- 2021 : extension cluster LSST (capacité +60%)
- 2021 : cluster pour WoK (RBD+CephFS+S3)
- 2022 : cluster pour remplacement GPFS (CephFS)

Ceph @CC : clusters

- OpenStack : Nautilus (Ceph v14), 600 To, 16 serveurs de disques (disques + SSDs & SSDs uniquement), 258 OSDs, réplication
- *IFB : Pacific (v16), 500 To, 4 serveurs de disques (disques + NVMe), 64 OSDs, réplication*
- WoK : Pacific, 150 To, 3 serveurs de disques, 72 OSDs, réplication
- LSST : Octopus (v15), 7 Po, 32 serveurs de disques (disques + SSDs), 640 OSDs, *erasure coding* (8+2) pour les données & réplication pour les métadonnées
- SPS : Quincy?, 7 Po, 30 serveurs de disques (disques + SSDs), 480 OSDs, CephFS, *erasure coding*
- 3 MONs (physiques) sur tous les clusters

Ceph @CC : configurations matérielles

- Dell R730xd, R740xd, R740xd2, R440 : 67 serveurs
- HPE A4200 & DL360 (pas encore en prod) : 33 serveurs
- Dell R740xd2 (cluster LSST), x32 :
  - 2 Xeon Silver 4210, 128 Gio
  - 20 disques 12 To SAS-NL, 4 SSDs 1.6 To SAS
  - 2 x 10G Eth
- HPE A4200, x30, similaire Dell R740xd2, sauf :
  - 16 disques 14 To SAS-NL, 6 SSDs 1 To SAS
  - 2 x 10/25G Eth

Ceph @CC : utilisations

- RBD pour OpenStack
- CephFS :
  - très limité pour le cluster OpenStack (Spark, etc.), Manila peu utilisé en production
  - WoK en préproduction
  - utilisation à assez grande échelle pour LSST & SPS (bientôt 2 clusters de ~5 Pio utiles chacun)
- RGW très limité pour OpenStack & WoK

Ceph @CC : CephFS/LSST [1/3]

- Migration de 1.1 Pio de données de GPFS & Isilon vers CephFS
- Ajout de 12 serveurs de disques (+1.8 Pio utiles) :
  - 11 jours de redistribution des données (~1.2 Pio)
  - sans interruption des jobs utilisateurs
- Performances accès mono-*thread*
- Bugs mineurs (documentation, MGR)
- Montage statique au lieu de l'automonteur
- Maturité de CephFS (vs GPFS) face à des utilisateurs
- Comportement des *hardlinks* (Anaconda, utilisateurs avancés)

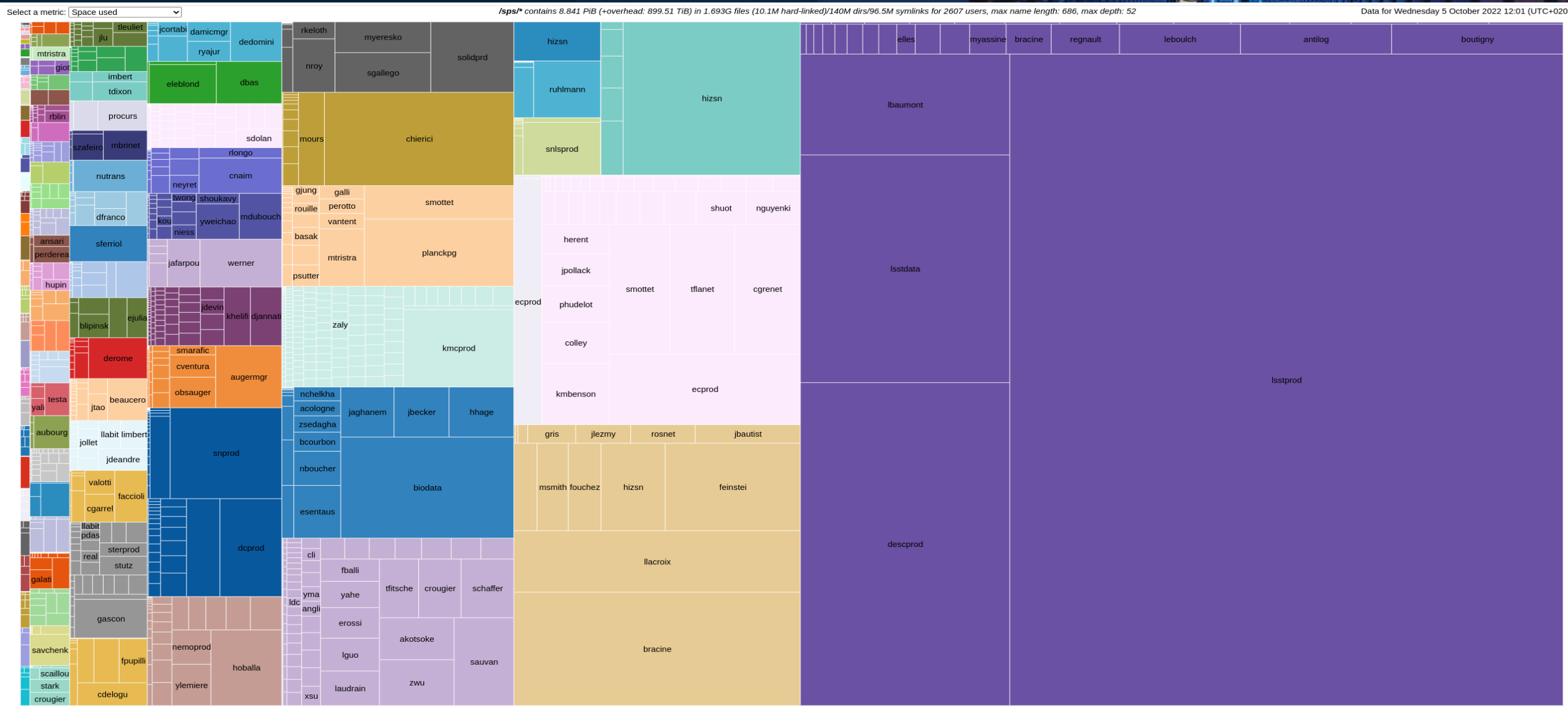
Ceph @CC : CephFS/LSST [2/3]

- Stabilité MDS (initialement 2 actifs + 1 *standby*) :
  - blocages détectés mais pas toujours leur source
    ⇒ redémarrage/basculement autoritaire MDS
  - pics de latence constatés par les utilisateurs
  - *scrub* (MDS pas OSD)
  - depuis début 2022, un seul MDS actif
  - `mds_cache_memory_limit` à 32 Gio
- Réduction du taux de vérification (`scrub_rate`)
- Grafana & monitoring Prometheus
- Liste de diffusion *ceph-users@ceph.io*

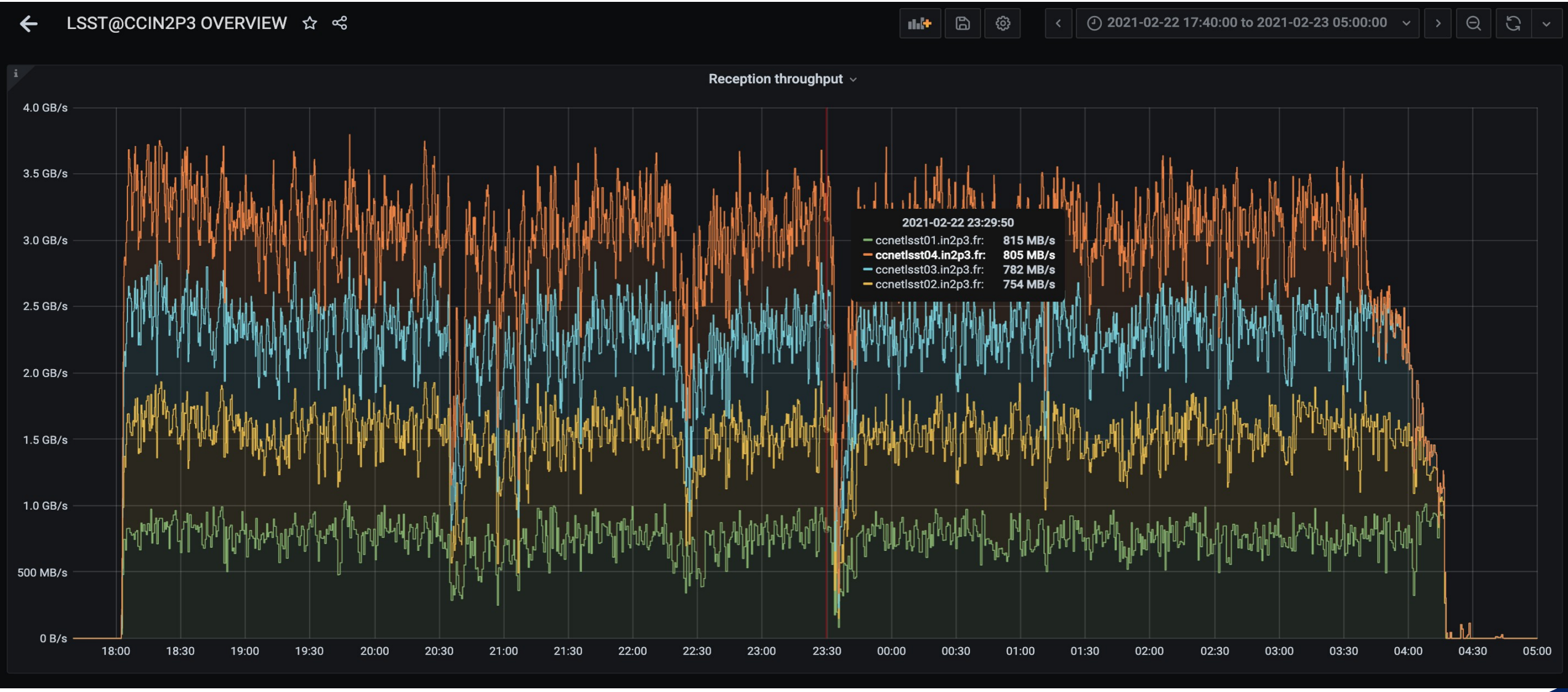
Ceph @CC : CephFS/LSST [3/3]

- Bonne résilience à un serveur de disques capricieux
  - `upmap-remapped.py` de Dan van der Ster (CERN) & `pgremapper` (Digital Ocean)
- Export Web de sous-répertoire utilisateur
- Collecte quotidienne métadonnées pour comptabilité & gestion d'espace
- Attributs étendus dynamiques (virtuels) CephFS :
  - `getfattr -n ceph.dir.rbytes $DIR`
  - ```
    for attr in rbytes rsubdirs rfiles rctime ; do
       getfattr -n ceph.dir.$attr $DIR
    done
    ```
    ```

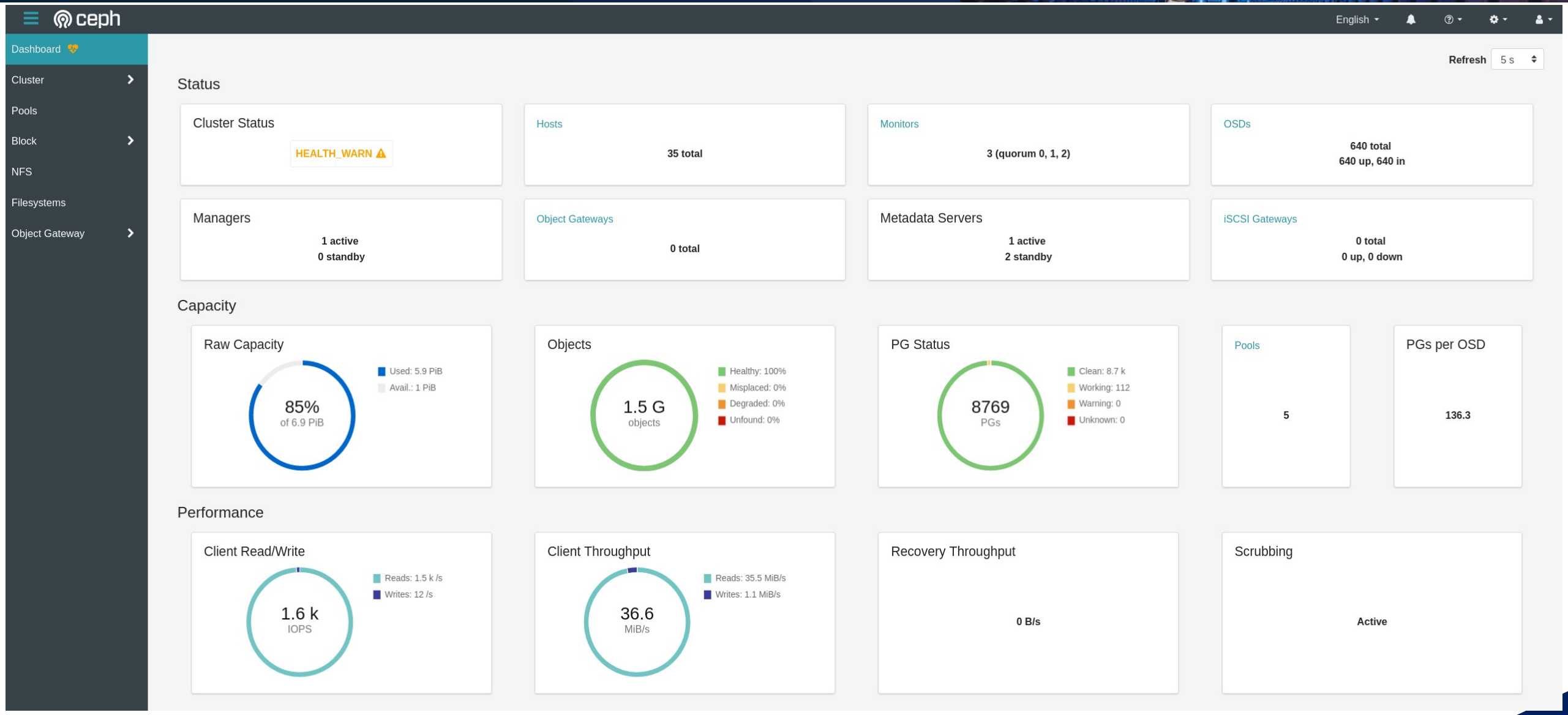
Ceph @CC : comptabilité/statistiques



Ceph @CC : transferts LSST CC ⇔ NERSC

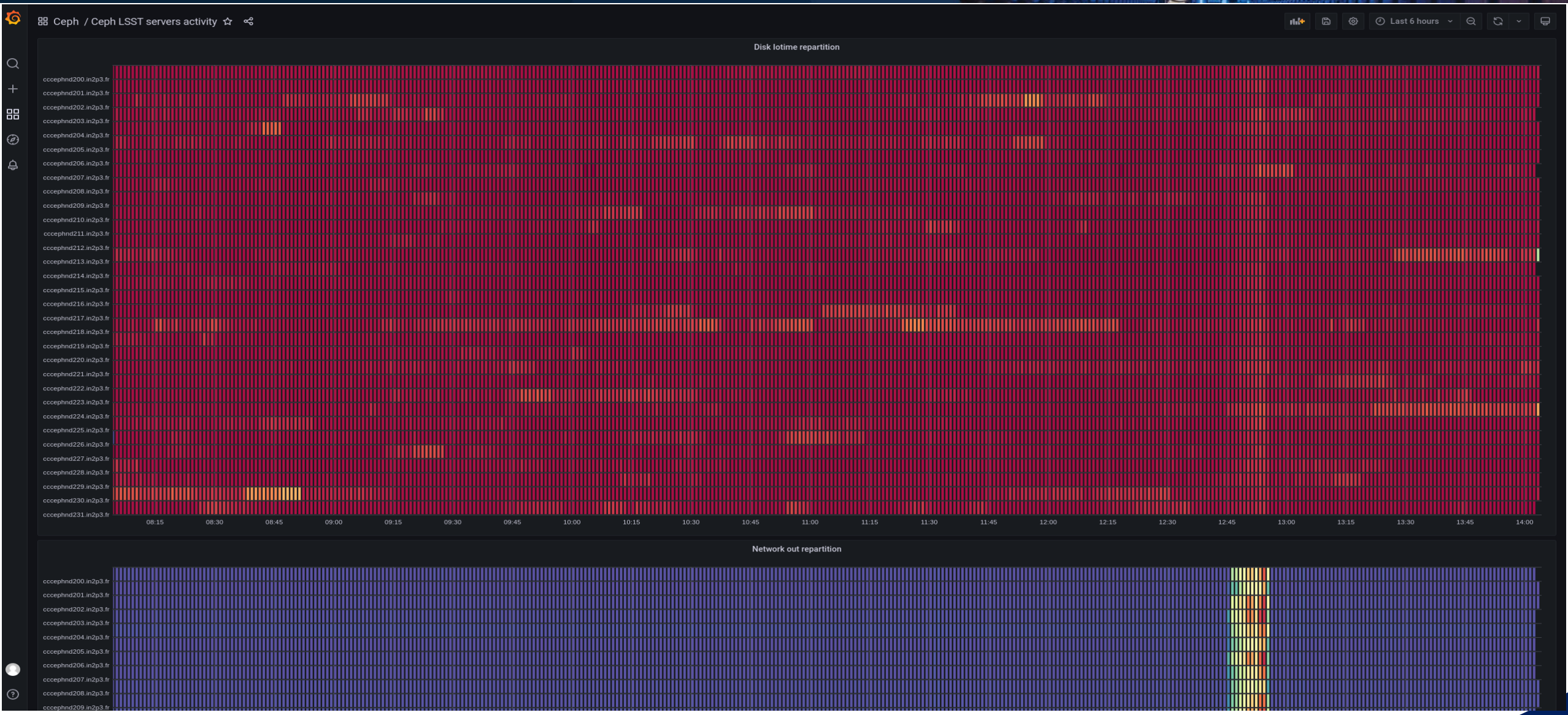


Ceph @CC : Dashboard Ceph



Ceph @CC : Grafana Ceph

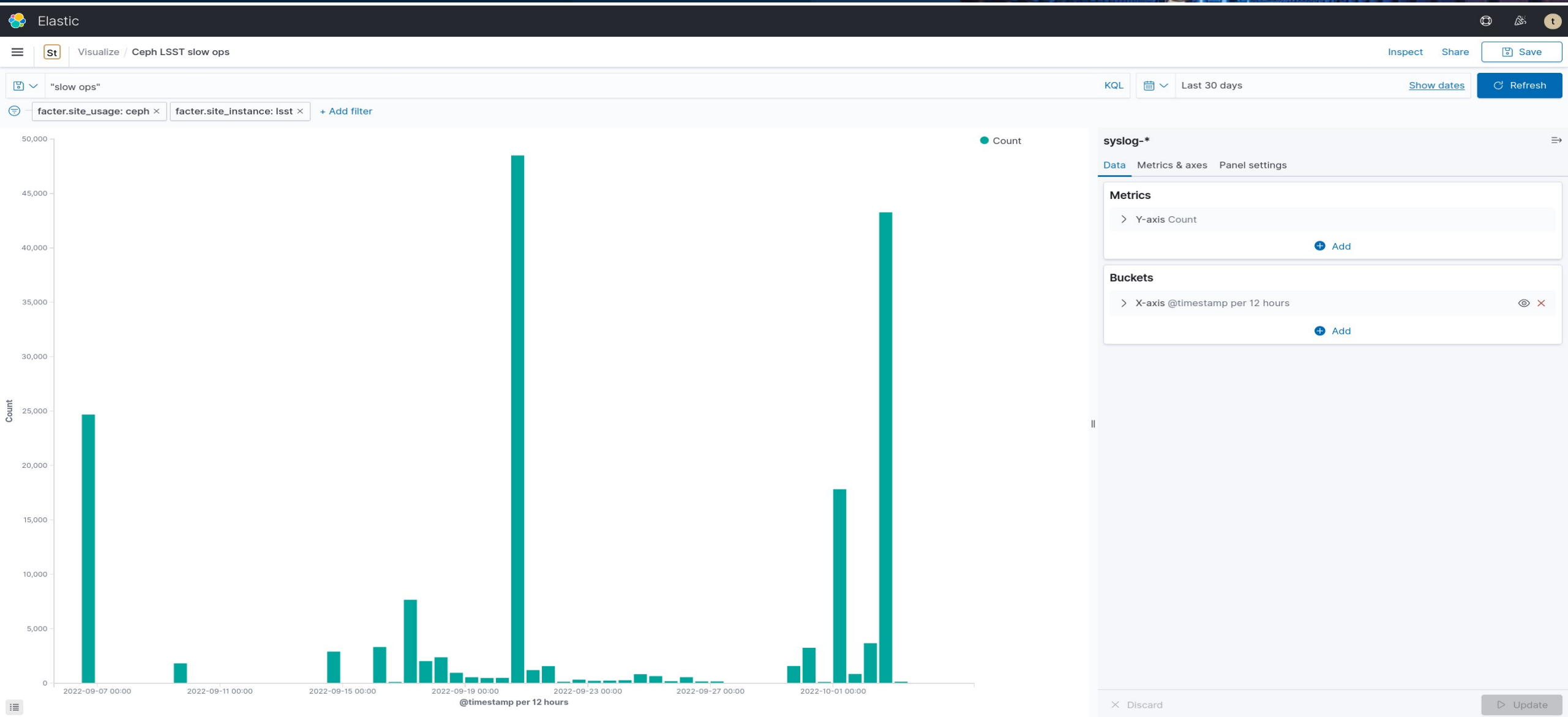
Ceph @CC : Grafana Ceph activity heatmap



Ceph @CC : Grafana CephFS

Ceph @CC : Grafana CephFS

Ceph @CC : Elastic Search CephFS slow ops



Ceph @CC : Grafana CephFS ?!?