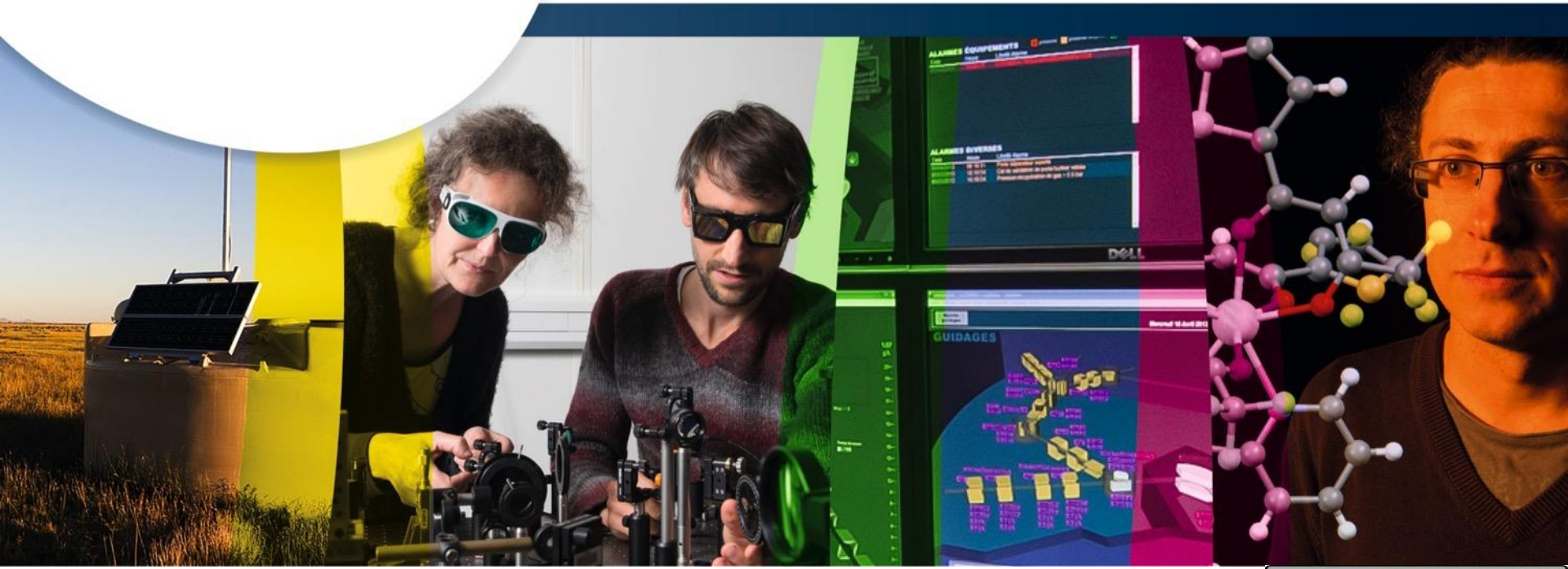
**Pour plus d'informations concernant ce document :**

- *http://www.ods.cnrs.fr/contacts.html*
- *https://aide.core-cloud.net/mycore/Pages/Accueil.aspx*

www.cnrs.fr



# Plan

- Description du contexte My CoRe
- Description de la plateforme technique
- Concepts d'architectures Scality
- RETEX
- Points de vigilance
- Conclusion
- Fond documentaire

www.cnrs.fr

# Contexte

# Contexte

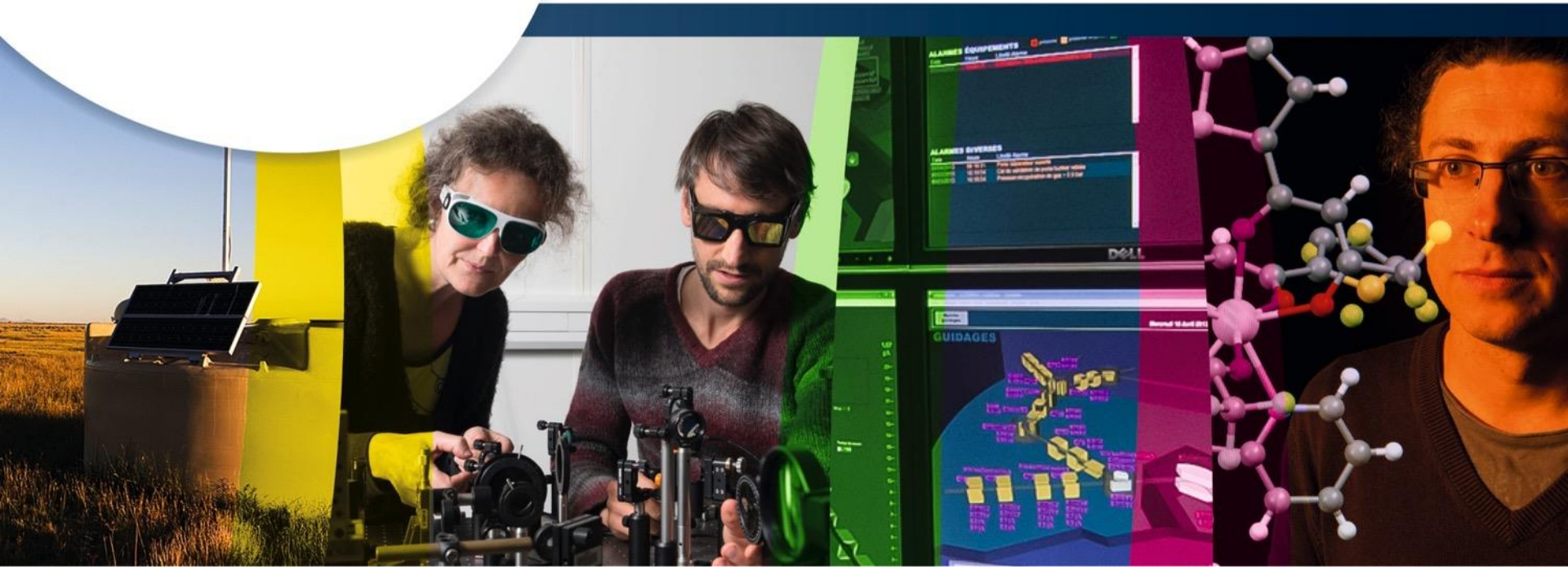
www.cnrs.fr

- Objet de cette nouvelle Offre De Service (ODS)
  - Répondre à un besoin existant et identifié d'une **ODS sécurisée de partage et de synchronisation de fichiers de travail**
  - Proposer une **ODS pour copier en ligne ses fichiers de travail locaux** (plus communément appelé « Mes documents »)
  - ODS à destination des **agents des unités CNRS**, avec **20 Go** d'espace utile gratuit par utilisateur

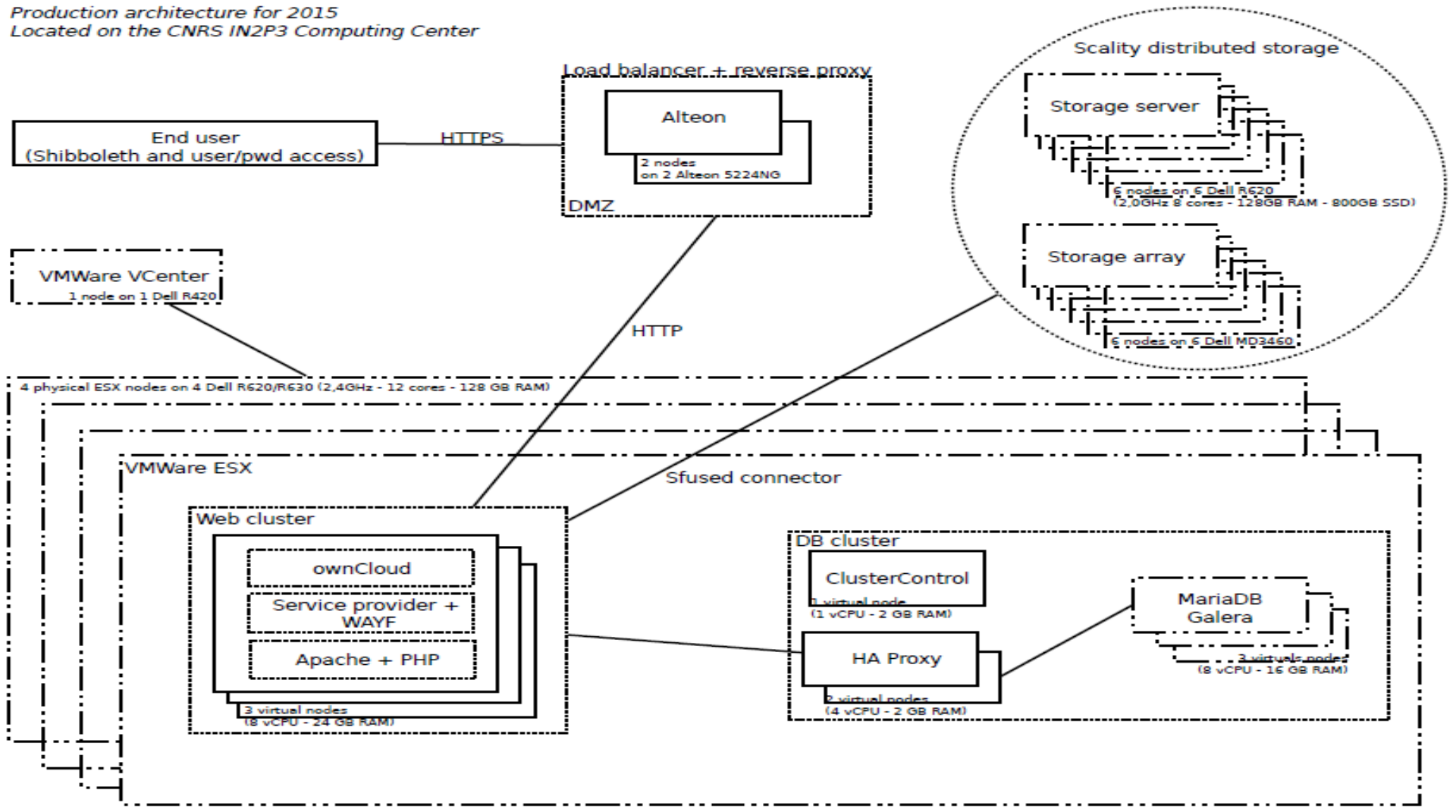
- Intérêt de cette nouvelle Offre De Service
  - ODS « **anti-Dropbox** » : offre sur un cloud souverain CNRS permettant de réduire les risques d'exploitation d'informations de recherche inhérentes à des solutions « Dropbox-like »
  - ODS qui facilitera le **nomadisme**

www.cnrs.fr

# Architecture Plateforme My CoRe



# Architecture



# Architecture

- Choix de la solution de backend de stockage pour My CoRe
  - une forte résilience, car pas de « sauvegarde » nécessaire
  - Très fortement scalable
  - qui soit unique pour toute la volumétrie
  - un accès « file system" stable
  - un coût au Go faible
  - la possibilité d'être multi sites
- Choix initiale de Scality en mai 2014

# Architecture

- Eléments Clés du Ring
  - Stockage logiciel d'objets
  - Capacité illimitée (scale-out)
  - Stockage mutualise
  - La notion d'ARC
- Les Points forts
  - Compatible tout serveur x86
  - Ratio brut/utile d'environ 1.6
  - Très hautement disponible
  - Pas de RAID matériel

# Architecture

1 Disques SSD 800 Go /serveur

R630 Dell

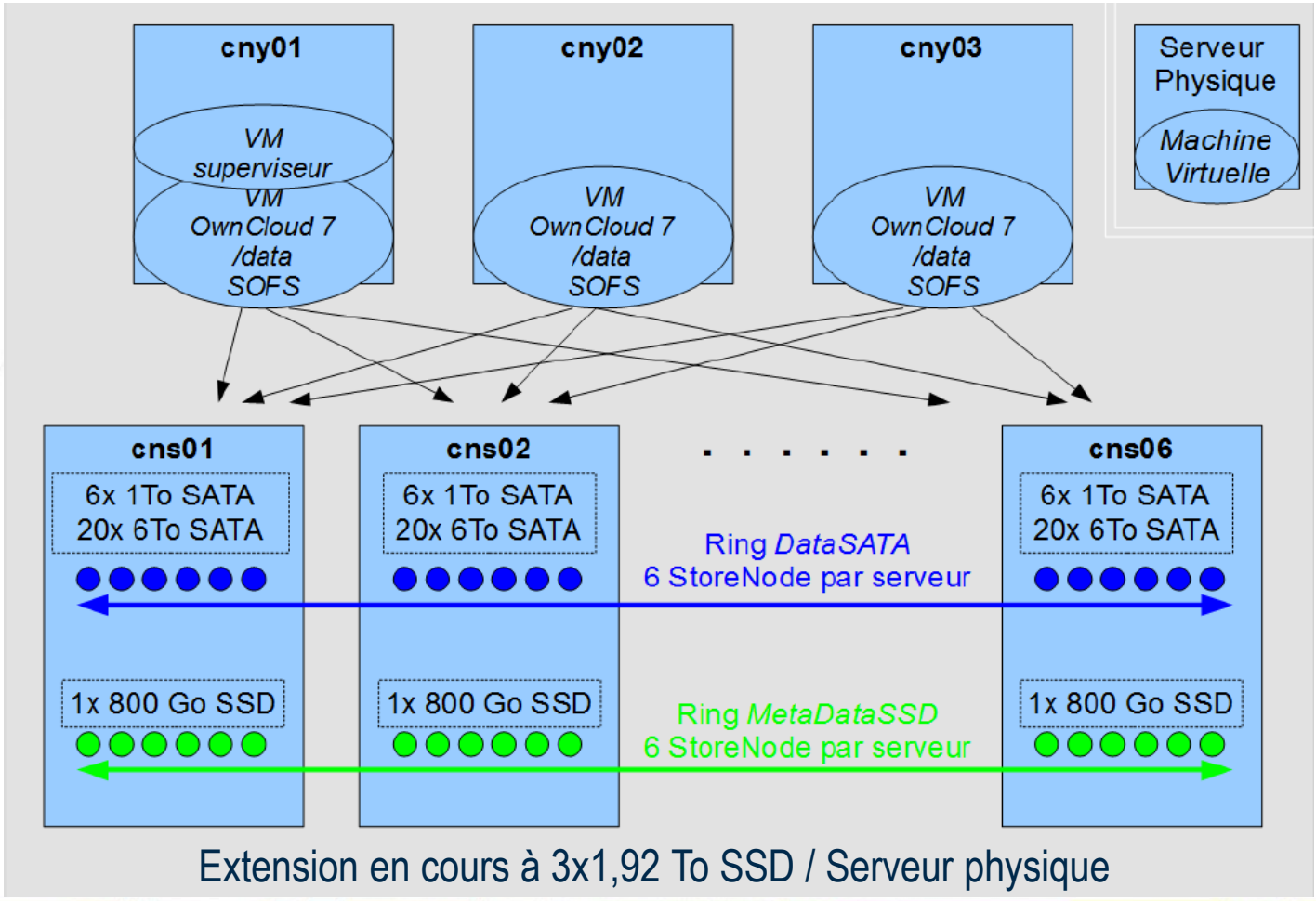
Châssis Dell MD3460

SCALITY

20 Disques de 6 To
6 Disques de 1 To
SATA 7,2 K

Nb: superviseur sur VM

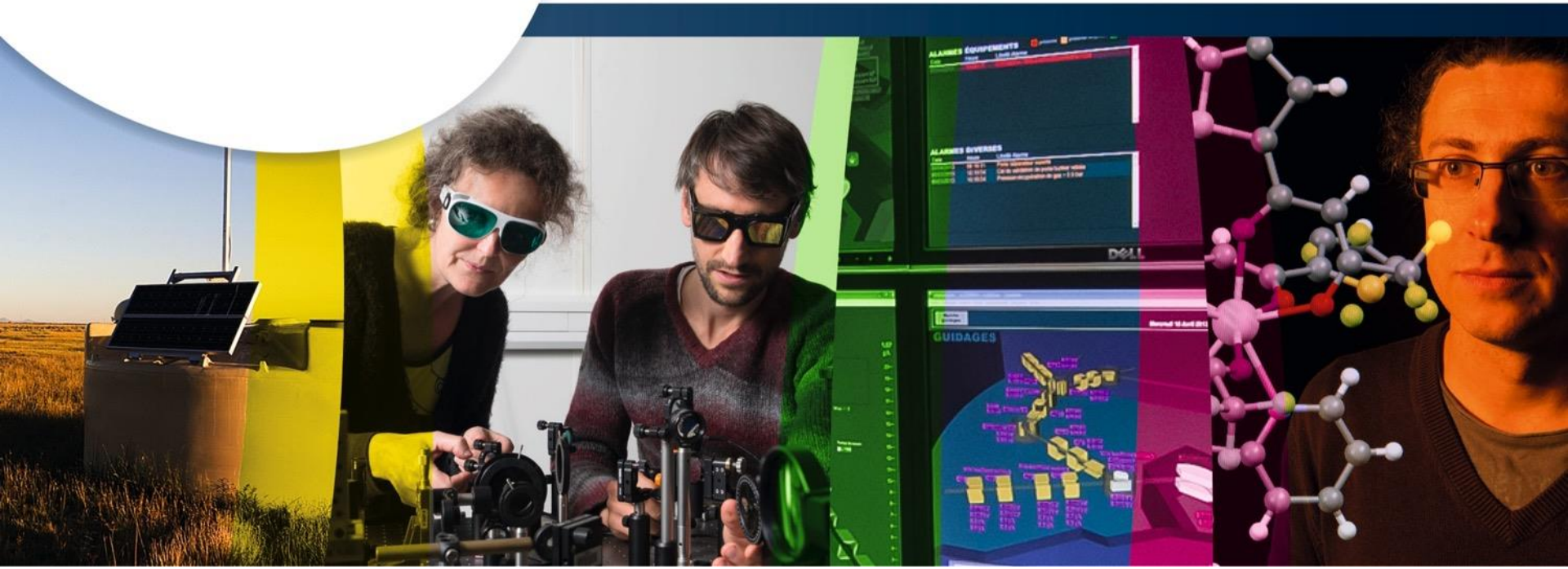
# Architecture




Extension en cours à 3x1,92 To SSD / Serveur physique

www.cnrs.fr

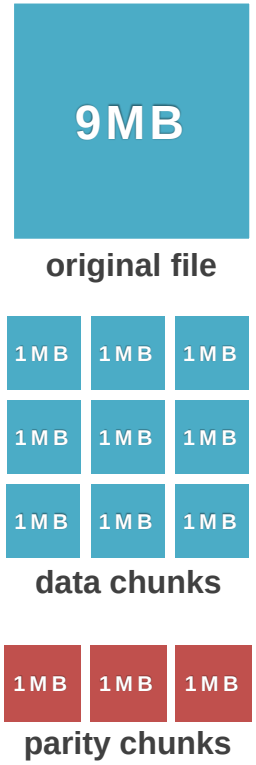
# Concepts d'architectures Scality



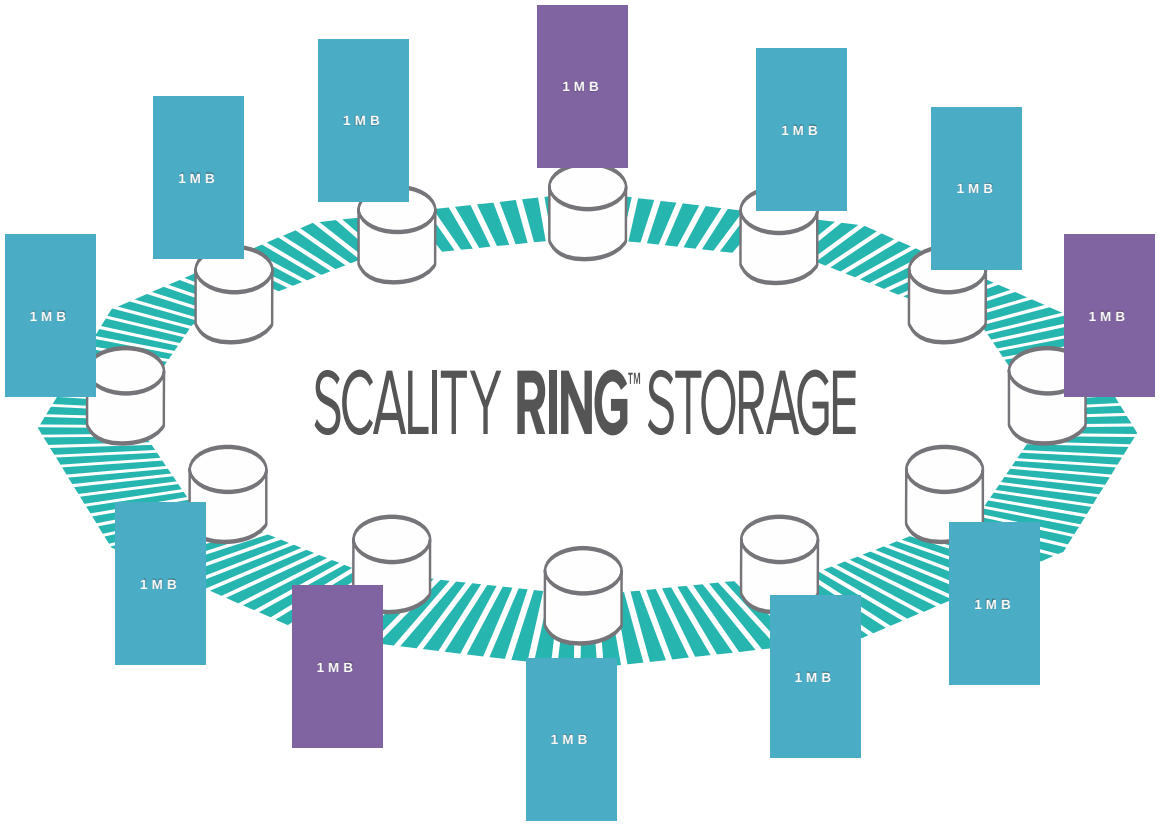
# Concepts

www.cnrs.fr

SCALITY

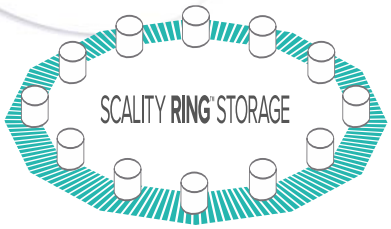



*Example: ARC(9,3)*
*Provides three-disk failure protection with ~33% overhead*

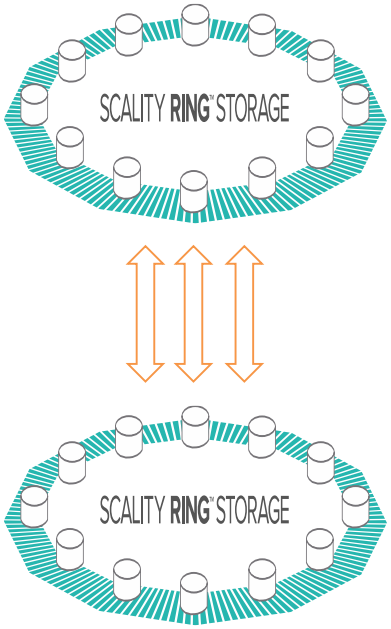


# Concepts – Topologie du Ring

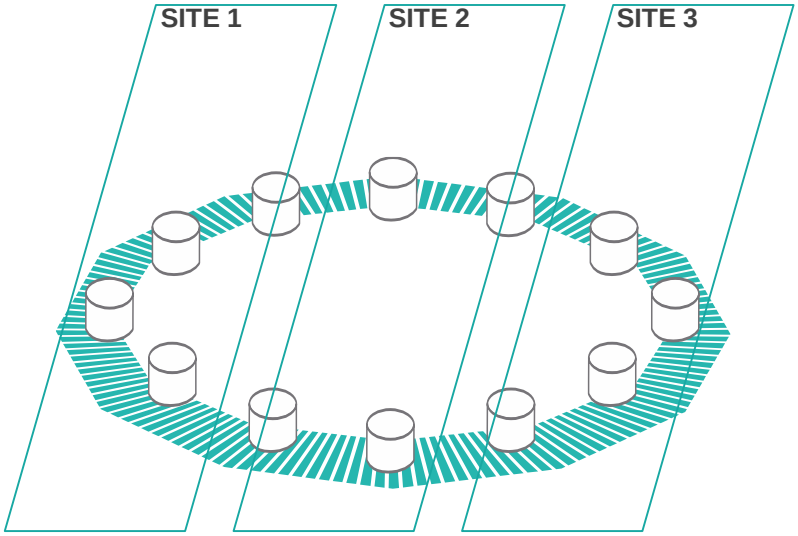
www.cnrs.fr




Single site durability




Multi-RING, multi-site durability, asynchronous data

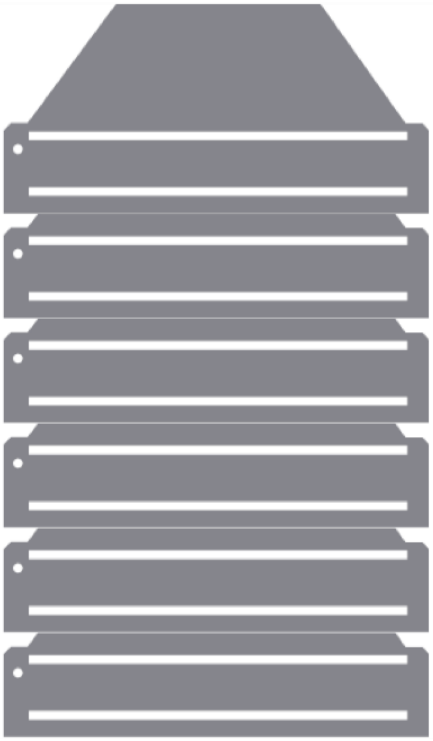



Multi-site durability, synchronous data

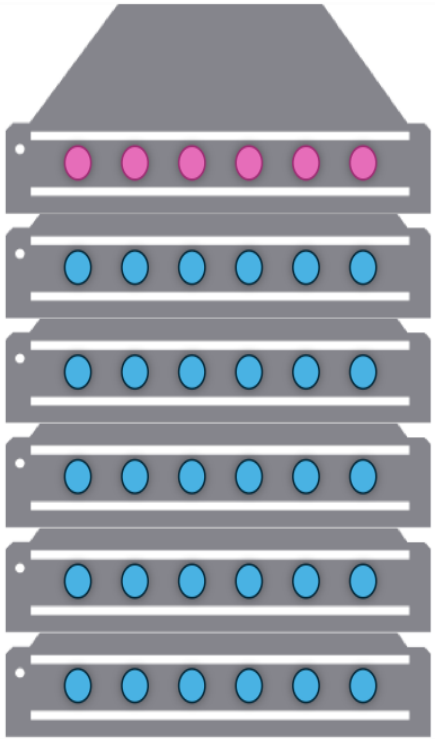
# Configuration

www.cnrs.fr

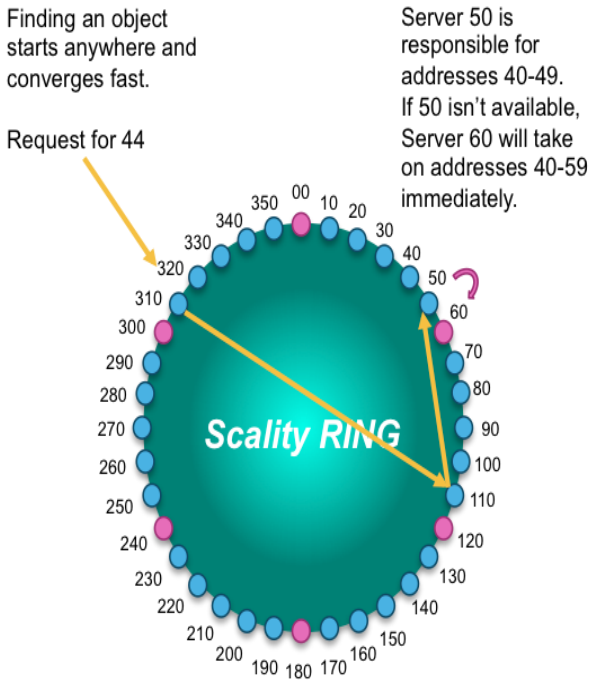
SCALITY



A minimum of six servers



Six *Storage Nodes* per server




**36 *Storage Nodes* logically organized into a binary addressing space**

Configuration Minimale : 6 Serveurs physiques 6 Storage nodes (process) / serveur

# Différents « Connectors »

www.cnrs.fr

SCALITY

**FILE**
NFS v3
SMB 2.0/3.0
Linux FS (Sfused)

**OBJECT**
REST (Sproxyd)
CDMI REST
S3 (RS2)

**OPENSTACK**
CINDER
GLANCE
SWIFT

# Performances observées des « Connectors »

Performances moyennes constatées

www.cnrs.fr

SCALITY

| FUSE | NFS | SMB | REST |
|---|---|---|---|
| READ: 700MB/s | READ: 250MB/s – 350MB/s | READ: 300MB/s – 700MB/s | READ: 2,500MB/s |
| WRITE: 450MB/s – 650MB/s | WRITE: 250MB/s – 350MB/s | WRITE: 250MB/s – 350MB/s | WRITE: 2,500MB/s |

Linear Performance Scaling

# Concepts – Les différents objets et process

- Le Ring-Data (disque à plateau)
  - Contient les données brutes des applications ou des utilisateurs
  - Organisé dans des containeurs
    - pas de limitation d'inodes
    - pas de limitation de block-size
    - pas de fragmentation, les données sont intelligemment réarrangés au fil du temps
  - La répartition des blocs est fonction de l'architecture de l'ARC
  - Utilise des disques à faible coûts

# Concepts – Les différents objets et process

- Le Ring MétaData
  - Contient les données utiles des applications ou des utilisateurs pour les fichiers ayant une taille inférieur à la taille du bloc défini lors de la mise en œuvre du Ring Datas
  - Pas de répartition de la données sur plusieurs disques
  - Contient les entrées et indexes de répertoires
  - Contient les liens symboliques
  - la liste des "morceaux" de fichiers bruts existants

# Concepts – Les différents objets et process

- Les bizobjs
  - Stockage sur les SSD en dehors du Ring METADATA
  - Contient les users metadata des données brutes pour un disque donné
  - Un fichier bizobj par disque et par ring géré sur ce disque
  - Données contenues
    - Filename
    - Crc (Cyclic Redundancy Check)
    - Permissions
    - Emplacement dans le conteneurs
    - Taille
    - Version
    - Etc…

# Concepts – Les différents objets et process

- Usage de la RAM ou stockage des Clés (Indexes)
  - La localisation de chaque bloc d'un ficher de data utile est stocké en RAM dans un fichiers dit de « Clé » celle-ci représente 47octets / fichier
  - Ces fichiers seront supprimés par les mécanismes de purges une fois le ficher de donnée source supprimé
- Les purges
  - Ces mécanismes positionnent un « flag » sur les différents items à supprimer sur les Rings (Méta ou Data)
- Eléments clés sur les purges
  - Clé (en RAM)
  - Méta Datas (dans le fichier bizobj)
  - Blocs de données utiles
  - Les mécanismes de purges sont réglables

# Concepts – Les différents objets et process

- Les réallocations
  - Les mécanismes de réallocations sont des processs d'arrière plan qui viennent compléter les process purges
  - Ce sont eux qui sont responsable de la libération de l'espace physique sur les disques
  - Il y a un process par disque physique, ceux-ci sont lancées par les process biziod
  - Les mécanismes de réallocation sont réglables
- Warning
  - Tout ces paramètres ont un impact sur les performances
  - Il convient de rester attentif après un changement afin de valider qu'il n'y a pas d'effet de bord

www.cnrs.fr

# RETEX

# RETEX

- Implémentation
  - Étapes du choix de Scality, avec délai et manpower faibles pour faire ce choix
  - Tour d'horizon sur "papier" de diverses solutions (HDFS, CephFS, GlusterFe S, iRODS)
  - Définition d'une short list
    - Dell Compellent : le plus économique mais volumétrie limitée à 2 Po et résilience limitée
    - EMC Isilon : le plug and play mais matériel spécifique
    - Scality : le plus souple
      - Grille de choix : cf. https://github.com/CNRS-DSI-Dev/mycore_press/blob/master/CNRS-INSERM-20160502.pdf, slide 11
  - Étude de Scality en détail : maquettage et récupérations de divers retours d'expérience sur la solution

# RETEX

- Implémentation
  - Étapes du choix de Scality, avec délai et manpower faibles pour faire ce choix
  - Tour d'horizon sur "papier" de diverses solutions (HDFS, CephFS, GlusterFe S, iRODS)
  - Définition d'une short list
    - Dell Compellent : le plus économique mais volumétrie limitée à 2 Po et résilience limitée
    - EMC Isilon : le plug and play mais matériel spécifique
    - Scality : le plus souple
      - Grille de choix : cf. https://github.com/CNRS-DSI-Dev/mycore_press/blob/master/CNRS-INSERM-20160502.pdf, slide 11
  - Étude de Scality en détail : maquettage et récupérations de divers retours d'expérience sur la solution

# RETEX

- Exploitation
  - Le récurent
    - Assez peu de charge en dehors des opérations de maintenance, le système tourne « tout seul » seulement rester attentif à la supervision
  - Les montées de versions
    - 2 montées de version déjà réalisées sans arrêt de service
    - Nécessite de faire monter les différents composants au fur et à mesure dans un ordre précis et dans le respect de la matrice de compatibilité
      - Superviseur
      - Connecteurs
      - Storage Node

# RETEX

www.cnrs.fr

- Nécessite de bien planifier les opérations en amont
- Support de l ’éditeur réactif et compétent
- Evolutions possibles
  - Passage du mode fichier (SFUSED) au mode objet depuis les VM clientes ownCloud si les améliorations du connecteurs SFUSED annoncées début 2016 ne sont pas suffisantes
  - Chirage du systeme de chiers via Scality

www.cnrs.fr

# Points de vigilance

# Points de vigilance

- La formation
  - Nécessite des connaissance de base en solution de stockage distribué
  - Nécessite des connaissances spécifiques pour appréhender correctement les différents éléments de l'architecture logicielle
  - Nécessite de connaitre le fonctionnement du superviseur
  - Nécessite de connaitre les commandes spécifiques ainsi que la façon de passer les options
- Dimensionnement initiale de l'ARC du Ring
  - Choix du nombre serveurs physiques
  - Définition de l'ARC
  - Choix du nombre de SSD

# Points de vigilance

- Les I/O sur les SSD
  - La lecture / écriture des fichiers sur le Ring Data passe par une lecture / écriture des fichiers bizobjs stockés sur les SSD
  - La lecture / écriture des fichiers stockés sur le Ring Méta-Data
- Le capacity planning du Ring Méta-Data
  - L'occupation de l'espace disque sur les SSD n'est pas linéaire et proportionnelle à la montée en charge du Ring Data
- Bien connaitre son application:
  - Les différents accès sur le Ring via les différents mécanismes de l'application
    - Listing de fichiers
    - Forte profondeur de l'arborescence de fichiers

www.cnrs.fr

# Conclusion

# Conclusion

www.cnrs.fr

- Les points positifs
  - Une technologie intéressante ouvrant des perspectives d'avenir (cf road map)
  - Un produit robuste
    - 2 incidents majeurs sans perte de données
  - Une technologie à maitriser
    - Attention aux effets de bords!!!
  - Suffisamment fiable et robuste pour conserver la technologie et penser à la pérennisation de la plateforme
  - Réflexions autour de nouveaux usages :
    - Stockage via une interface S3 pour une nouvelle offre de service
    - Stockage via Api Rest

# Conclusion

www.cnrs.fr

- Les points moins positifs
  - Limitations du connecteur SFUSED (mauvaise performance des accès au de la de 2 millions de fichiers lors de certaines opérations, type listing, et corruption de fichiers lors d'accès concurrents en écriture sur un même objet)

www.cnrs.fr

# Fond documentaire

# Fond documentaire

www.cnrs.fr

- Présentation préalables de David Rousse:
  - mycore_press/**CNRS-JoSy-20140519.pdf**
  - mycore_press/**CNRS-INSERM-20160502.pdf**
- Fond documentaire Scality:
  - Documents avant vente
  - Documents de formation