EXPLOR:
Un Ensemble de Calcul Scientifique Pour la LORraine

Journées méso-centres - 25 septembre 2017

Présentation

- Ressources HPC / salle machine
- Architecture réseau
- Stockage
- Vue utilisateur / Vue administrateur
- Perspectives

Le Centre EXPLOR

Maison de la Simulation Lorraine

- Appui au calcul/Expertise
- Réseau métier « calcul »
- Entreprises
- Formation
- Animation Scientifique

Mésocentre

- Infrastructure HPC
  - Calcul
  - Stockage
  - Analyse
- Services HPC

HPC

Détails HPC : 107 Tflops

STD	64 C6320	2x E5-2683v4, 2.1 GHz, 16c 128 GB DDR4, 2 400Go SSD MU, OPA PCIex16	2048 cores **68.8 Tflops**
HF	16 R630	2x E5-2637v4, 3.5 GHz, 4c 128GB DDR4, 2 1.2To, 10GbE 10k	128 cores **7.1 Tflops**
Ivy	12 C6220	2x E5-2640v2, 2.0 GHz, 8c 64GB DDR3, 1.0To 7.2k, IB QDR PCIe2.0	192 cores **3.1 Tflops**
k20	5 Supermicro	4x K20m, 2x E5-26xx, 8c 64GB DDR3, 1.0To 7.2k, IB QDR PCIe2.0	20 K20m **24.7 Tflops**
Phi	1 R720	2x E5-2640v2, 2.0 GHz, 8c, 64GB, 1.0To IB, PCI16x Xeon Phi 5110P + Xeon Phi 7220P	144 cores **3.6 Tflops**

données projets HOME: brut 960 To, utile 708 To

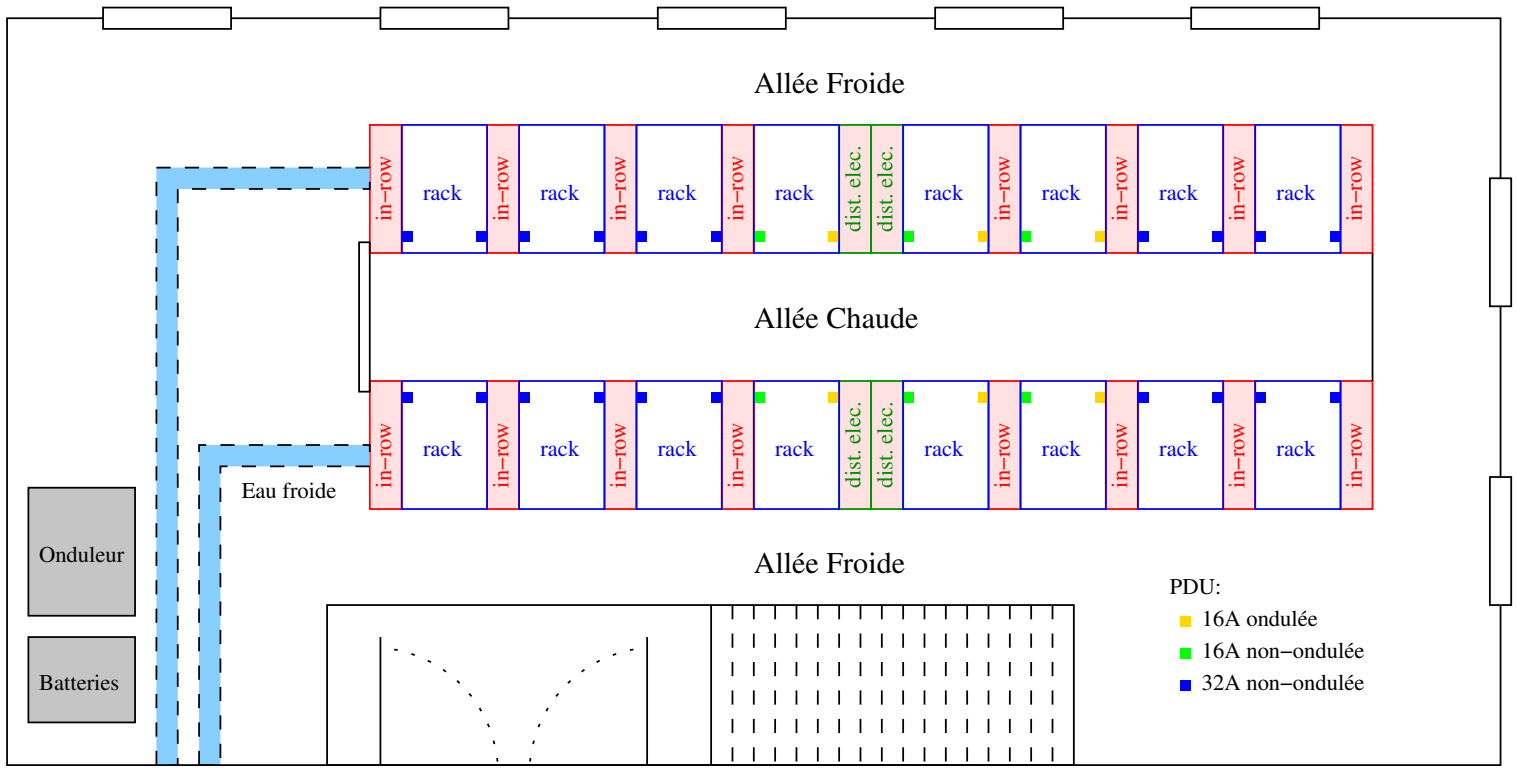
2 Storage configurations	1x MD3460 (20 DD) + 2 x MD3060e (20 DD)
120 Disks in storage array	Hot Plug 8To NL-SAS 12Gbit/s 512e 7.2k 3,5", PI
8 NFS servers	2x E5-2630v4 (2.2G, 10c) 128Go RAM 2x300G 15k RAID1, 4x10Gb ethernet

Détails HPC

Élements Informatiques			Somme par modèle			
Type	Modèle	Quantité	W (100%)	Poids (kg)	CFM	U
Networking	N4064F 10Gb	15	11310	158.4	705	15
	N2048 1Gb	3	258	11.1	50	3
	S6000-ON 40Gb	2	742	14.6	94	2
	S5030Q IB QDR	1	0	0	0	1
	H1048 OPA	6	2448	40.2	420	6
Serveurs	R620	3	0	0	0	3
	R720	1	0	0	0	1
	R630	28	11340	473	854	28
	C6220 II	3	3135	122	0	6
	C6320	16	22592	656	1861	32
	S.M. X9DRG-HF	5	0	0	0	10
Stockage	MD3460	2	2406	210	462	8
	MD3060e	4	4900	420	924	16
	MD3400	1	431	9	28	2
	R730xd	1	481	32	18	2

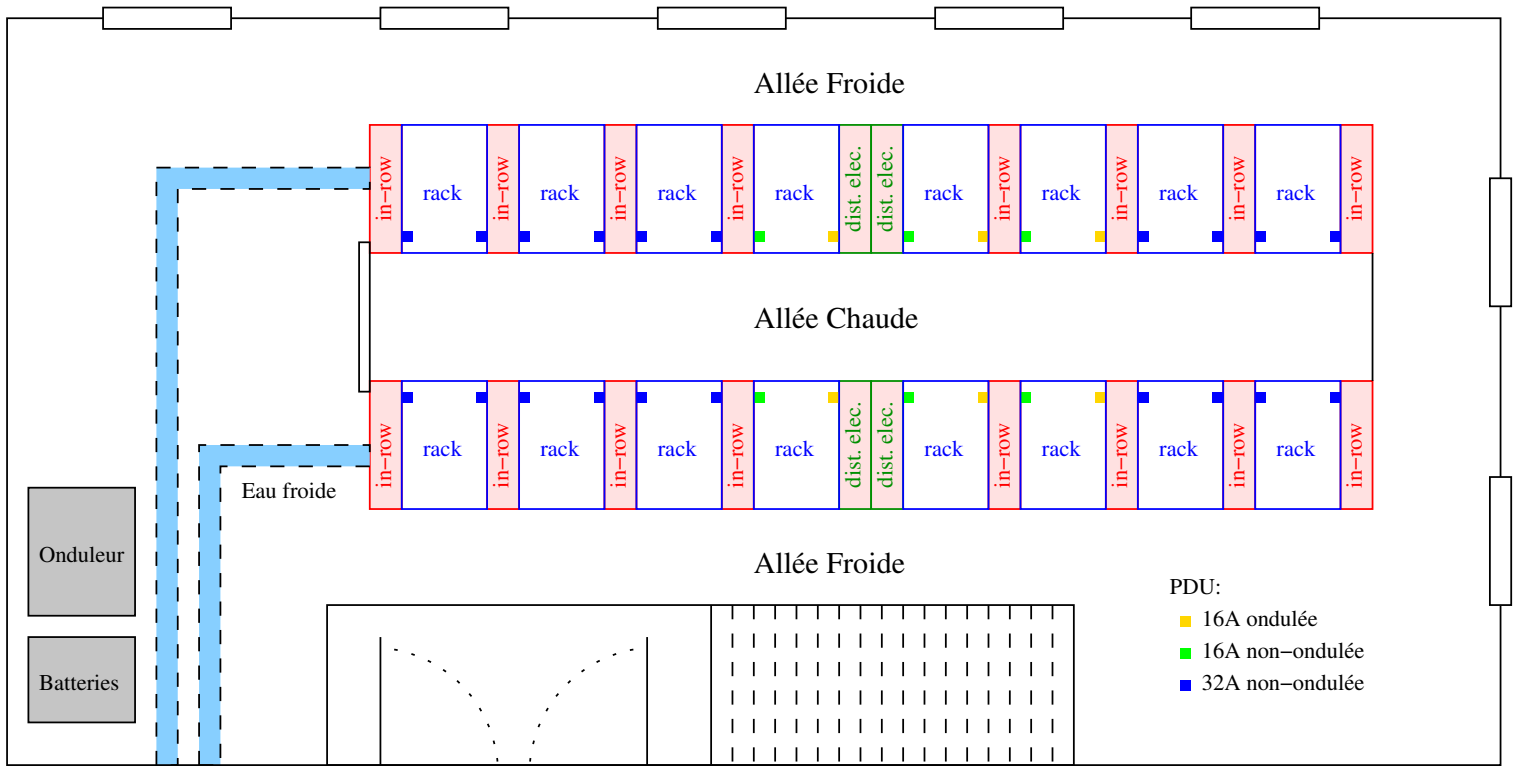
60 kW 19%	2147 kg	5415.5	139 U 20%

salle EXPLOR



- 93 m^2, 293 m^3, allée chaude confinée 23 m^3
- 320kW de Froid: 16 Inrows de 20kW

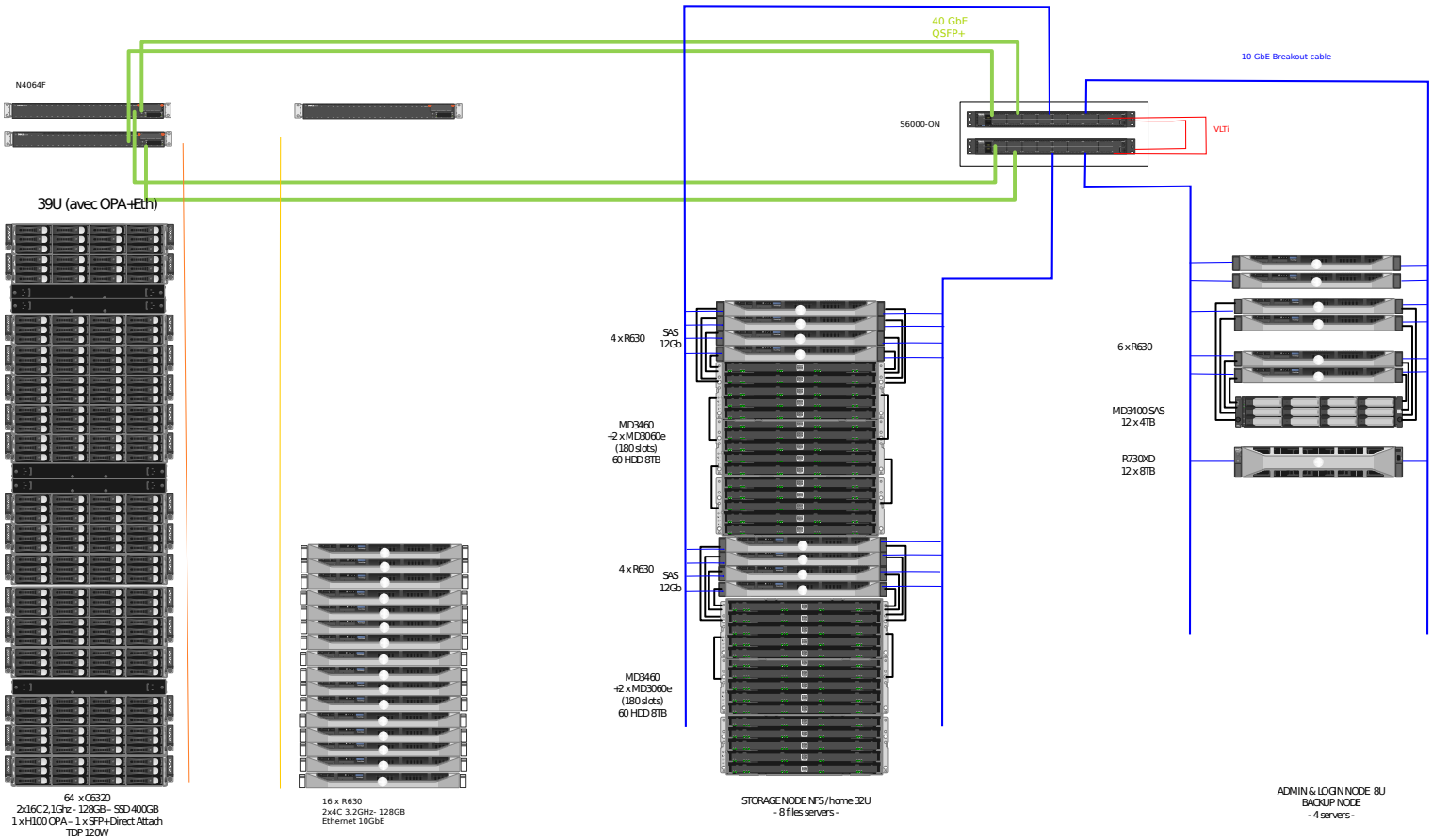
salle EXPLOR



- 93 m^2, 293 m^3, allée chaude confinée 23 m^3
- 320kW de Froid: 16 Inrows de 20kW

- Onduleur 80 kVA
  - ✓ 6 racks de service (1PDU/2)
  - ✓ 16 Inrows
  - ✓ suvi par SNMP

Schéma physique globale



Réseau physique

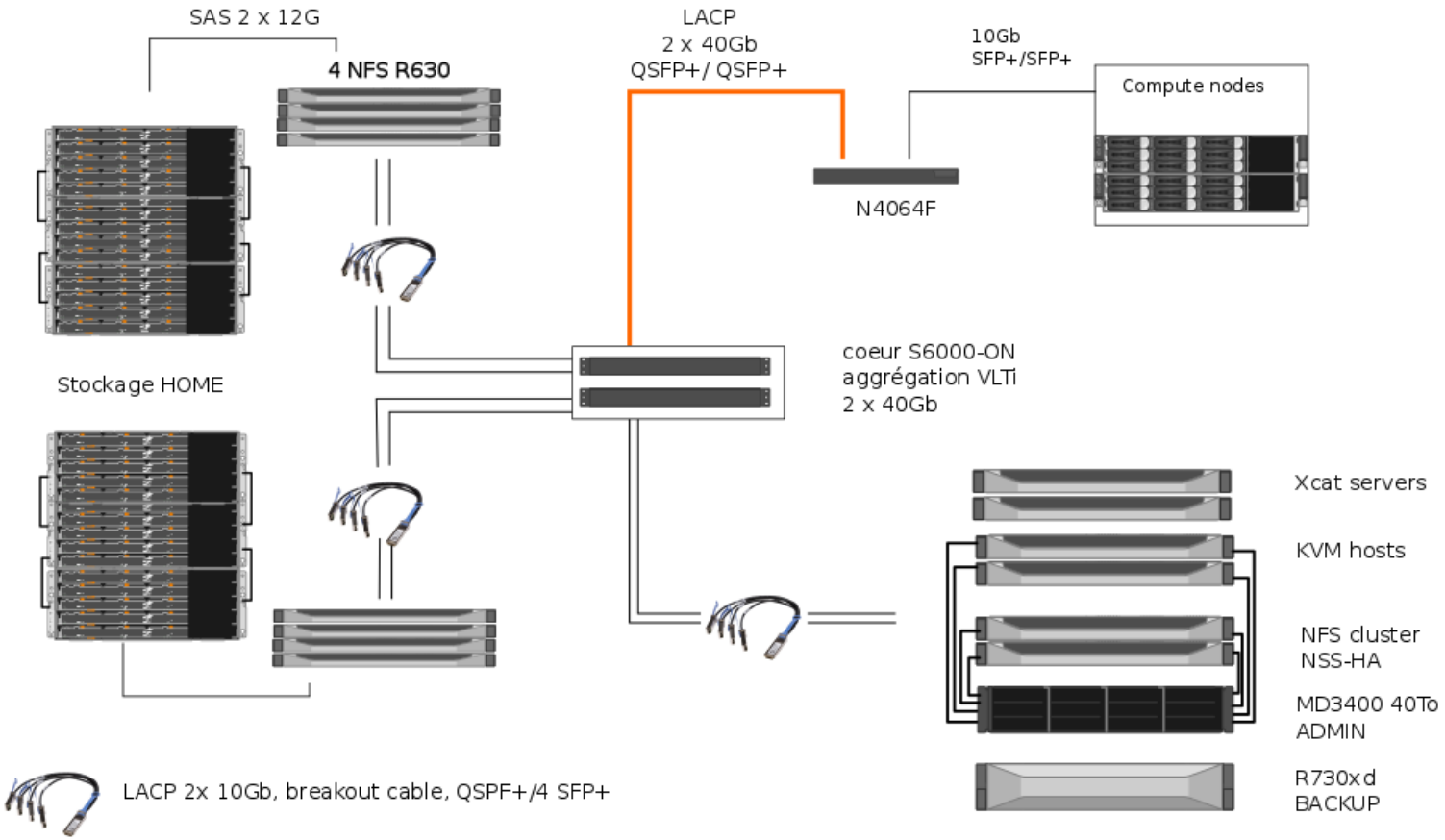
Ethernet : transport data HOME

- Inter-co université en 2 x 10Gb (LACP)
- 2 switches de coeur (S6000-ON) 2 x 40 Gb agrégés avec un VLT
- 16 switches TOR (N4064F) 2 x 40Gb (LACP) vers le coeur
- Machines de service 2 x 10Gb (LACP) vers le coeur
- Noeuds de calcul simple attachement 10Gb (1Gb pour matériel ancien)
- ☞ Economie de câbles, partage d'interface pour le management iDRAC

Faible latence : inter communication entre noeuds

- 5 switches OPA (H1048-OPF) en non bloquant pour le rack Standard
- 1 switch OPA pour l'ADMIN et demain le stockage BeeGFS
- 1 switch Infiniband Mellanox QDR

Stockage HOME



- stockage HOME des projets
- Scratch local sur les noeuds
- stockage SAVEDIR 10Go par utilisateur
- Sauvegarde du SAVEDIR sur R730xd 96To Brut

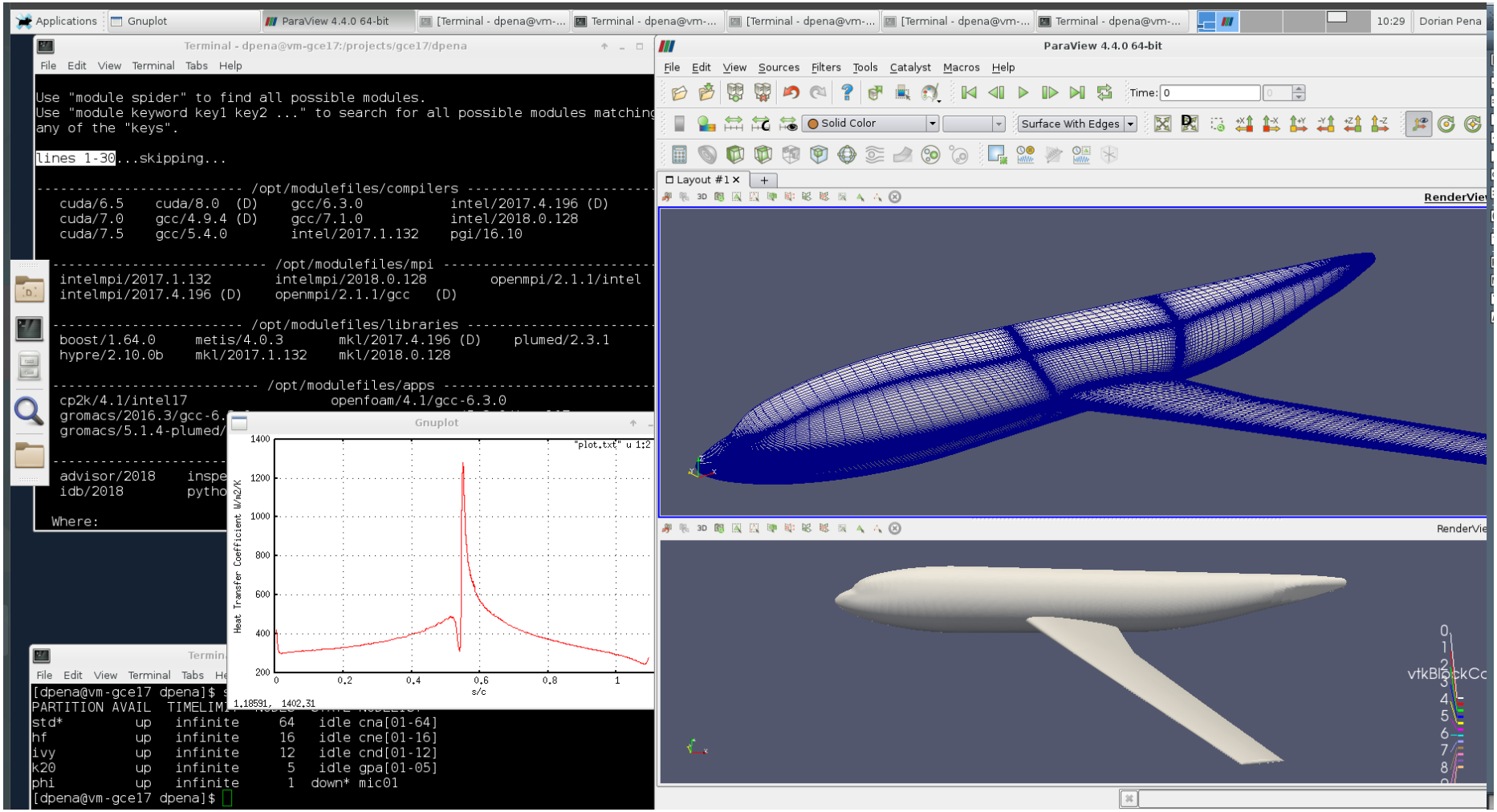
Vue utilisateur

- Connexion à distance par SSH ou X2Go (environnement graphique Xfce)

☞ `ssh -p 50399 zoro@gw39.explor.univ-lorraine.fr`

- Machine virtuelle propre au projet
- 1 compte anonyme par utilisateur
- 1 projet anonyme avec un espace de partage commun (1To $\longrightarrow$ 4To)
- Environnement HPC : montage NFS (avec autofs) du stockage
- Environnement de Module : Lmod 7.4
- File de soumission : Slurm 17.02

Vue utilisateur X2Go



Vue administrateur

OS

- Noeuds : Centos 7.2
- Machines de services, frontales : Centos 7.3
- Logiciels libres et outils opensource
- Sécurité : Segmentation, protection par contexte, pare-feu,...

Technologie et outils

- Virtualisation avec KVM (script virt-install, kickstart)
- Xcat: Déploiement des OS et MAJ des matériels
- Authentifcation centralisée : Openldap avec la rfc2307bis
- Supervision/Monitoring : Shinken/Graphite/Carbon, Ganglia
- Ticket/Inventaire/Plan d'adressage : glpi, netbox
- Documentation technique interne et journal de bord : mediawiki
- Sauvegarde : Bacula
- Orchestration : Ansible

Vue administrateur : Réseau logique

Pare-feu

- Machine virtuelle branchée sur 4 bridges
- Pare-feu avec firewalld
  - ✓ Fonction de routeur
  - ✓ Place les règles proprement dans les chaines PREROUNTING, POSTROUTING et FORWARD
  - ✓ Exemple : Translation de ports et d'adresses, ex: 1 port SSH par environnement virtuel

Zones

- Utilisation de VLAN pour séparer l'administration de la production
- ZONE DMZ : VM utilisateur (frontale)
- ZONE ADMIN : administration des noeuds, de l'onduleur, ...
- ZONE PROD : réseau commun

Vue administrateur : LDAP

Serveurs et Version

- Deux VM sur 2 KVM hosts CentOS release 7.3.1611
- OpenLDAP daemon version : slapd 2.4.40

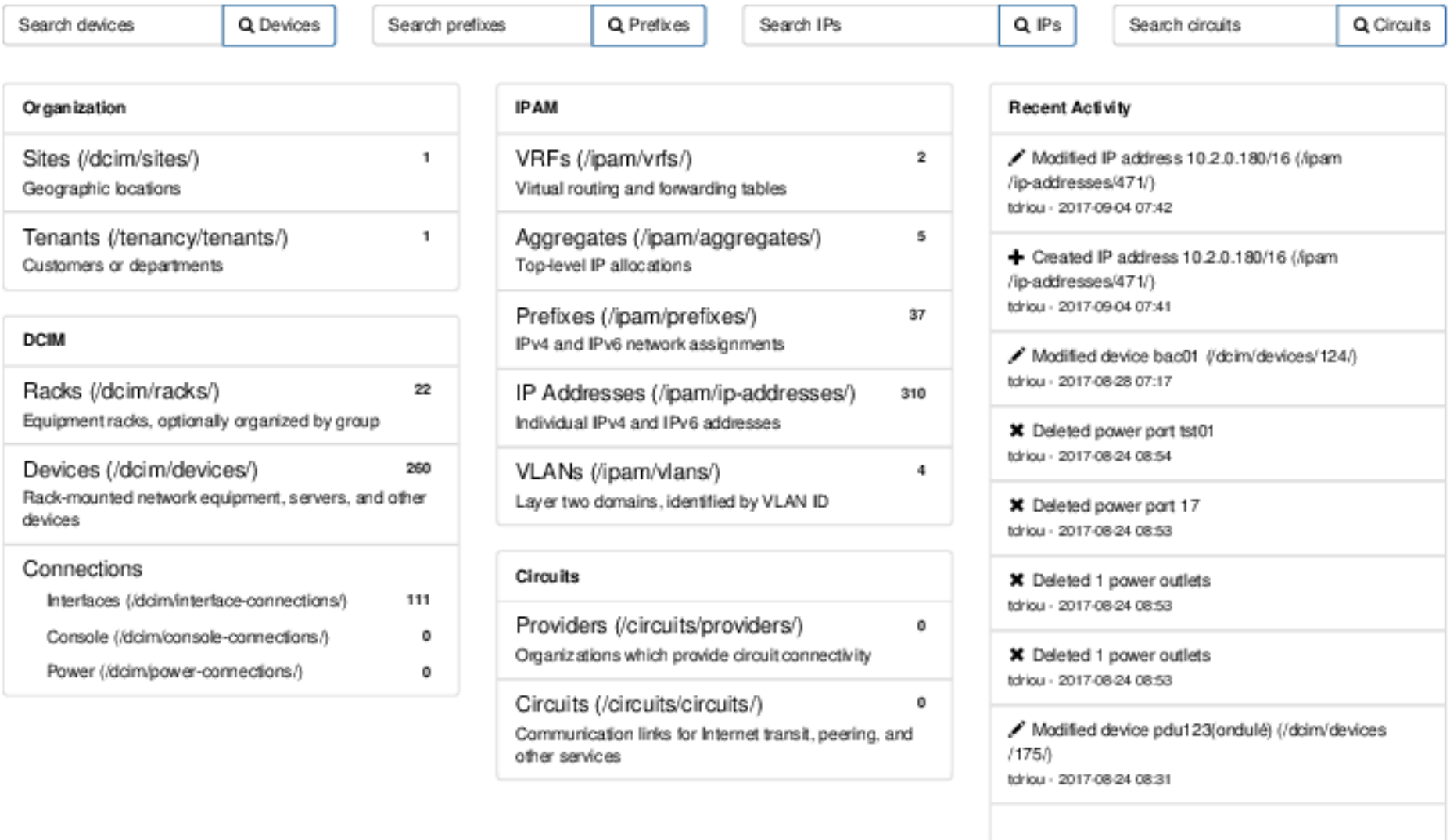
Fonctionnalités

- Annuaires chiffrés avec StartTLS
- Replication multi-maîtres (Config et data) $\longrightarrow$ overlay olcSyncRepl
- Gestion dynamique des groupes et utilisateurs $\longrightarrow$ overlays memberOf, refint
- Création dynamique de groupes

Gestion centralisée avec SSSD

- Authentification
- Filtre d'accès
- Montage automatique avec AutoFS
- Gestion droits SUDO

Gestion des plans d'adressage et du matériels



Gestion des plans d'adressage et du matériels

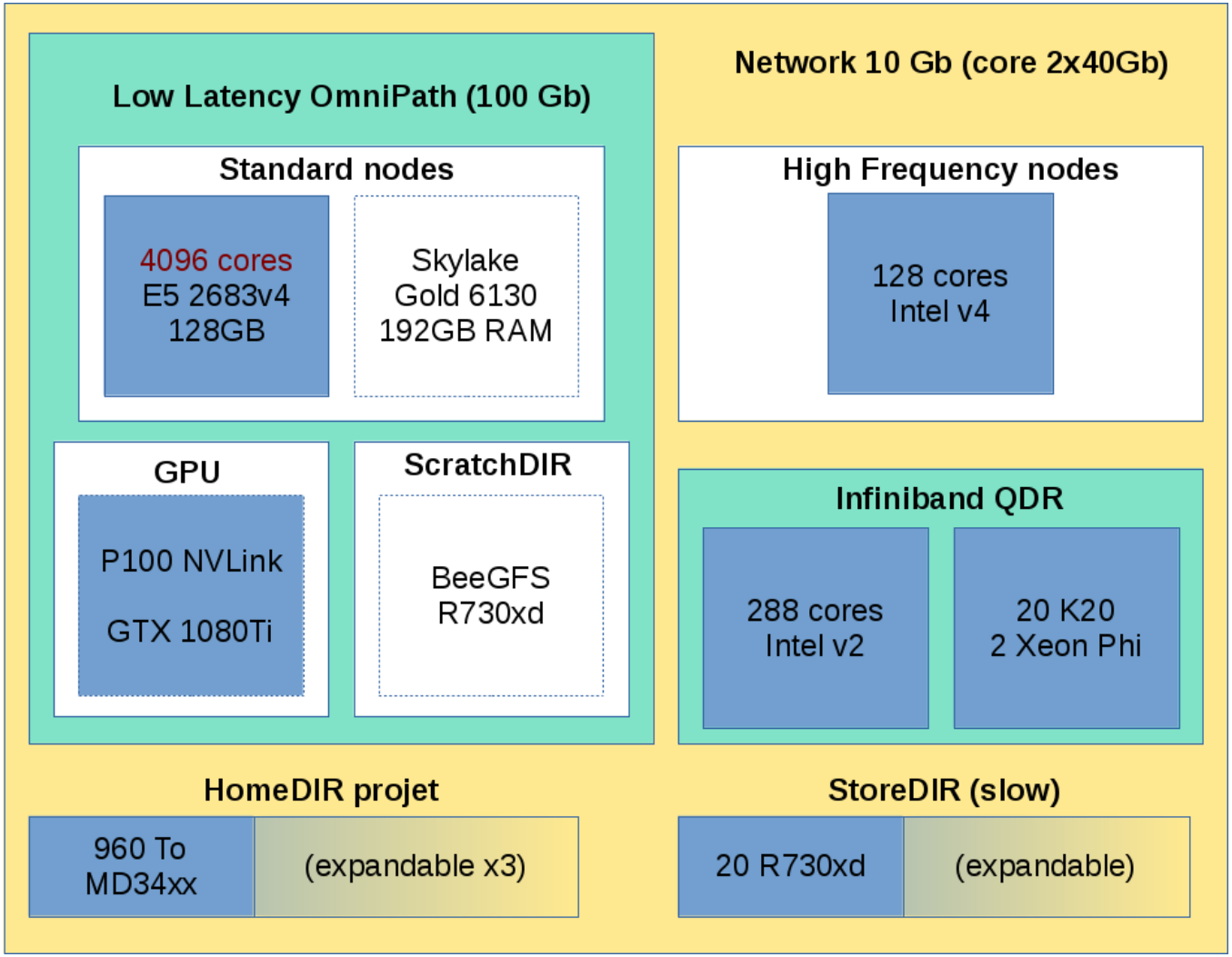
Racks

Name	Site	Group	Facility ID	Tenant	Role	Height	Devices	Utilization
C01	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	19	40%
C02	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	89	92%
C03	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	3	2%
C04	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	3	2%
C05	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	3	2%
C06	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	3	2%
C07	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	3	2%
C08	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	3	2%
C09	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	28	54%
C10	SITE ARTEM	—	—	EXPLOR team	Compute nodes	42U	3	2%
S01	SITE ARTEM	—	—	EXPLOR team	Services	42U	4	4%
S02	SITE ARTEM	—	—	EXPLOR team	Services	42U	29	26%
S03	SITE ARTEM	—	—	EXPLOR team	Services	42U	11	42%
S04	SITE ARTEM	—	—	EXPLOR team	Services	42U	11	42%
S05	SITE ARTEM	—	—	EXPLOR team	Services	42U	5	9%
S06	SITE ARTEM	—	—	EXPLOR team	Services	42U	3	2%

Supervision

- Solution Shinken / graphite / carbon
- 377 sondes,
- Surveillance des matériels par snmp, ipmi, SSH : switchs, onduleur, PDPM, PDU, idrac serveurs, stockage, OS
- Développement de nos sondes à partir des MIBS

Evolutions


Low Latency OmniPath (100 Gb)
Network 10 Gb (core 2x40Gb)
Standard nodes
4096 cores
E5 2683v4
128GB
Skylake
Gold 6130
192GB RAM
High Frequency nodes
128 cores
Intel v4
GPU
P100 NVLink
GTX 1080Ti
ScratchDIR
BeeGFS
R730xd
Infiniband QDR
288 cores
Intel v2
20 K20
2 Xeon Phi
HomeDIR projet
960 To
MD34xx
(expandable x3)
StoreDIR (slow)
20 R730xd
(expandable)

Evolutions

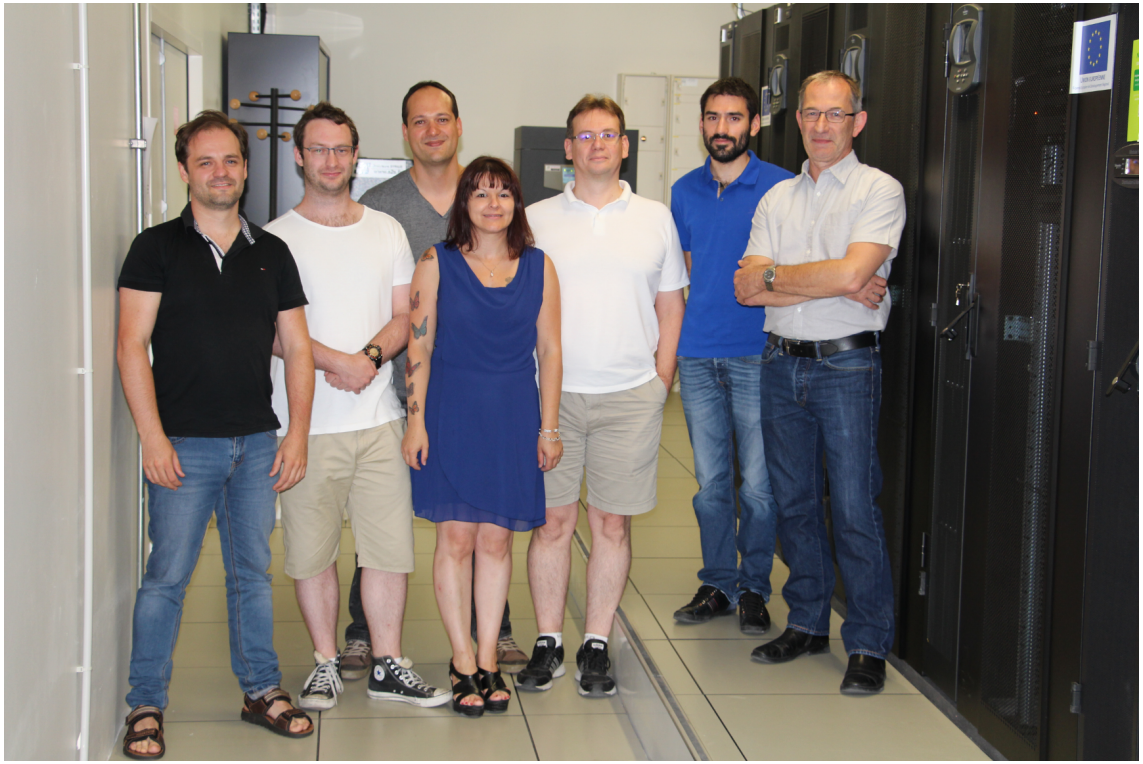
Matériels

- CPU std x 2 ☞ 4096 coeurs
- Noeuds GPU P100 + noeuds GTX 1080Ti avec OPA
- Stockage BeegFS pour le SCRATCHDIR avec OPA
- Stockage lent, serveurs R730xd
- 8 Machines de services dont visualisation graphique

Services

- Visualisation avec acceleration graphique X2go + VirtualGL
- Centralisation des logs, rapports
- Base de données centralisée avec workflow
- Evolution vers Kerberos
- Ajout VLAN sur le coeur de réseau, redémarrage à distance

Equipe et remerciements



- S. Bonenberger, G Monard, L. Sauder, D. Pena.,A. Willas et T. Driou
- UL, CNRS, Etat, Région, FEDER
- Grid5000, Romeo, Alsa calcul, CRIHANN, les centres nationaux