# Solutions pour le stockage et l'archivage chez Mathrice

## ANF2018 - MAT1

Laurent Azema

4 Décembre 2018

mathrice

# La PLM de Mathrice

- Mathrice : réseau métier des personnels informatiques des laboratoires de mathématiques
- Mathrice : GDS2754, groupement de service (structure légère du CNRS regroupant des volontaires pour proposer des services)
- PLM la Plateforme en Ligne pour les Mathématiques
- Ensemble de services à destination des chercheurs et des ASR des laboratoires de la communauté de recherche en maths en France
- PLM-team : 12 personnes de 9 laboratoires participent à son exploitation

# L'infrastructure de la PLM

- Opérée sur sa propre infrastructure
- Hébergée par des laboratoires de maths
- 4 sites de production : Bordeaux, Lille , Angers et Lyon
- 1 serveur de fichiers NFS par site : filer
- Des serveurs hôtes de machines virtuelles (VM) : hyper
- 1 site de sauvegarde : Grenoble
- Évolution de la solution en fonction des besoins

mathrice

# Première génération

- Configuration du filer :
  - serveur Dell PowerEdge R510
  - disques SAS 10ktpm constituant un RAID5
  - partitionné en plusieurs LV ext4
- Exports NFS :
  - Images-disque des VM montées sur des hyperviseurs VMWare ESX
  - Données montées depuis des VM : répertoires de travail, boîtes aux lettres, sites web...
- Sécurisation :
  - rsnapshot depuis serveur sauvegarde d'un autre site
  - restauration distante donc lente

# Deuxième génération

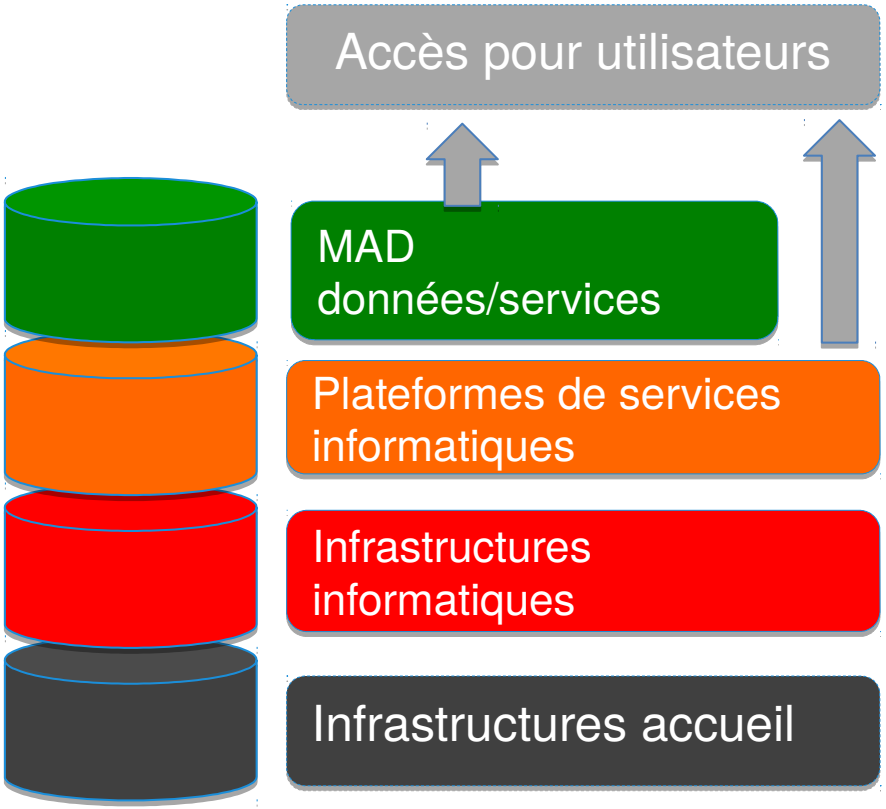
- Configuration du filer :
  - serveur Dell PowerEdge R720xd
  - disques SAS 10ktpm en JBOD (ou un RAID0 par disque)
  - tests HammerFS sur DragonFly, avant ZfsOnLinux sur Debian puis ZFS sur FreeBSD
  - 1 zpool de plusieurs raidz2 (RAID6) avec plusieurs dataset ZFS
- Exports NFS :
  - Images-disque des VM montées sur des hyperviseurs libvirt/KVM
  - Partitions data des VM demandant de l'espace de stockage
- Sécurisation :
  - $2^{nd}$ serveur sur site identique au filer : replica
  - copie régulière de snapshot ZFS
  - restauration ou échange des rôles avec filer selon l'incident
- Sauvegarde distante via L3VPN RENATER (interconnexion des réseaux locaux)
  - Dell R730xd avec 11 disques 10T SASNL pour 1 zpool raidz3
  - BackupPC de partitions des filer et VM
  - copie de snapshot ZFS pour les données $>$ 500Go sur filer

# Évolutions à venir

- Suivre la croissance de l'espace de stockage utilisateur :
  - modularité des baies de stockage externes
- Plateformes de test OpenStack et OpenShift
  - service d'hébergement de VM à la demande
  - service de conteneurs pour PLMlab/CI
  - service hébergement web sur conteneurs
  - service notebook sur conteneurs
- Évolution du stockage :
  - cinder NFS/ZFS (Storage as a Service)
  - hyperconvergence
  - stockage objet
- Devenir client d'une infrastructure ?
  - Voir projet Infranum 2020 du ministère [1]

1. https://jcad2018.sciencesconf.org/data/7_20180921_infranum_jcad_2018.pptx

# Vision InfraNum 2020




Métiers, Communautés, Organismes, établissements, entreprises….

Données, métadonnées, applications et services dédiés

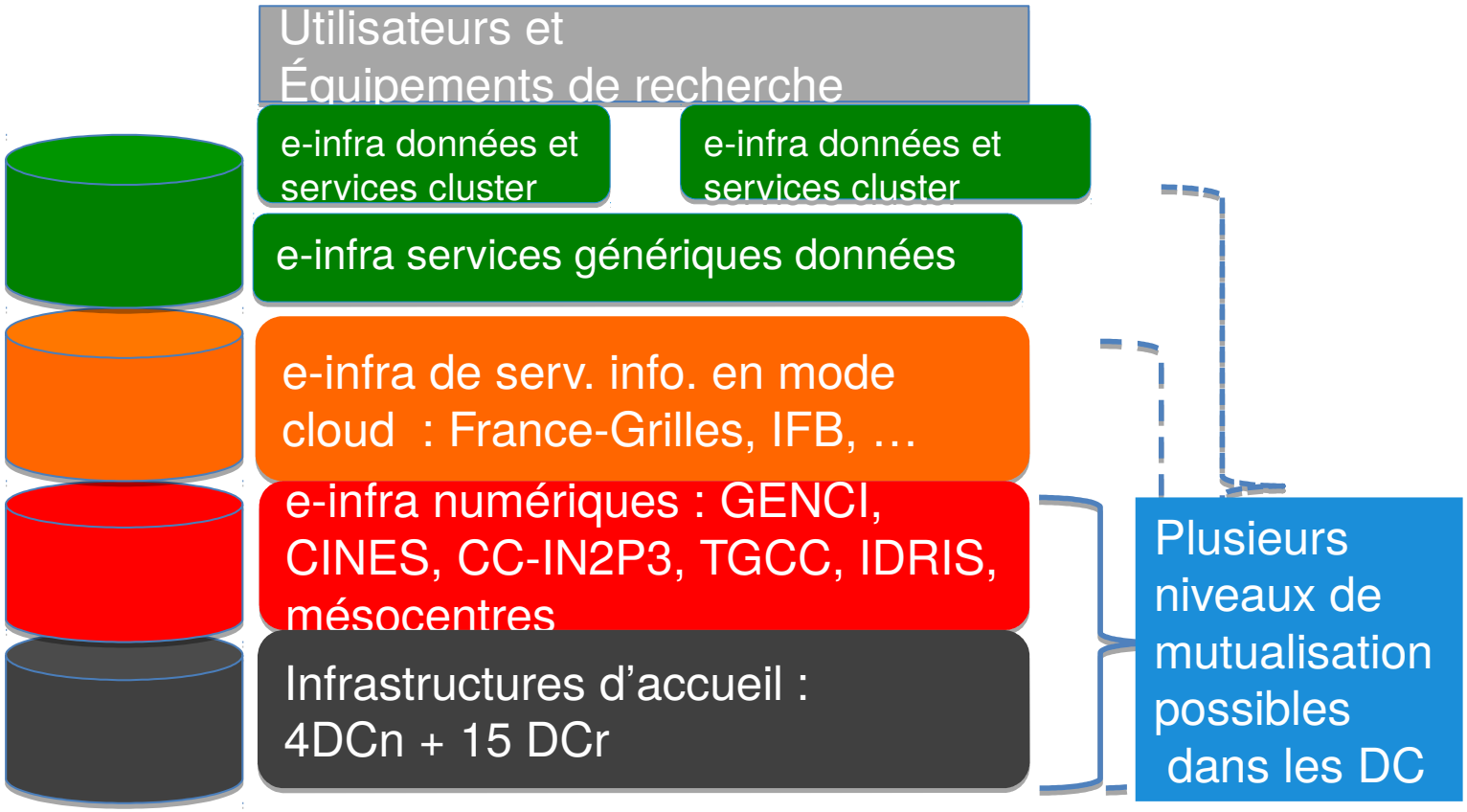
Architectures logicielles systèmes & réseaux, accès au calcul, traitement des données…

Architectures matérielles systèmes & réseaux, calcul, stockage…

Immobilier, énergie, climatisation, réseau, gardiennage…

Source JCAD2018 : Marie-Christine Plançon, *chef de projet "modernisation des infrastructures et des services numériques" au MESRI*

# Vision InfraNum 2020




Source JCAD2018 : Marie-Christine Plançon, *chef de projet "modernisation des infrastructures et des services numériques" au MESRI*

# Les données dans la PLM

Les données des différents services se trouvent sous différentes formes

- De fichiers-disque de VM
- De fichiers texte
- De Systèmes de Gestion de Bases de Données

# Des données sous forme de blocs

Virtualisation libvirt/KVM avec des fichiers RAW pour les disques

- 1 fichier = ensemble de blocs d'un disque monté sur une VM
- LVM pour étendre les partitions par agrégation de fichiers RAW
- Système de fichiers ext4 gère les blocs de chaque volume logique
- 1 export NFS pour chaque VM (filer vers hôtes (hyper))
- entrée/sortie sur la VM => accès NFS de la machine hôte

Sécurisation sur la PLM :

- Snapshot ZFS sur filer envoyé sur replica (ZFS send/receive)
- Perte du filer => bascule manuelle montages NFS depuis replica
- Equivalent à démarrage de VM dans version passée après arrêt brutal
- Perte de données en mémoire ; perte des écritures depuis sauvegarde

mathrice

# Des fichiers texte

Certaines applications utilisent directement des fichiers texte.

- Mail : postfix, dovecot
- Fichiers froids de disque : sftp
- Pages web simples personnelles sur disque : apache
- Sites web simples hébergés : apache
- Listes de diffusion : postfix, sympalist

D'autres gèrent eux-même les données dans des fichiers

- Dépôts git de PLMlab : gitlab
- Fichiers chauds de PLMbox : seafile

# Le mail : postfix et dovecot

Mode de stockage :

- Files entre modules postfix = répertoires (/var/spool/postfix/)
- Boîtes aux Lettres au format Maildir [2] (/home/login/Maildir/)
- 1 fichier = 1 mail => facilite l'accès par plusieurs applications
- Index de chaque dossier pour accélérer l'accès aux mails [3]

Localisation des données selon les usages :

- client POP : référence de la BaL = client
- client IMAP : référence de la BaL = serveur
  - travail hors-ligne du client sur copie locale des mails
- interface web : le serveur webmail se connecte en IMAP à la BaL

2. https://wiki1.dovecot.org/MailboxFormat/Maildir
3. https://wiki1.dovecot.org/IndexFiles

# Espace de stockage personnel

Plusieurs usages de disque.math.cnrs.fr :

- Stockage à moyen terme (fichiers froids) de travaux de recherche
- Sauvegarde d'un portable (rsync, rsnapshot...)
- Travail en ligne sur montage réseau webDAV
- Pages web personnelles statiques (pas de PHP)

Fonctionnement :

- 1 dataset ZFS du filer exporté en NFS par utilisateur avec son quota
- Accès par copies SSH (shell scponly) : sftp, rsync, unison, filezilla...
- Répertoire upload/ pour stockage individuel
- Répertoire public/ => `http://loginPLM.perso.math.cnrs.fr/`

# L'hébergement de sites web

Hébergement de sites web institutionnels (colloques, projets, équipes...)
Fonctionnement :

- Répertoire htdocs par site en partage NFS par filer
- Gestion des droits sur fichiers avec NFSv4ACL
- Frontal SSH cms.math.cnrs.fr pour les webmestres
- Affectation site web à 1 VM web (différent apache/PHP)
- Mandataire inverse rprox.mathrice.fr des VM web

Sécurisation de ces 3 services (mail, disque et web) :

- BackupPC des filer depuis sauvegrenoble2
- Répliqua filer localement pour reprise rapide
- Restauration répertoire/fichier possible depuis replica

# Les listes de diffusion : sympalist

Plusieurs moteurs :

- listes.math.cnrs.fr, listes.mathrice.fr
- de laboratoires : listes.cpht.polytechnique.fr, listes.idpoisson.fr...
- de partenaires : listes.rnbm.org, smf.emath.fr, listes.resinfo.org...

Mode de stockage :

- Base MySQL pour configuration des listes et gestion des abonnements
- Répertoires et fichiers mail pour routage sympalist et postfix
- Fonction archive (MHonArc) => stockage des envois sur serveur

Sécurisation sur la PLM :

- BackupPC des archives sur la VM
- Restauration distante possible sur demande

# Les dépôts Git : PLMlab

Gestion de projets : `https://plmlab.math.cnrs.fr/`
Chaque projet comprend par défaut 2 dépôts git (dépôt et wiki)

Différents usages de l'outil :

- interface web : travail en ligne sur les données du serveur
  - webIDE (édition en ligne non collaborative)
  - gollum (interface du wiki)
- copie de travail locale couplée au dépôt du serveur :
  - arborescence de fichiers gérés (ou non) par Git
  - validation locale des modifications pour créer une version (commit)
  - serveur connaît seulement les modifications poussées vers lui
  - attention, avec les branches, la synchronisation peut être partielle

# Les dépôts Git : PLMlab

Localisation du stockage des données d'un dépôt :

- sur la copie de travail : sous .git/objects
- sur le serveur plmlab : sous data/repositories/ (montage NFS depuis filer)

Structure des données [4] :

- 1 objet = 1 fichier (compressé avec zlib)
- répertoire+nom objet = hash SHA1(entête+contenu non compressé)
- 3 types d'objet
  - blob : contenu d'un fichier
  - tree : contenu d'un répertoire (liste d'objets de type blob ou tree)
  - commit : informations sur une version
    - un objet racine de type tree
    - un objet parent de type commit
    - des informations texte (auteur, mail, commentaire...)

mathrice

4. https://git-scm.com/book/en/v2/Git-Internals-Git-Objects

# La PLMbox : seafile

Espace synchronisé voire partagé de fichiers (données chaudes)

Interface `https://plmbox.math.cnrs.fr/`

- Gestion individuelle de ses bibliothèques de fichiers
- Gestion droits d'accès des autres utilisateurs
- Création de liens de téléchargement (voire télédépose ! ! danger)
- Politique d'historique (attention à la cohérence avec le quota)
- Accès à l'historique des fichiers pour restaurations ponctuelles
- Editeur de fichiers en ligne

Synchronisation locale de bibliothèques

- Répertoire local entièrement synchronisé avec une bibliothèque
- Possibilité de travail hors-ligne
- Résolution des conflits : renommage de chaque version avec date

# La PLMbox : seafile

Fonctionnement :

- Gestion d'objets sous forme de fichiers comparable à git
- *blocks* <=> *blob* de git
- *fs* <=> *tree* de git
- *commits* <=> *commit* de git
- Destruction des objets inutiles par rammasse-miettes

Sécurisation sur la PLM :

- snapshot ZFS envoyé vers replica (marche normale)
- snapshot ZFS envoyé vers sauvegrenoble2 (marche normale)

# Plusieurs familles de bases de données

Un site propose un classement des moteurs de bases de données :
`https://db-engines.com/en/ranking`
La PLM utilise 3 familles :

- les bases relationnelles
- les tables clé-valeur
- les bases d'objects (documents)

# Les bases de données relationnelles

MySQL, MariaDB, PostgreSQL ont facilité l'usage de bases relationnelle :

- robustesse, performance, transaction ACID
- format de stockage MyISAM [5] (.frm, .MYI, .MYD), InnoDB [6] (ibdata1) ou PGDATA [7]
- langage de requête normalisé Simple Query Language
- utilisées même en absence de relations entre les tables

Applications utilisant des bases SQL :

- pour gérer leurs configurations :
  - seafile, sympalist, onlyoffice, jabber, openmeeting
  - plmconfig puppet : faits, catalogues, exports, rapports
  - jetons de licences : gestion filtres iptables selon souscriptions
- pour aussi stocker des données utilisateurs :
  - PLMlab, hébergement web
  - indico : contenu des événements
  - webmail horde : agenda personnel

5. https://mariadb.com/kb/en/library/myisam-overview/
6. https://dev.mysql.com/doc/internals/en/innodb.html
7. https://www.postgresql.org/docs/10/storage-file-layout.html

# Les bases SQL pour sites web

Exemple avec l'option base SQL de l'hébergement web :

- Création d'une ou plusieurs bases sur demande support@math.cnrs.fr
- VM db3 spécifique pour serveur MariaDB
- Requêtes SQL depuis les VM web via réseau interne

Sécurisation :

- Extraction quotidienne des bases de chaque site dans htdocs/site/.dbdump/
- Extraction quotidienne complète dans htdocs/.dbdump/
- État cohérent dans BackupPC du filer
- Restauration d'une base depuis dbdump sauvegardé

mathrice

# PLMlab

La base SQL est une des formes de stockage des données

- métadata de l'application
- signalement de problème ou demande de fonctionnalité (issue)
- proposition de modification de code (merge request)
- suivi des travaux d'intégration continue

Difficulté de conserver un état cohérent entre toutes les formes de stockage

Des pistes pour fiabiliser le service :

`https://docs.gitlab.com/omnibus/roles/README.html`

`https://docs.gitlab.com/ee/administration/high_availability/README.html`

mathrice

# Les tables clé-valeur

Propriétés des tables clé-valeur :

- Optimiser l'accès à des tables indexées
- Performance sur de gros volumes de données
- Base persistante sur disque couplée à un cache en mémoire
- Réplication et/ou répartition entre serveurs
- Données de session ; Cache information temporaire ; Files de travaux

- redis [8] : PLMlab, PLMlatex, PLMofficedoc (OnlyOffice)
- couchDB [9] et memcached : `https://portail.math.cnrs.fr`

mathrice

8. `https://redis.io/`
9. `http://couchdb.apache.org/`

# Les bases d'objects : Not only SQL

Propriétés du NoSQL :

- Stockage d'objets dont la structure peut évoluer
- Indexation des objets
- Réplication et/ou répartition possible entre serveurs

PLMlatex utilise MongoDB [10]

- Stocker les documents LaTeX
- Sécurisation avec réplication MongoDB sur une seconde VM

indico.mathrice.fr (Indico v1) utilise Zope Object DataBase ZODB [11]

- Outil développé par le CERN
- Abandon de cette base de données pour PosgreSQL à partir de la v2

10. https://www.mongodb.com/
11. http://www.zodb.org/

# Questions

Quelques questions avant de passer aux ateliers PLM ?

# Les ateliers PLM

A. Espace de stockage données froides sur la PLM (disque)
B. PLMBox : stockage de données chaudes avec seafile, travail collaboratif, édition OnlyOffice en ligne
C. PLMlab : à quoi ça sert ? Dépôt git, pages, intégration continue.
D. Hébergement web sur la PLM
E. PLMlatex : édition de latex collaboratif en ligne
F. Indico : gestion d'une conférence (site web, programme, inscriptions), alimentation de l'agenda des maths
G. Limesurvey : enquêtes, sondages et questionnaires

mathrice

# Les ateliers PLM

A. Disque
B. PLMBox
C. PLMlab
D. Web
E. PLMlatex
F. Indico
G. LimeSurvey

Programme des 3 ateliers de 40min :

| Salle | 14 : 45 | 15 : 25 | 16 : 30 |
|---|---|---|---|
| Auditorium | B | B | F |
| Bibliothèque | C | A | A |
| Chapelle | D | E | G |

Pause : 16 : 05 – 16 : 30