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

Evolution of GPFS

and exploration of other storage technologies

FJPPL Meeting, CC-IN2P3, 15 February 2017, Loïc Tortay

GPFS @CC-IN2P3

- In use since early 2006
- Spectrum Scale 4.1.1 (upgrade to 4.2.2 in March)
- Main cluster with about 900 nodes:
  - all computing and login nodes, a few service nodes
  - 61 disk servers in production, 3 being commissioned
  - 2.2 PiB usable (+435 TiB soon & 2017 increase TBD)
  - 37 filesystems
  - 600M files for about 1700 users in 100 groups
- Small cluster for OpenStack backend:
  - Glance & Instances for some tenants
  - 10 servers & 10 clients (hypervisors)
  - 45 TiB (SNC w/ 3-way replication)

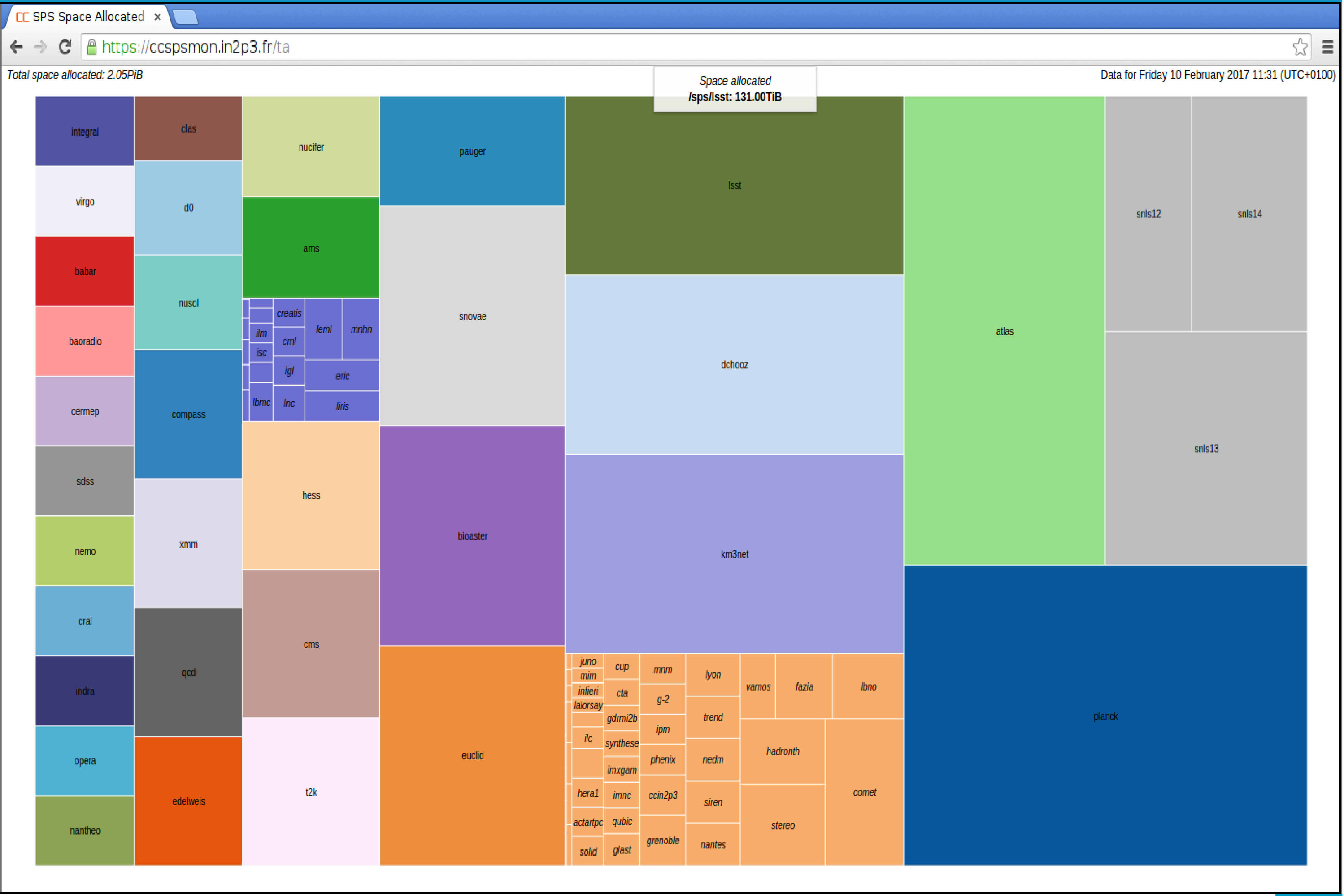
GPFS @CC-IN2P3: Semi-permanent Storage (aka SPS)

- Shared group space for active (data) files, w/ large capacity and **no** backup
- Original plan was to have generalized automated cleanups of inactive files:
  - in practice many groups only (& often reluctantly) clean when there is no space left
  - HSM integration (GHI or otherwise) initially deemed undesirable
- Plan for lightweight HSM integration, either:
  - simple ILM external migration (w/o transparent recall)
  - Spectrum Scale Transparent Cloud Tiering, maybe using Swift (or something w/ a Swift-like interface)
- Tool needed to allow users to migrate files (TBD)

GPFS @CC-IN2P3: Filesystems

- Historically 1 FS per collaboration/experiment:
  - 62 FS in 2012, 37 now, aim is 4 FS
- Administration simplification:
  - 1 fileset per group
  - finer space allocation granularity
  - tiering *(Fileheat)* w/ reasonable hardware requirements
  - mutualized *inactive* or specialized pools
  - more servers for more concurrent batch jobs
- Consolidation :
  - Initially manual migration w/ a rsync-like tool
  - AFM migration used to minimize downtime & allow multiple FS changes (block & inode size): 100M files, incremental migration, several issues (solved in 4.2.2?)

GPFS @CC-IN2P3: Filesystems (and filesets)

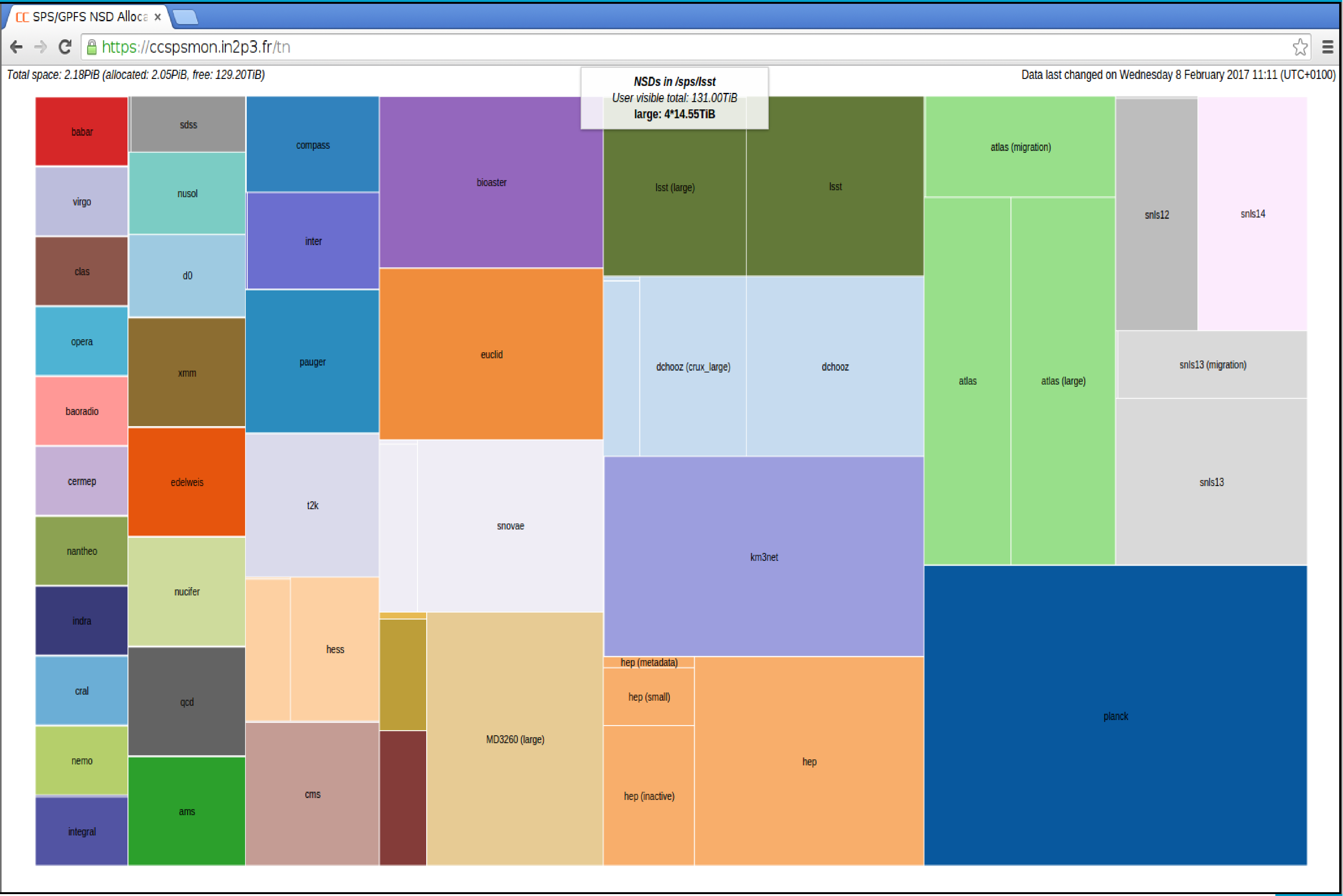


Total space allocated: 2.05PiB

Data for Friday 10 February 2017 11:31 (UTC+0100)

Space allocated
/sps/lsst: 131.00TiB

GPFS @CC-IN2P3: Filesystems (NSDs & pools)

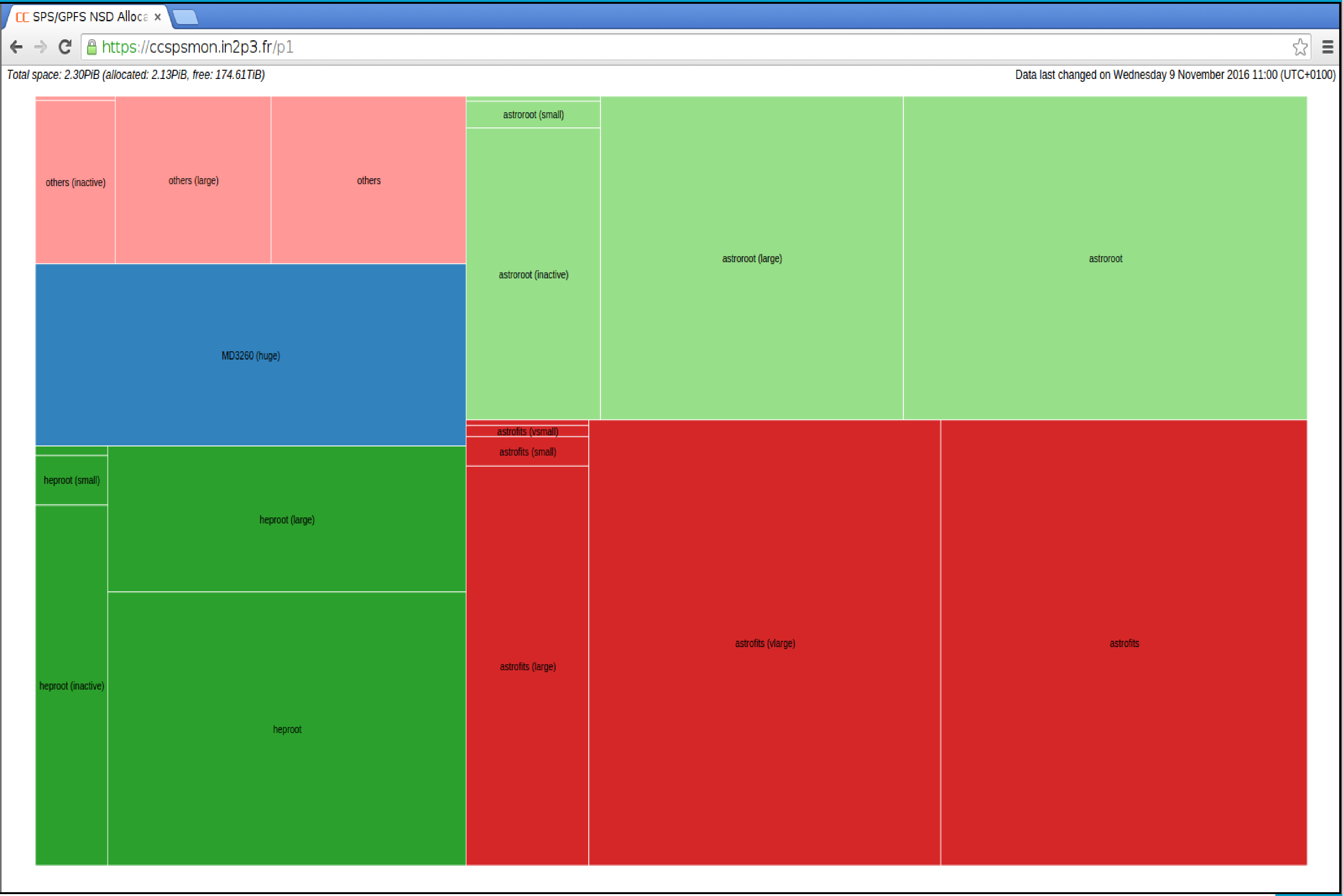


Total space: 2.18PiB (allocated: 2.05PiB, free: 129.20TiB)

Data last changed on Wednesday 8 February 2017 11:11 (UTC+0100)

NSDs in /sps/lsst
User visible total: 131.00TiB
large: 4*14.55TiB

GPFS @CC-IN2P3: Less filesystems (NSDs & pools)



Total space: 2.30PiB (allocated: 2.13PiB, free: 174.61TiB)

Data last changed on Wednesday 9 November 2016 11:00 (UTC+0100)

GPFS @CC-IN2P3: Spectrum Scale features

- Plan to use Zimon + Grafana to replace some locally developed monitoring tools
- Non intensive use of ILM
- Newer features enabled after AFM migration:
  - Fileheat & HAWC
  - maybe LROC (LSST specific nodes)
- Plan to use CES for NFSv4 w/ KRB5 auth:
  - access GPFS from VMs w/ user controlled images
  - maybe replace native client for some nodes
- TCT & CES require CCR, which has some issues in our environment

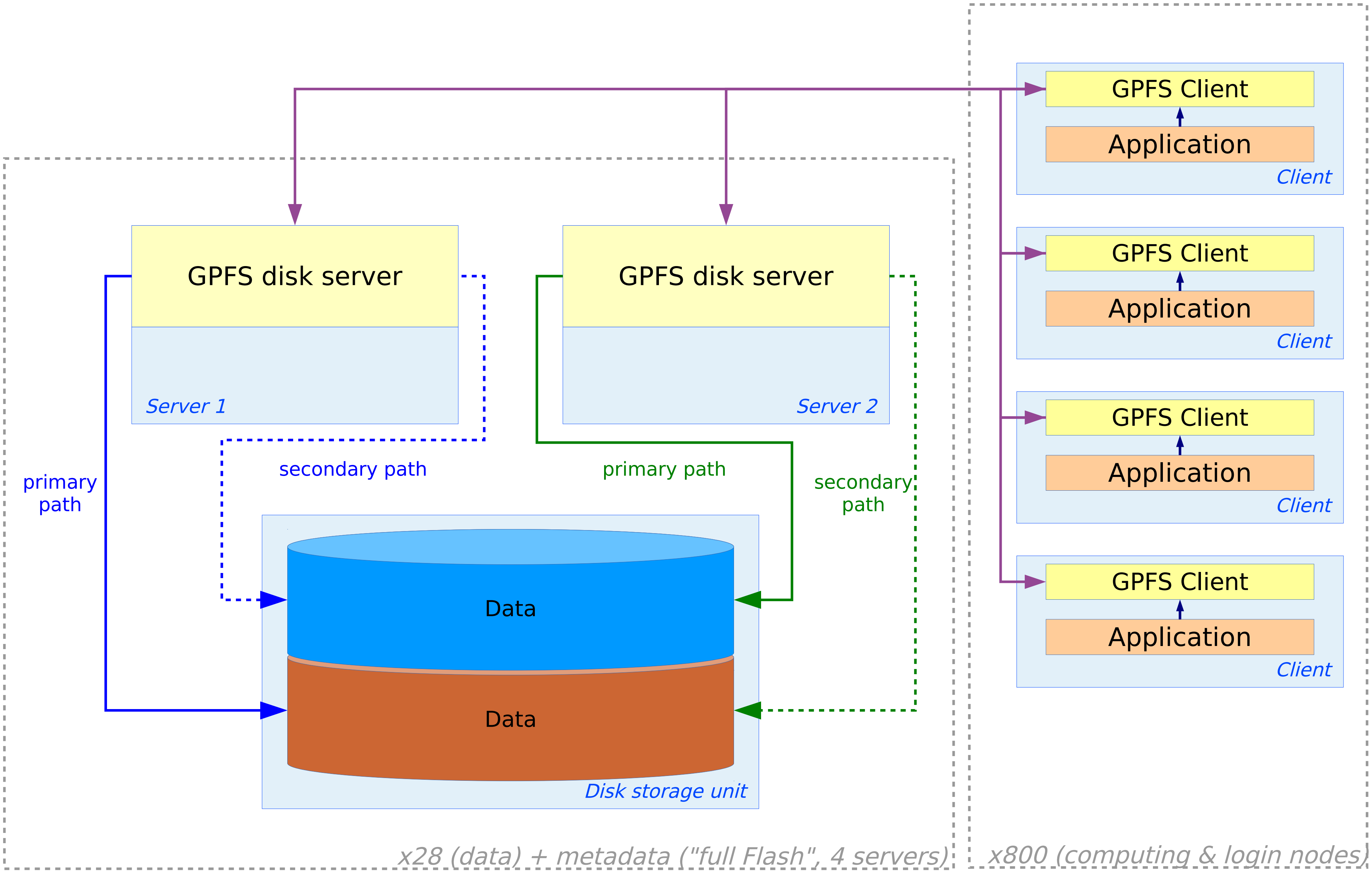
GPFS @CC-IN2P3: More

- Local developments for:
  - quota management delegation (& filesets creation/destruction)
  - multiple monitoring tools
- License costs:
  - Licenses for IN2P3 & CEA-IRFU, not just CC-IN2P3; > 100 server nodes, > 1400 client nodes;
  - TCT (Spectrum Scale Advanced)
  - GHI
- Considering wether an appliance (IBM ESS or DDN GridScaler) would make sense, both in terms of volume and license costs

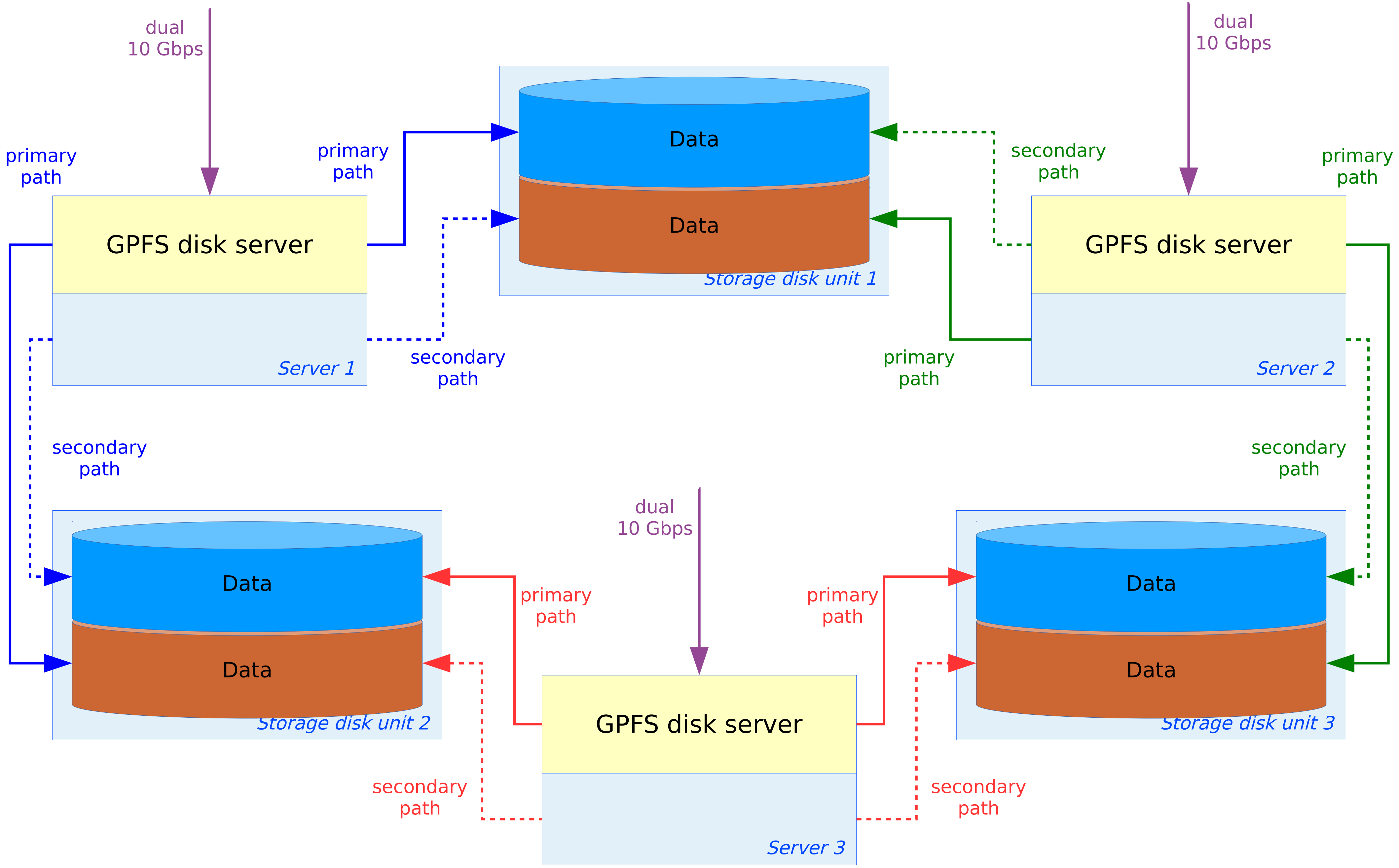
GPFS @CC-IN2P3: Hardware infrastructure

- Data on Dell MD3260 & 2 IBM DCS3700 disk units w/ 2 servers each
- Dell MD3x60/IBM DCS3xy0: LSI/NetApp OEM disk units, w/ 60 drives & dual redundant controllers
- 5th storage hardware generation for SPS: Dell MD3460 w/ 2 servers
- Newest procurement: 3 servers share 3 MD3460
- Metadata on full-Flash storage (LSI/NetApp EF560)
- All disk servers connected w/ 10 Gbps interface(s)
- Client nodes w/ 1 Gbps except parallel computing & login nodes (10Gbps)
- Cheap server for **very** inactive data: Dell R510

GPFS @CC-IN2P3: Hardware for I/O nodes



GPFS @CC-IN2P3: Hardware for new I/O nodes



Ceph @CC-IN2P3: WiP

- Use cases:
  - OpenStack Cinder replacement
  - maybe storage for Spark jobs
- RBD only (for starters)
- Ceph release: Kraken
- 6 x Dell R730xd disk servers, each w/:
  - 10 x SAS-NL 8 TB disks
  - 2 x SAS 400 GB write intensive SSDs (for journals)
  - 10 Gbps Ethernet
- 3 VMs for Ceph monitors
- 3 way replication (EC not available for RBD)