RozoFS

Erasure Code Based Scale-out NAS

La solution unique

Un stockage innovant pour le Cloud privé

Présentation JoSy 2014 « Cloud privé dans l'enseignement et la recherche »

Pierre Evenou & Christophe de La Guerrande

① Présentation

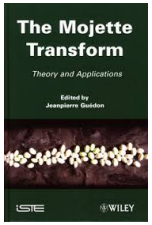
② Stockage et FEC : RozoFS

③ Cas client : Projet CARMIN

Fizians et son produit RozoFS

PRÉSENTATION

Chronologie



1995-Prof J.P Guédon
Mojette théorie

1997-Prof N. Normand
Mojette Algorithme

2003-P. Evenou
Mojette Stockage

Création de Fizians SAS

Développement du Produit

Mise sur le Marché De RozoFS

6 références en production

Commercialisation De RozoFS

Evangélisation et prospection

Développement Fonctionalités Produit

Etre un acteur majeur

Developpement à l'international

2010

2013

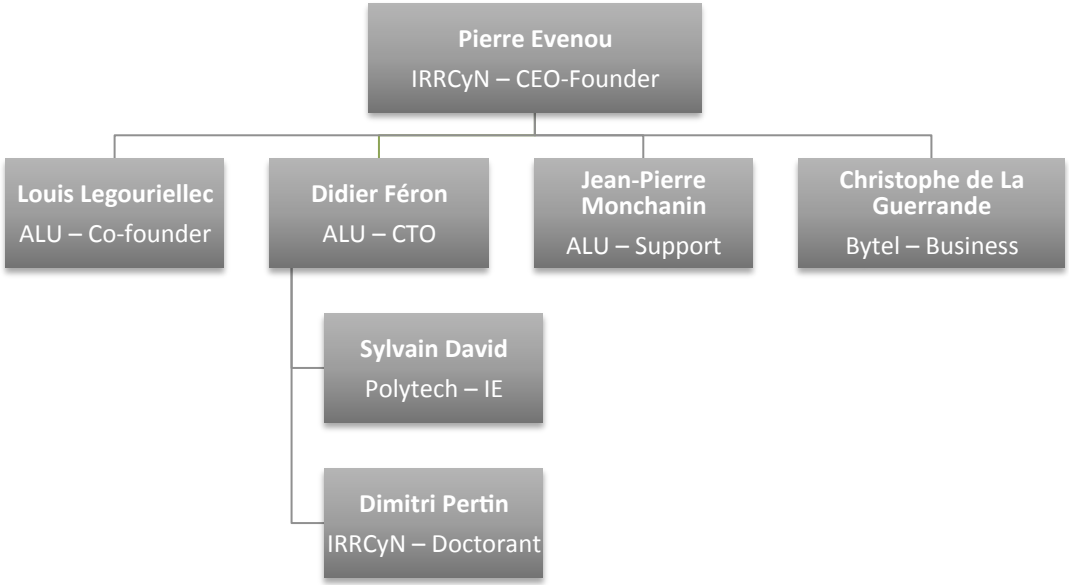
2014

2015

Overview

Fizians SAS :

- Réunion de la recherche et de l'industrie Télécom.
- 7 personnes.
- Clients ; CNED, Université de Nantes, CUN, Puy du Fou, Polytechnique, IHP, IHES, CMLS.
- Cibles ; Multimédia, Cloud Computing, HPC, Big Data.
- Devenir un acteur majeur dans le stockage Scale-Out NAS.



Problématique

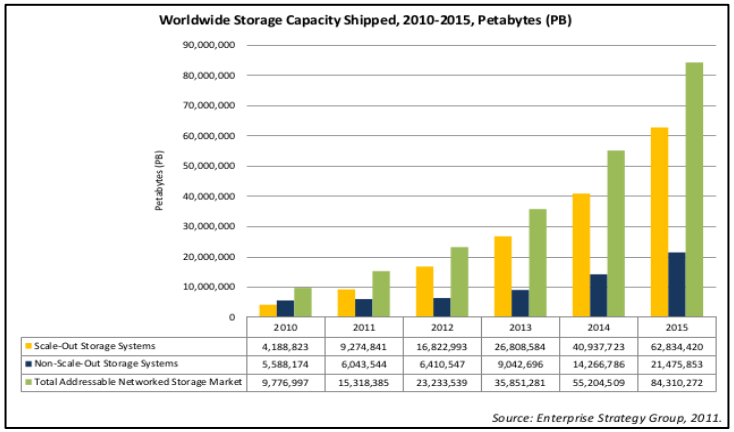
Une croissance exponentielle des données, (100% tous les 18 mois).

- Engendrent des difficultés pour stocker et organiser.
- Une évolutivité couteuse et complexe.
- L'indisponibilité des données est souvent synonyme de perte de CA.

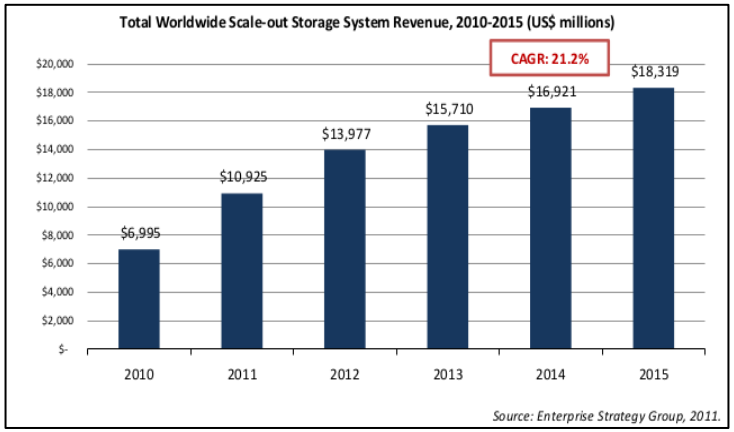
On assiste à un virage technologique, les solutions NAS, SAN, DAS et RAID ne répondant plus aux enjeux.

Marché

- Marché Mondial : **68 000 Po pour 18 Milliards de $ en 2015 pour le Scale-Out.**



	2010	2011	2012	2013	2014	2015
Scale-Out Storage Systems	4,188,823	9,274,841	16,822,993	26,808,584	40,937,723	62,834,420
Non-Scale-Out Storage Systems	5,588,174	6,043,544	6,410,547	9,042,696	14,266,786	21,475,853
Total Addressable Networked Storage Market	9,776,997	15,318,385	23,233,539	35,851,281	55,204,509	84,310,272



- Concurrence
  - Logiciel : GlusterFS, Caringo, Ceph, Scality.
  - Matériel : EMC, Net App…

Deux réponses technologiques émergeantes :

1) Le code à effacement (fiabilité et réduction des coûts),

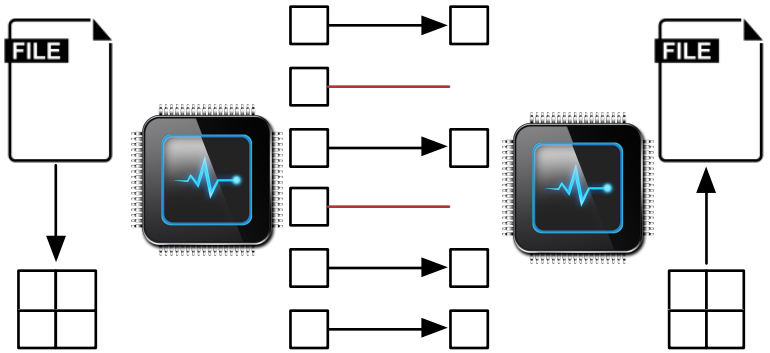
2) Scale-Out NAS (File System évolutif et simple),

Transformée Mojette

CODE À EFFACEMENT ET STOCKAGE

Erasure Code

- Haute protection
- $(n / k) - 1$ overhead



k data blocks

k data blocks

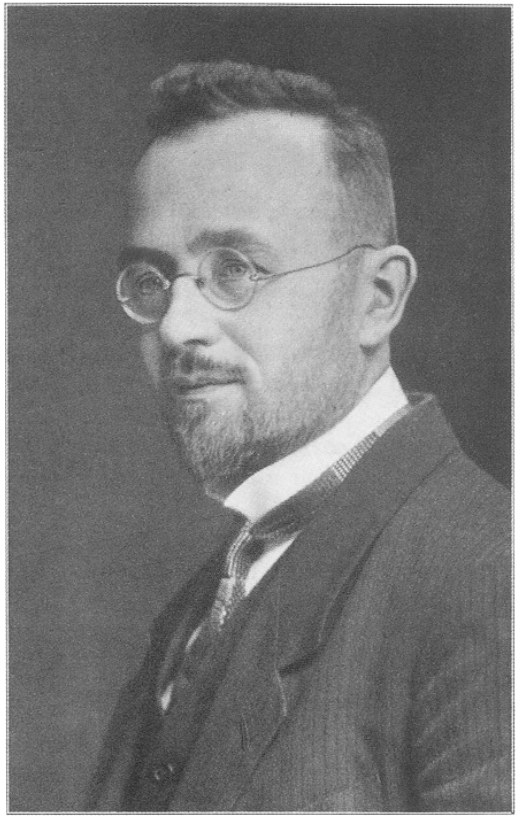
Code (6, 4) : 50% d'overhead

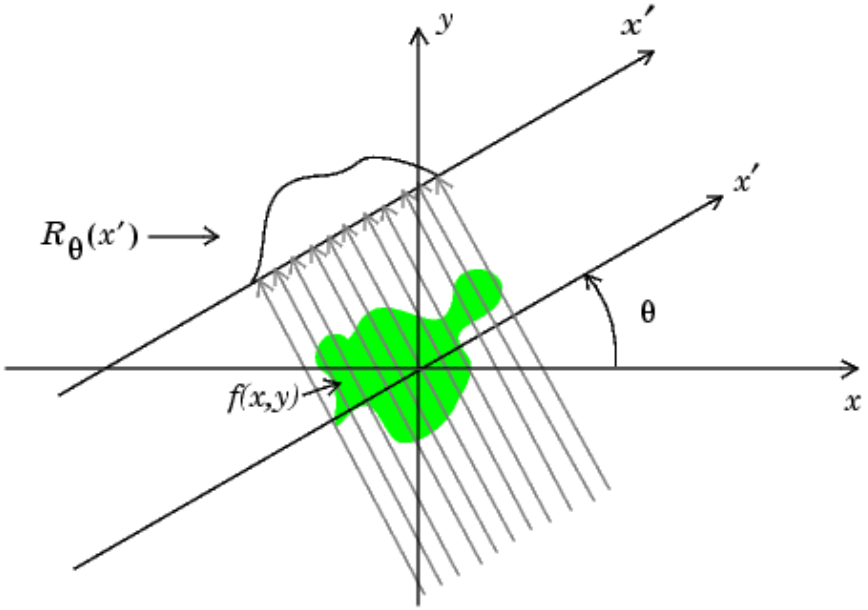
Adapté aux architectures distribuées
Efficace pour réduire les coûts

Trop lent et consommateur de CPU
Limité aux données froides à accès séquentiels

Mojette

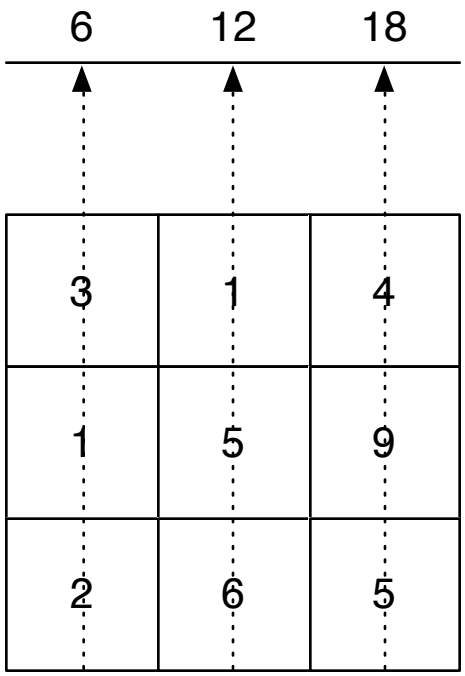
Une transformée de Radon





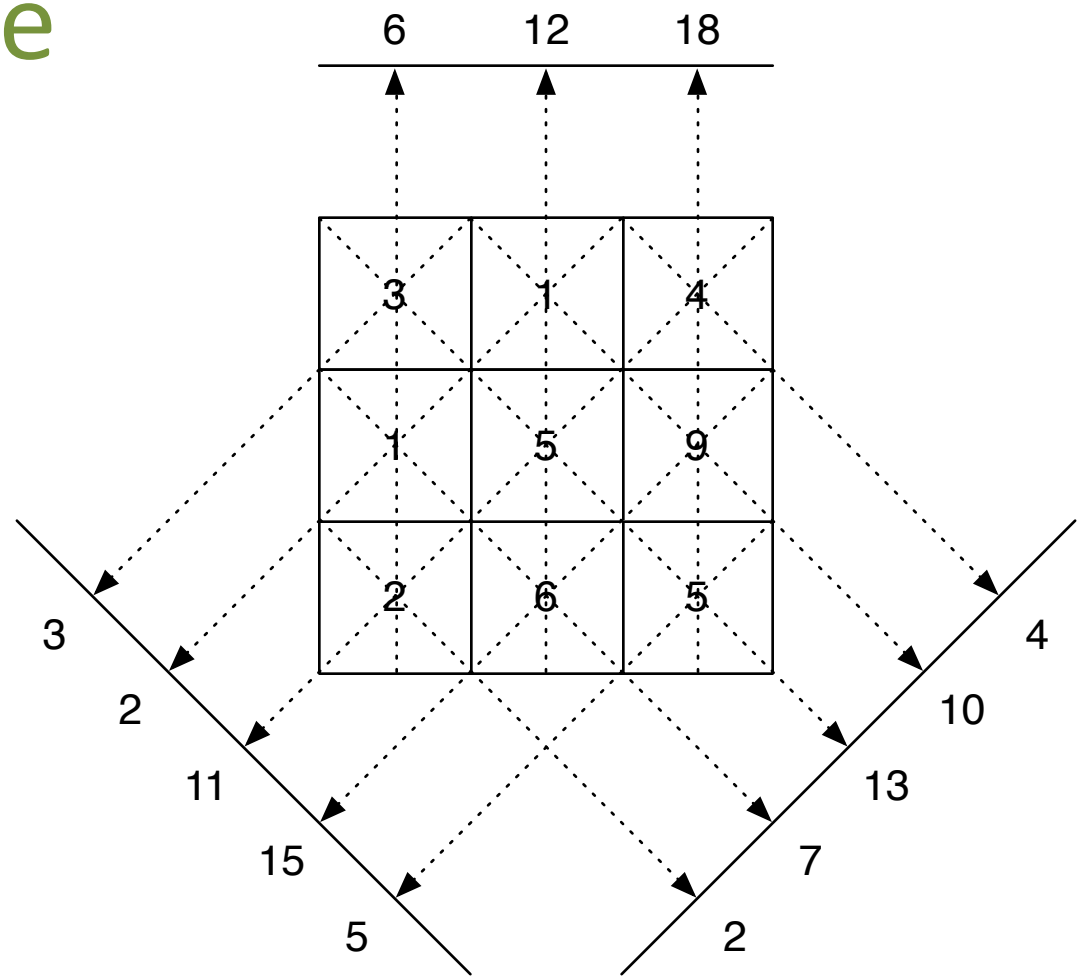
Une transformée de Radon

discrète



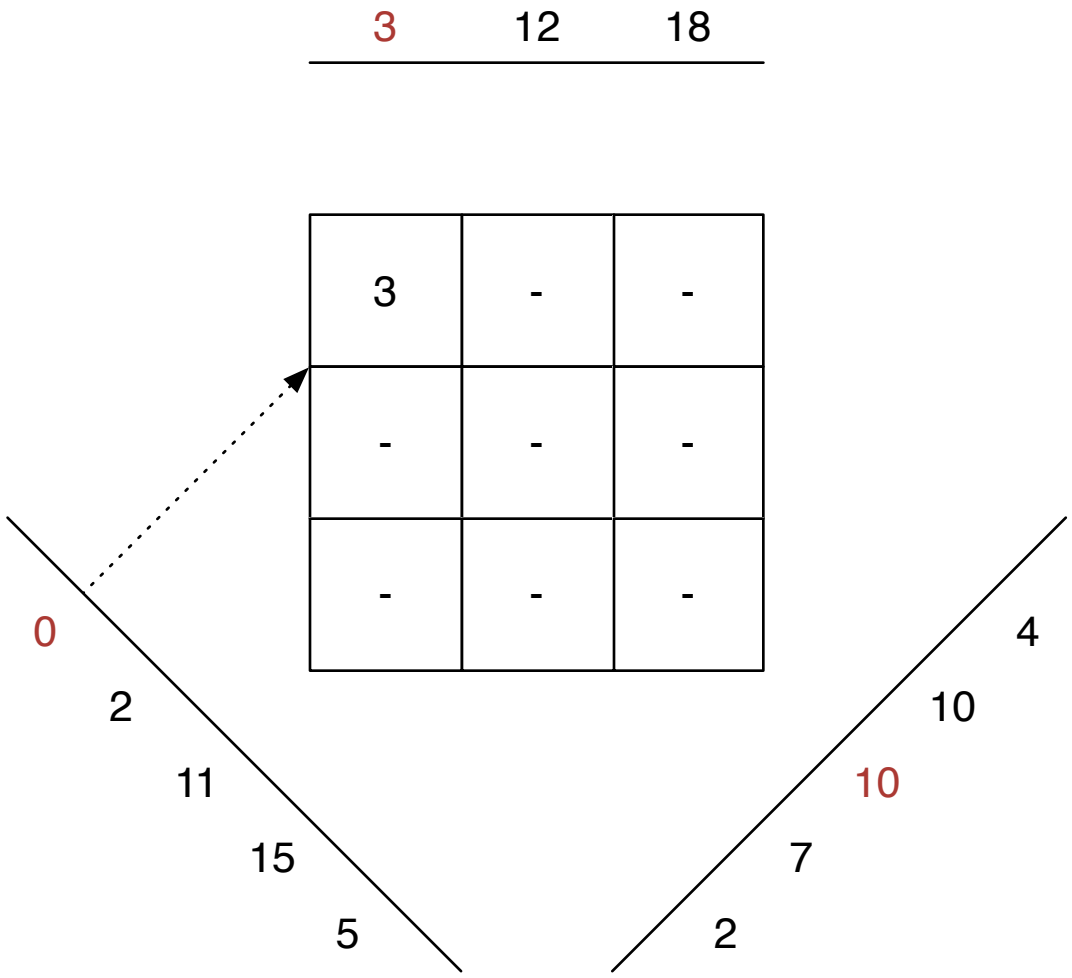
Une transformée de Radon

discrète



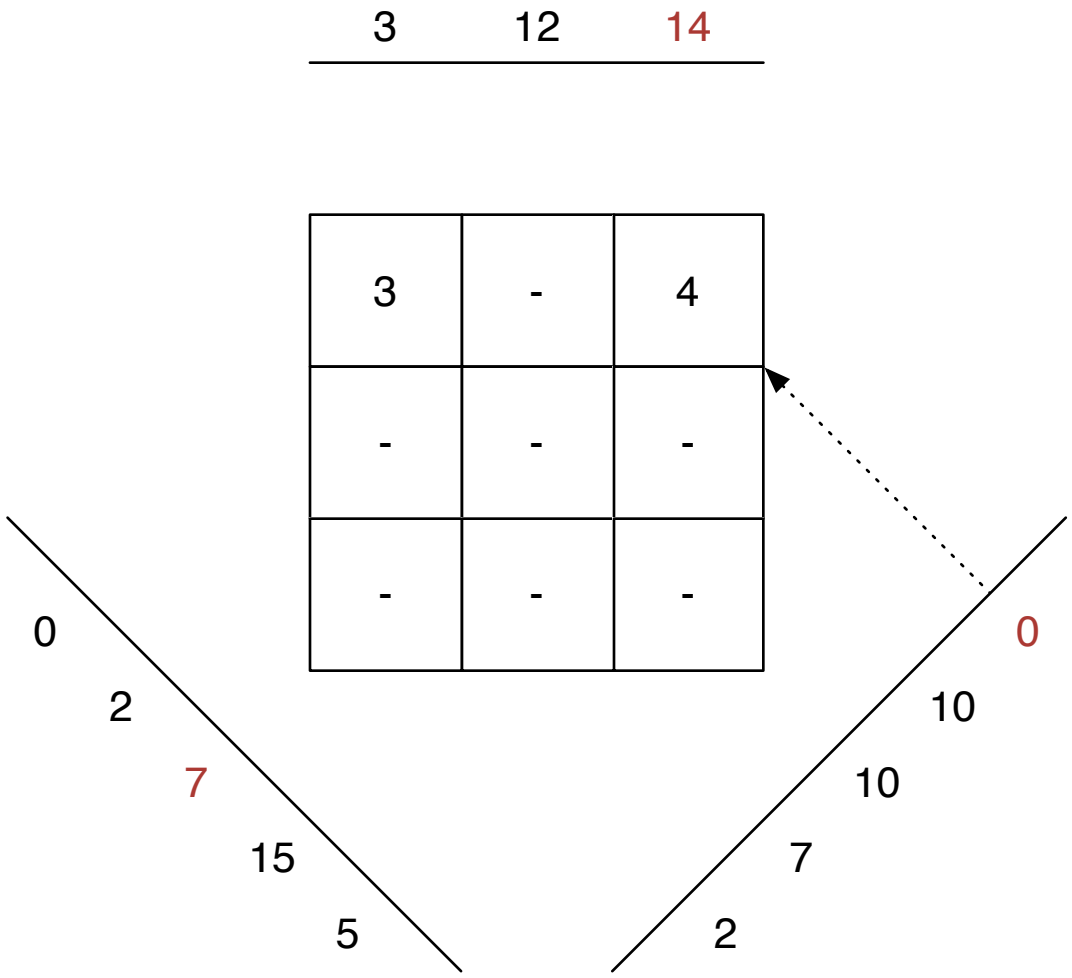
Une transformée de Radon discrète

exacte



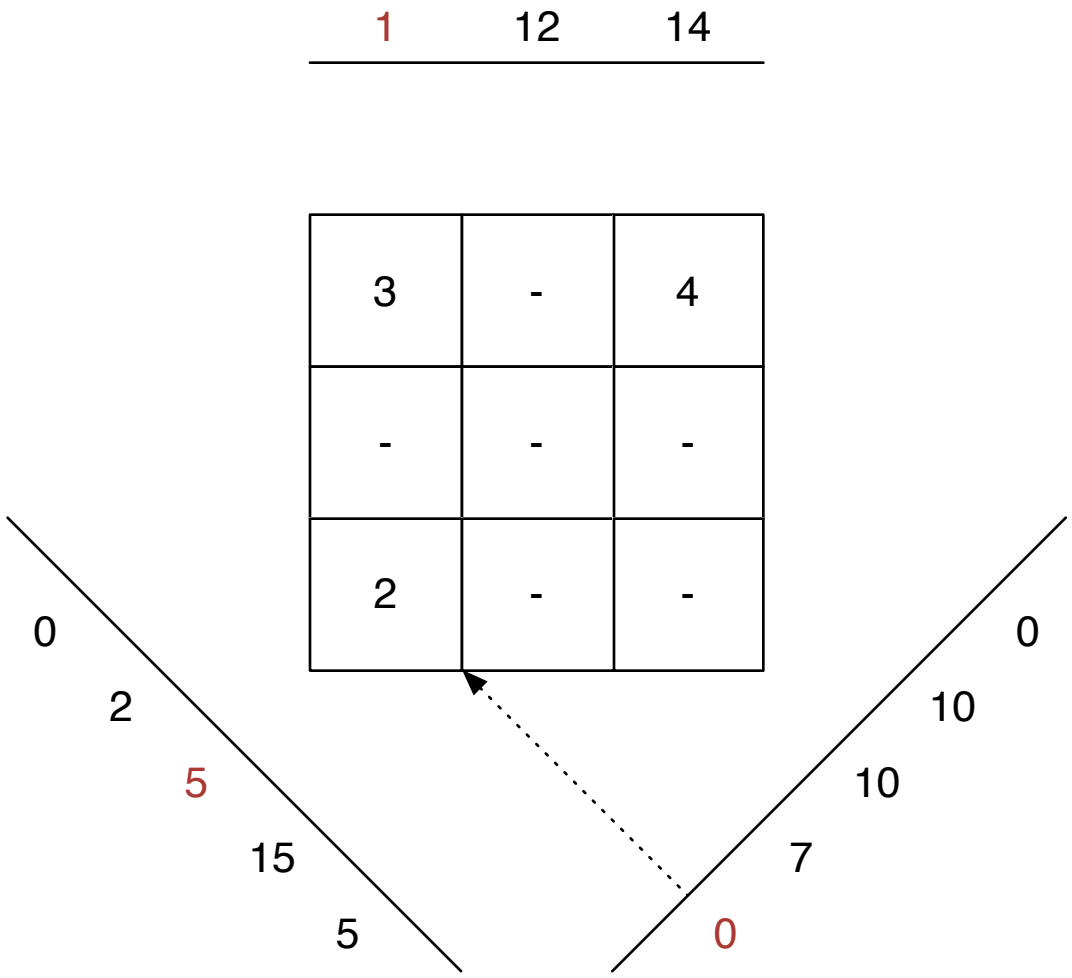
Une transformée de Radon discrète

exacte



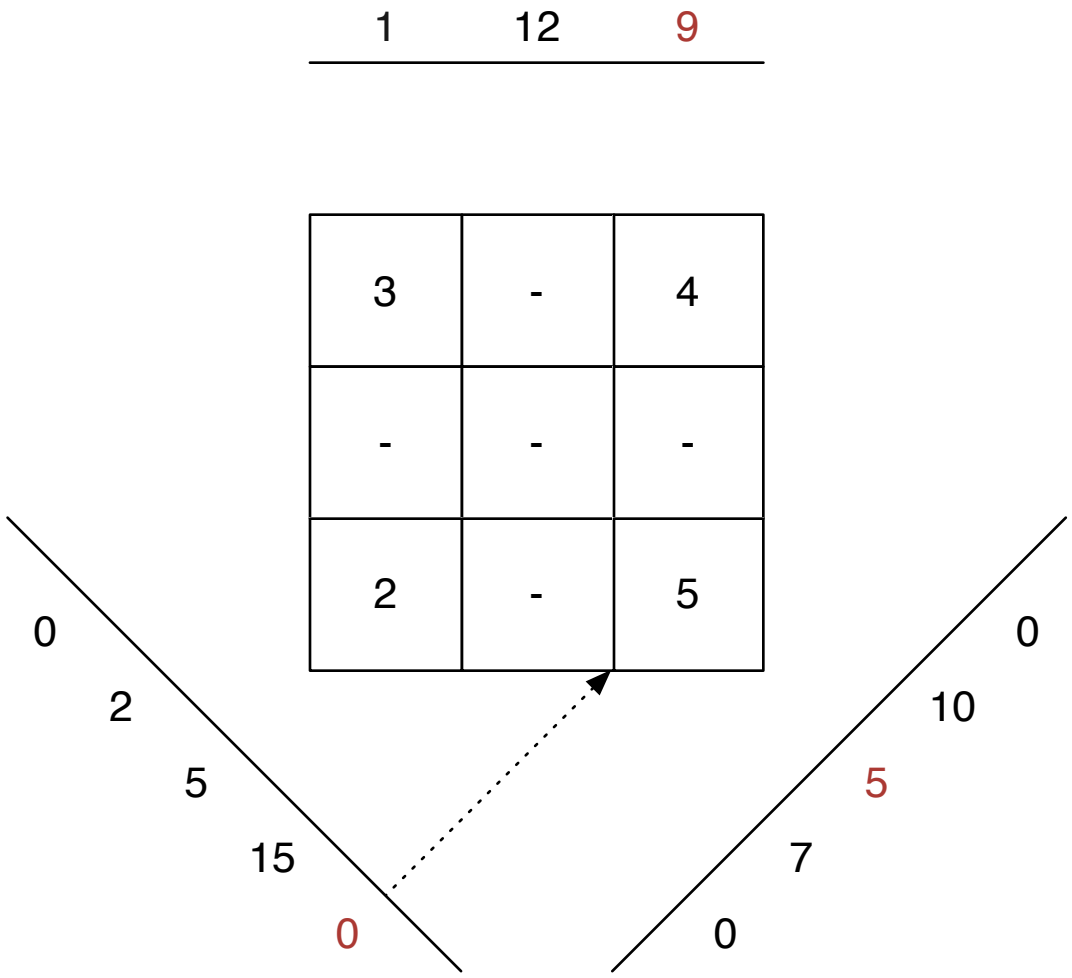
Une transformée de Radon discrète

exacte



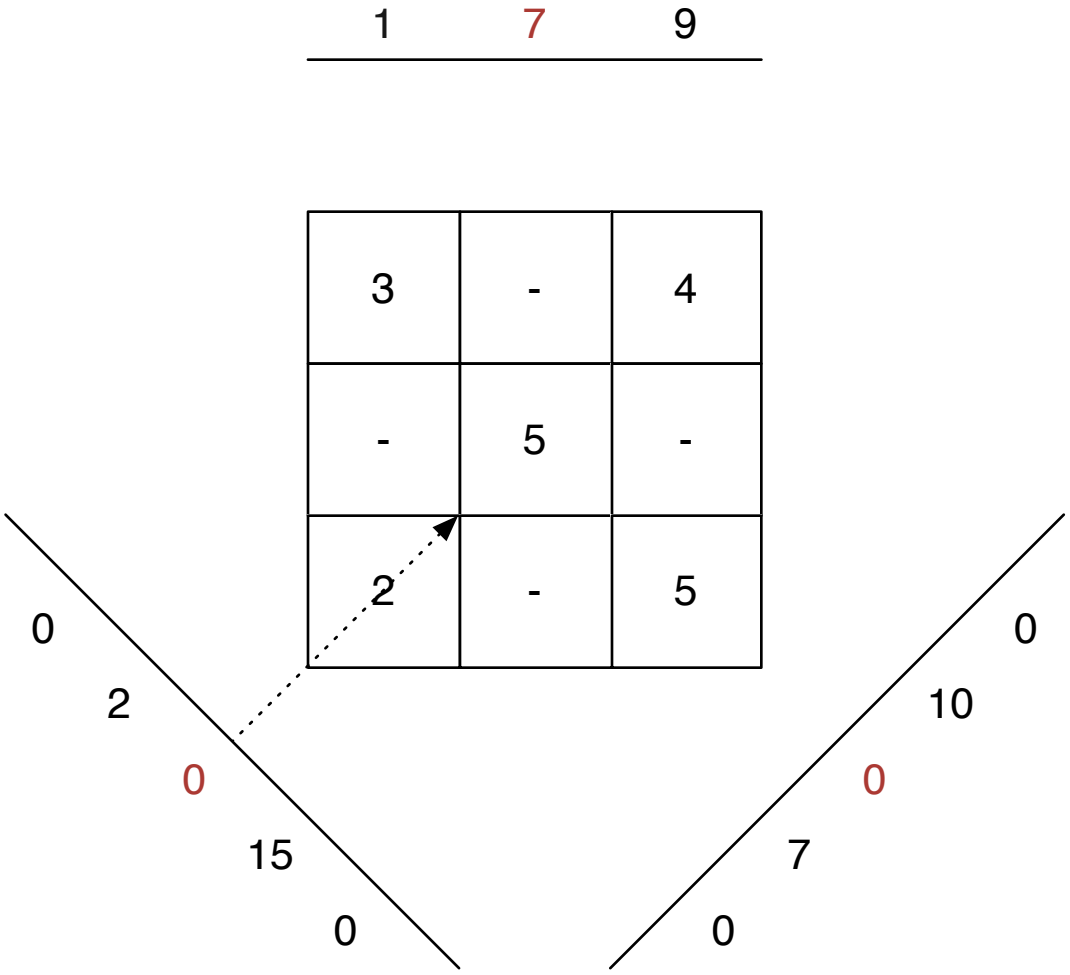
Une transformée de Radon discrète

exacte



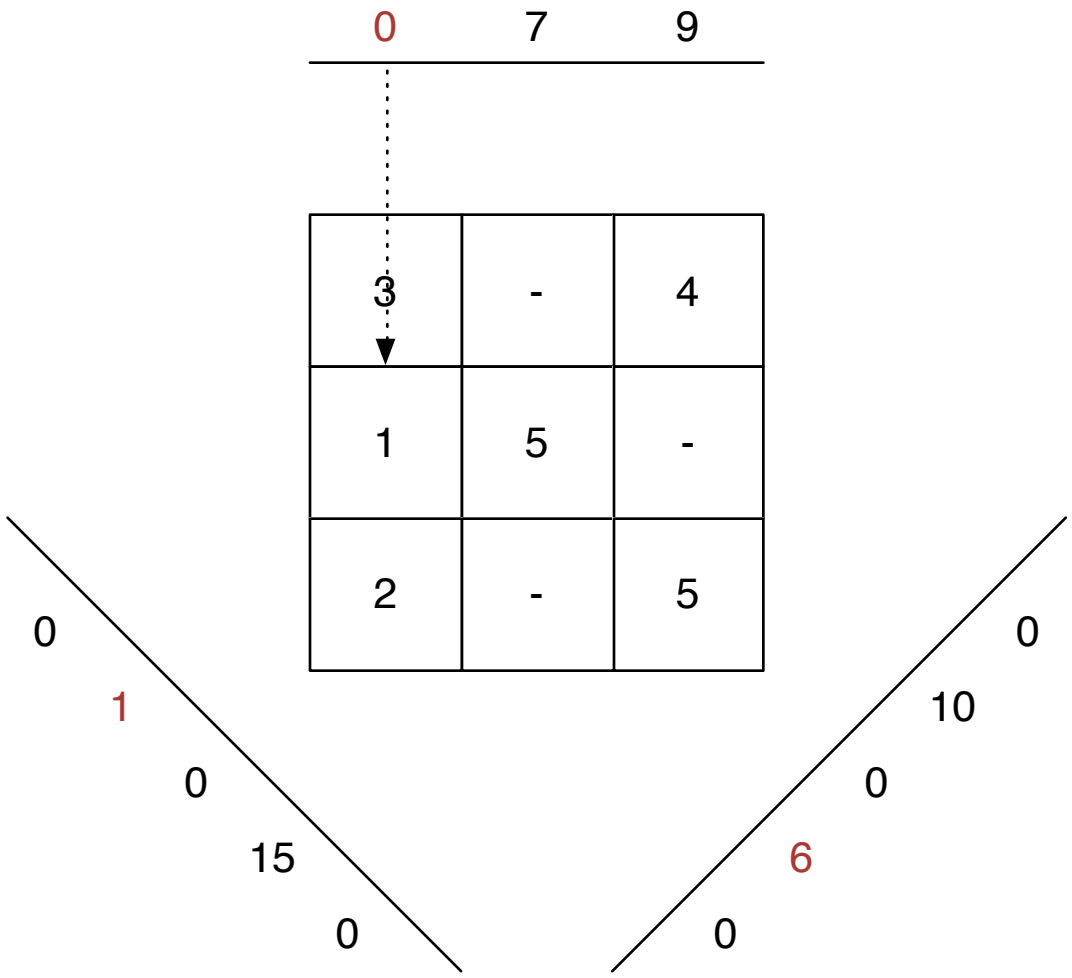
Une transformée de Radon discrète

exacte



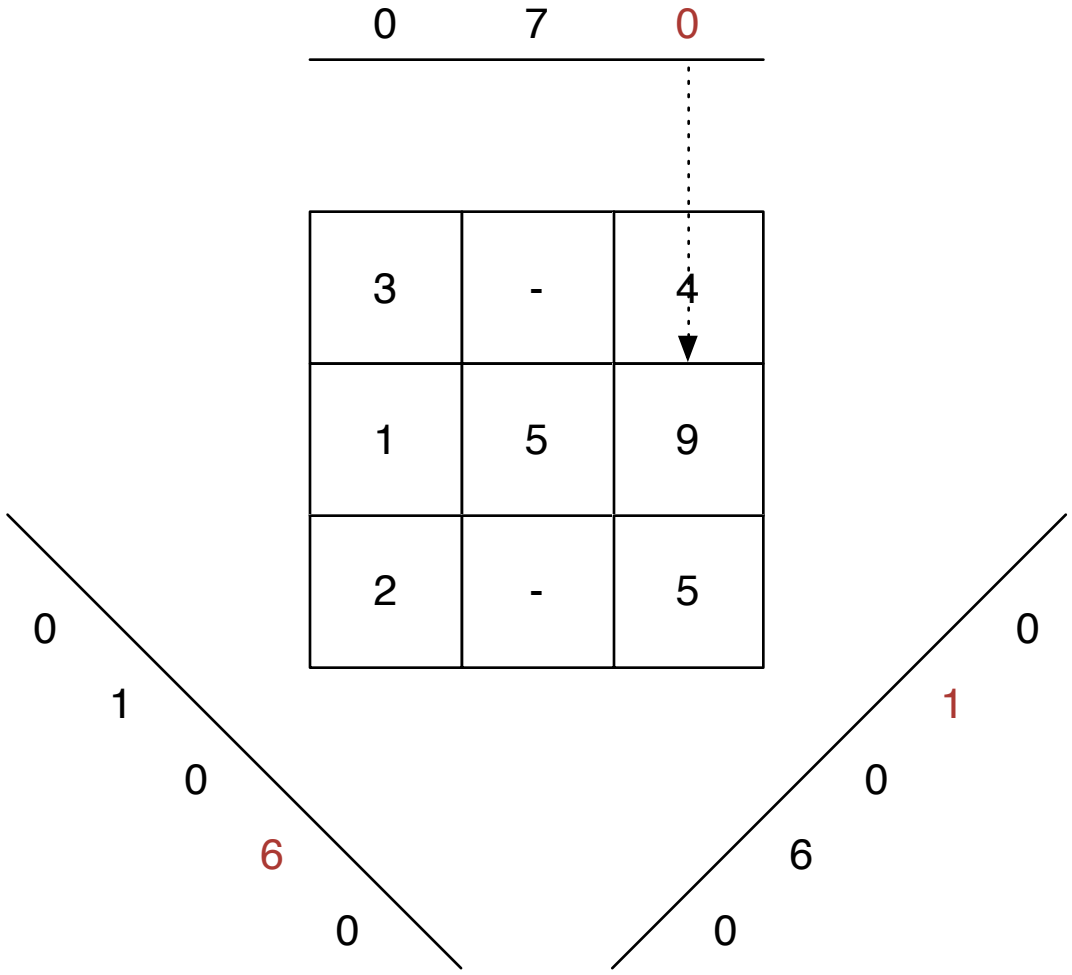
Une transformée de Radon discrète

exacte



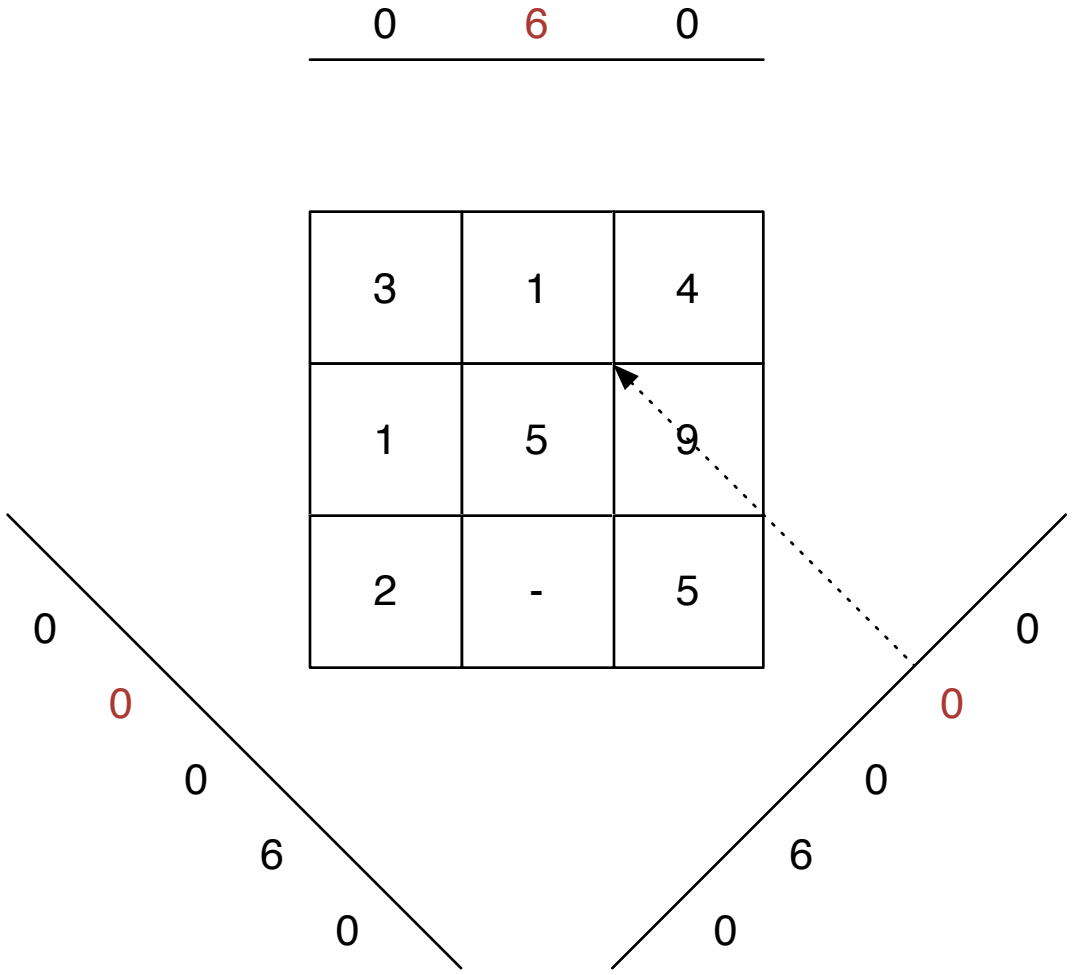
Une transformée de Radon discrète

exacte



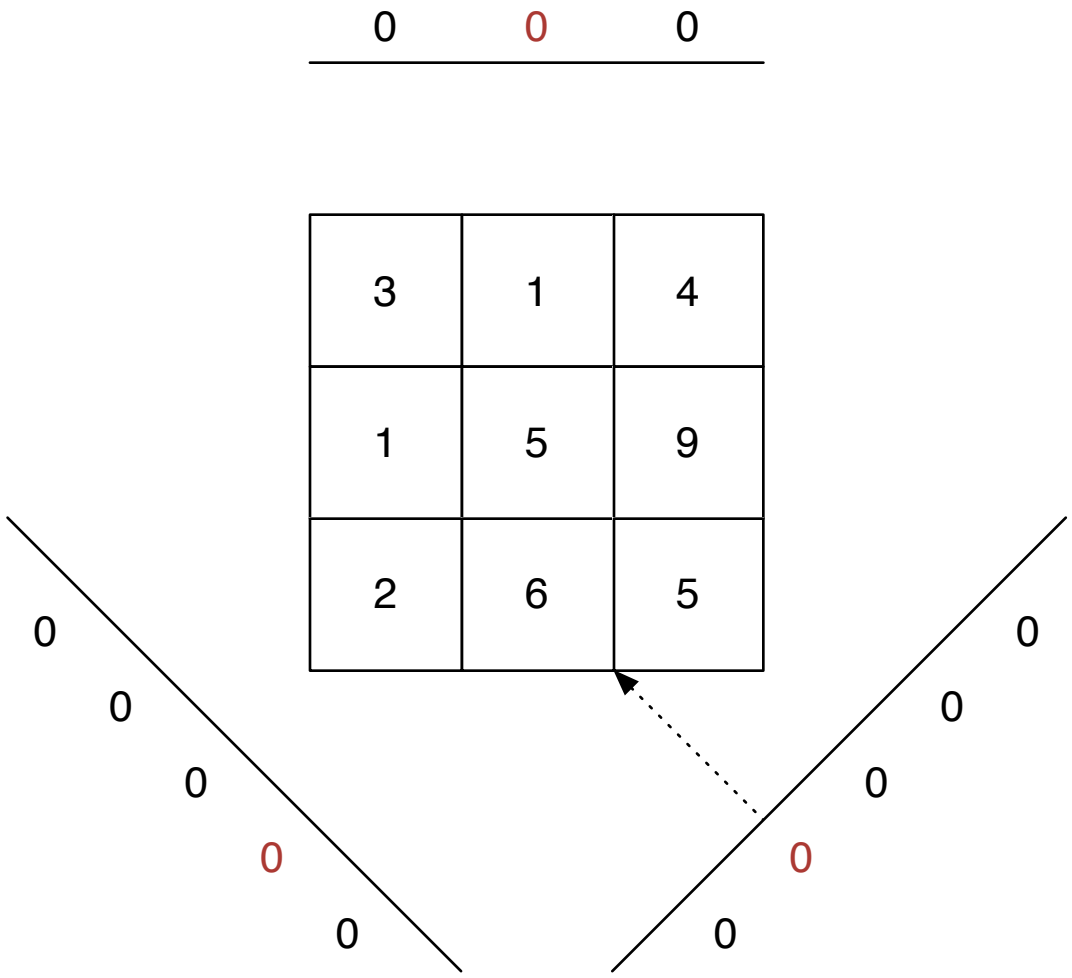
Une transformée de Radon discrète

exacte



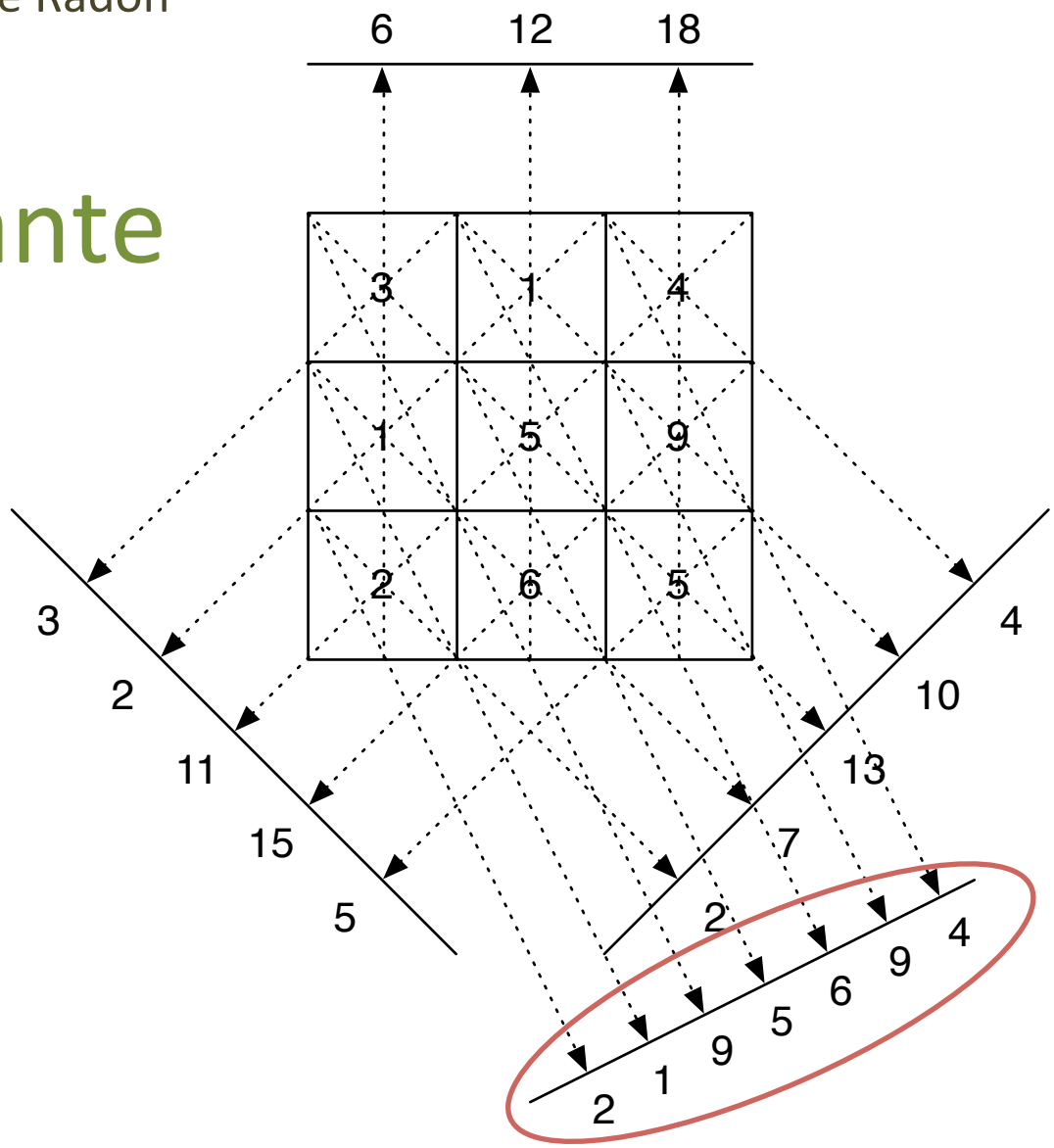
Une transformée de Radon discrète

exacte

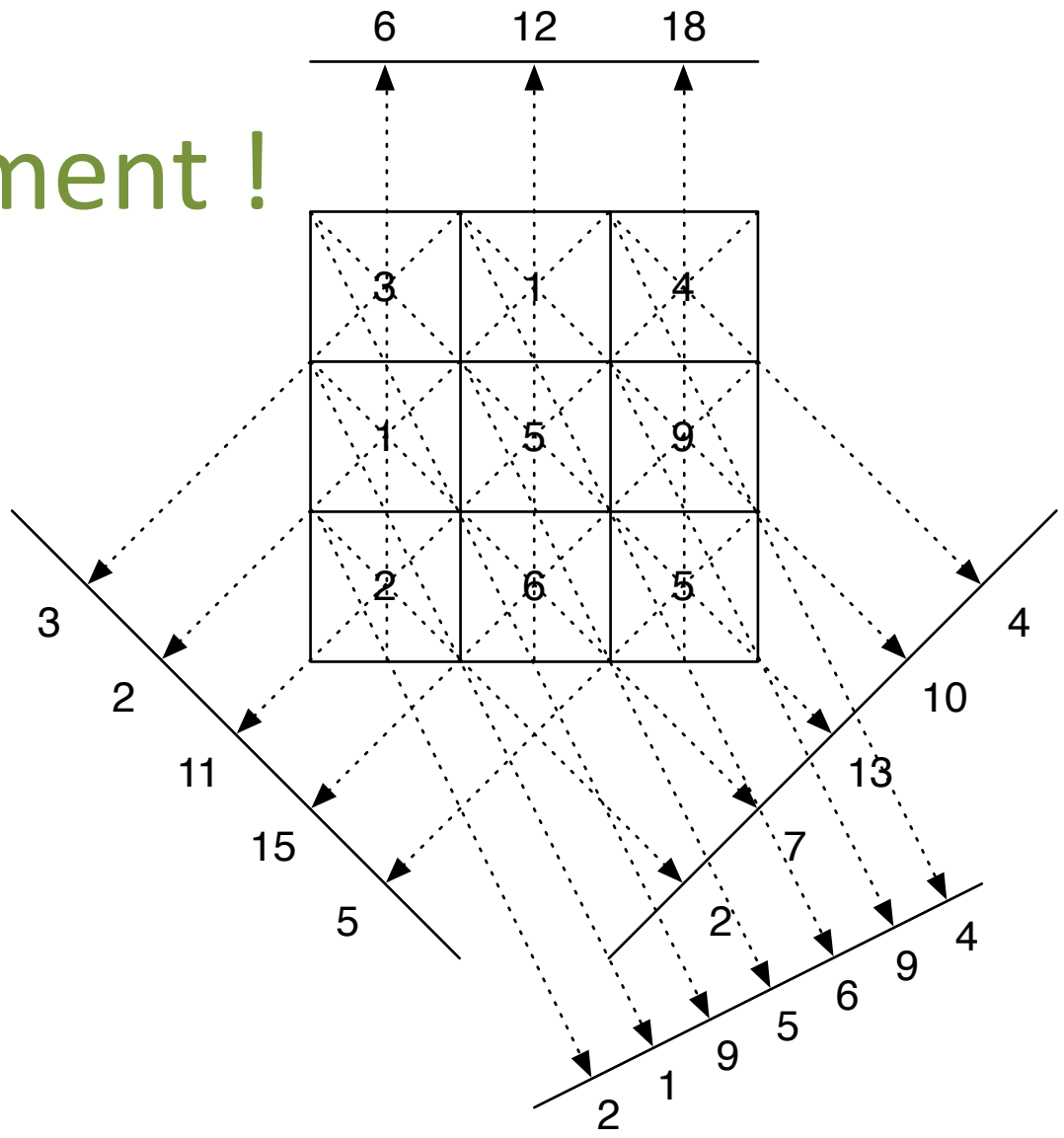


Une transformée de Radon discrète exacte

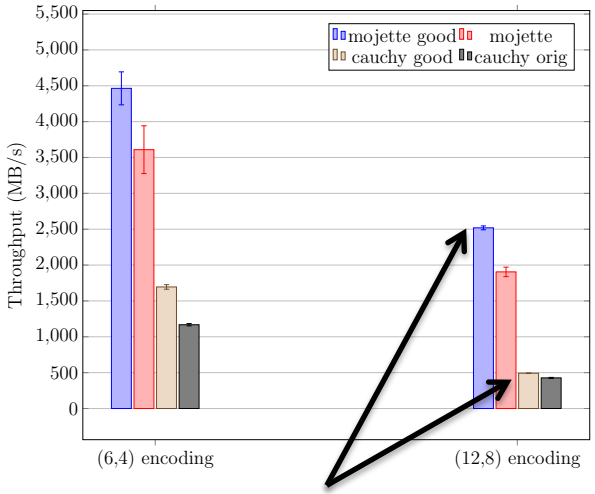
Redondante



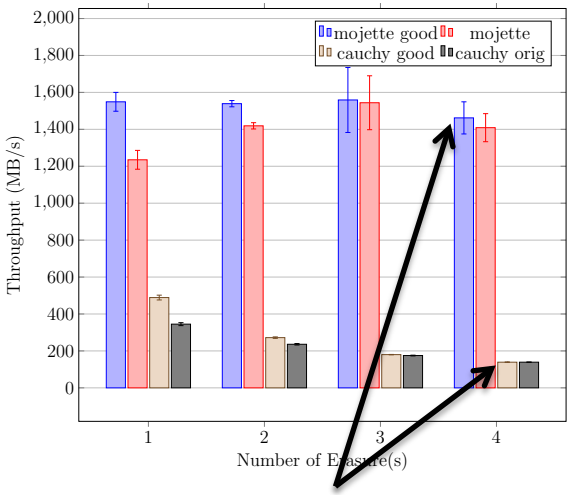
Un code à effacement !



Efficace sur des « petits blocs »



Mojette 5× plus rapide



Mojette 10× plus rapide

Applicable aux données chaudes à accès aléatoires

Deux réponses technologiques émergeantes :

1) Le code à effacement (fiabilité et réduction des coûts),
2) Scale-Out NAS (File System évolutif et simple),

De manière distincte, les deux technologies ne répondent pas entièrement aux besoins (Coûts, Performances et évolutivité) et elles étaient incompatibles.

Fizians a donc créé un code à effacement beaucoup plus performant et un file system adapté, et les a réuni en une solution unique :

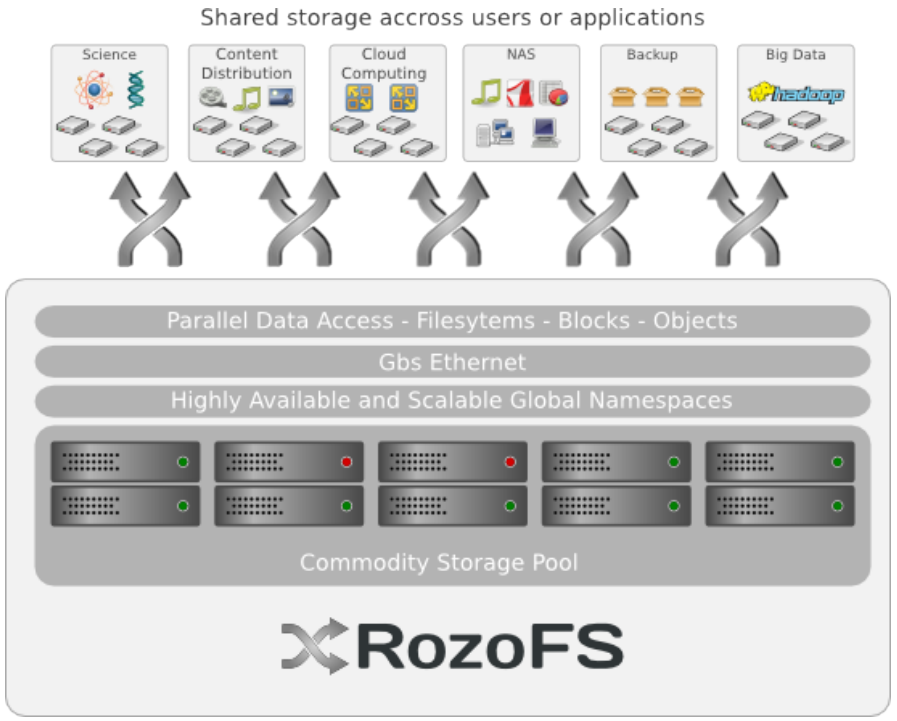
RozoFS
Erasure Code Based Scale-out NAS

Vous offrant une solution économique, simple, fiable et performante.

RozoFS
Erasure Code Based Scale-out NAS

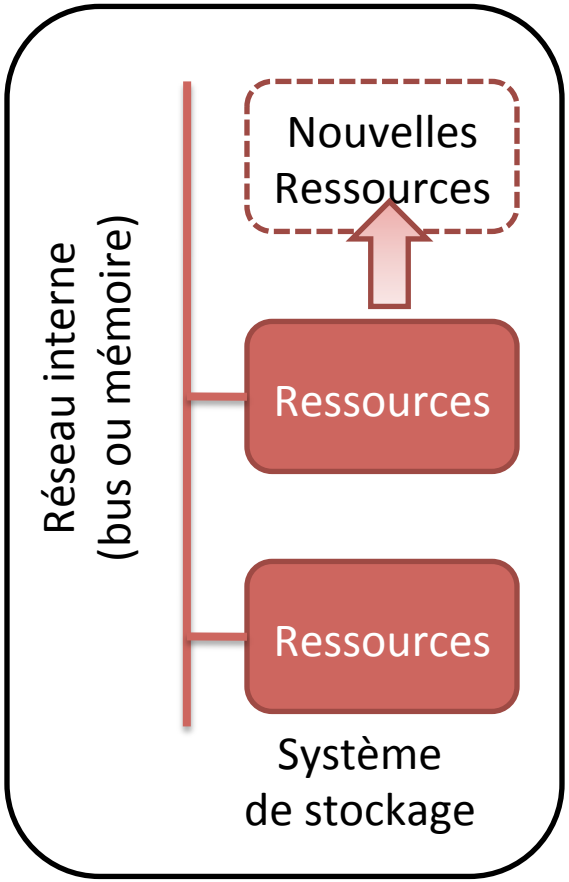
Scale-out NAS avec code à effacement

- ✓**Algorithme rapide et adapté aux systèmes de fichiers.**
- ✓**Temps de reconstruction constant des Blocs**
  - ✓ Any projection is equivalent, storage node failure is the regular case.
- ✓**Sans aucun point de panne isolé,**
  - ✓ Projections read/write are performed in parallel
- ✓**Faible latence**
  - ✓ Use several physical paths for projections read/write.
- ✓**Data-path split-brain free**

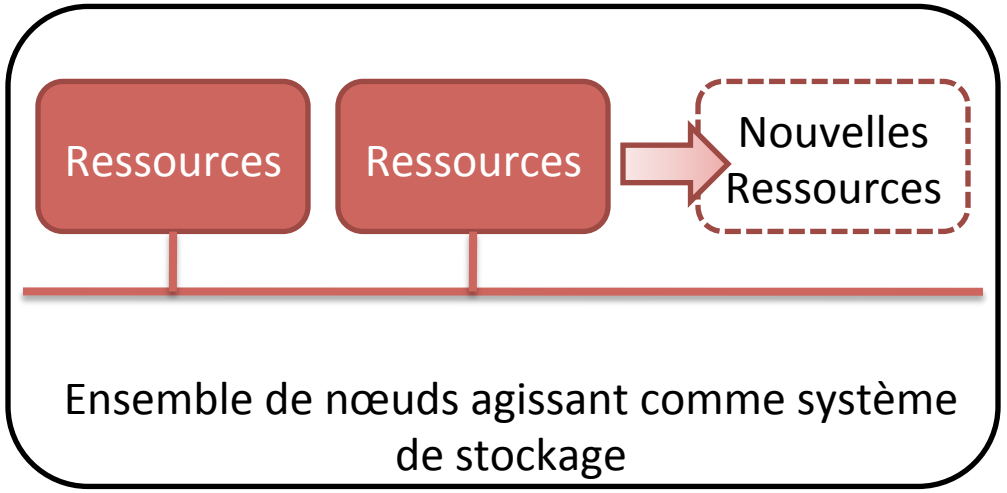


SCALE-OUT STORAGE

SCALE UP



SCALE OUT



- Plus extensible
- Plus performant
- Plus économique

La panne est la norme

Les volumes explosent

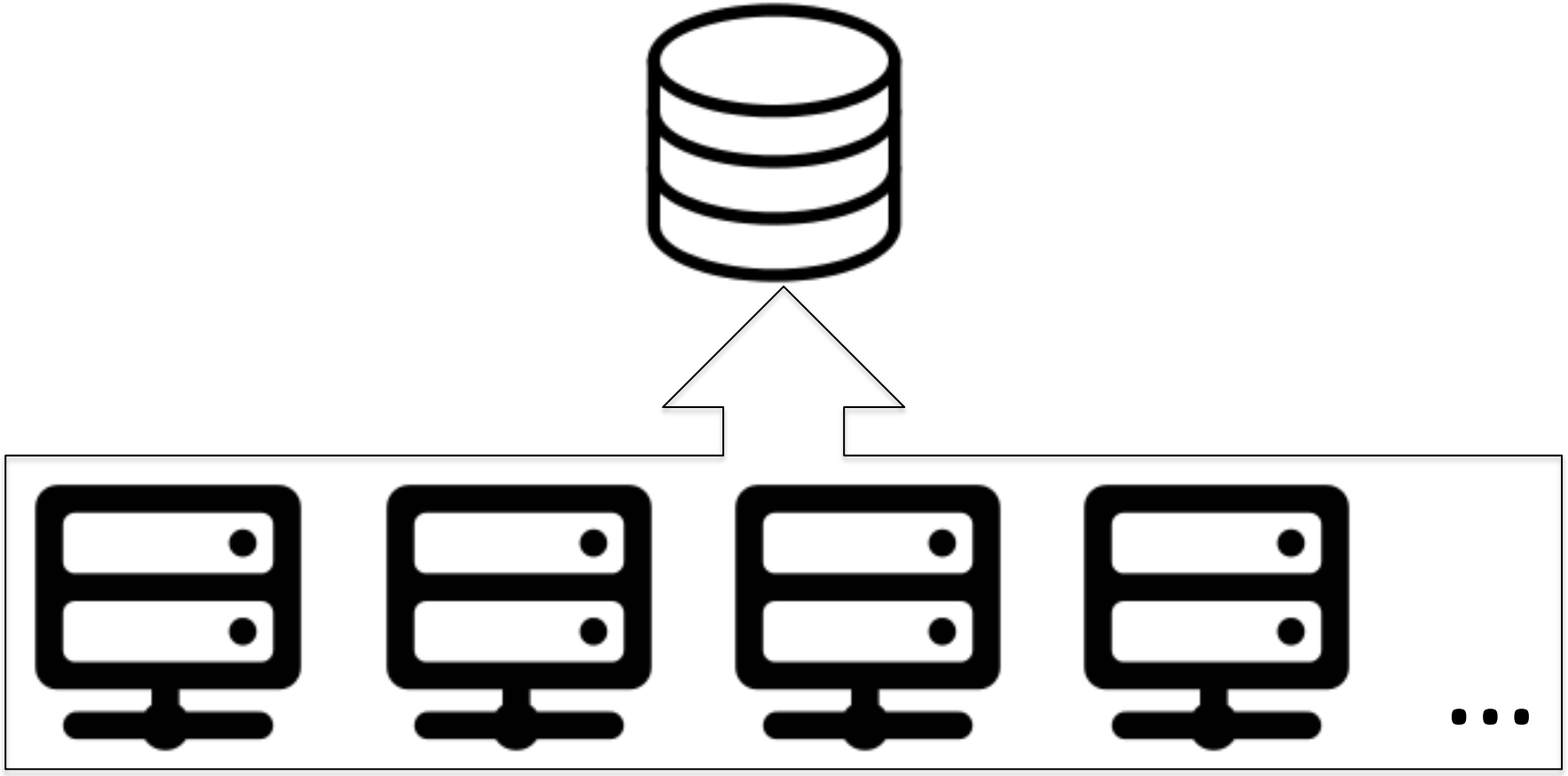
Le code à effacement devient incontournable

ROZOFS

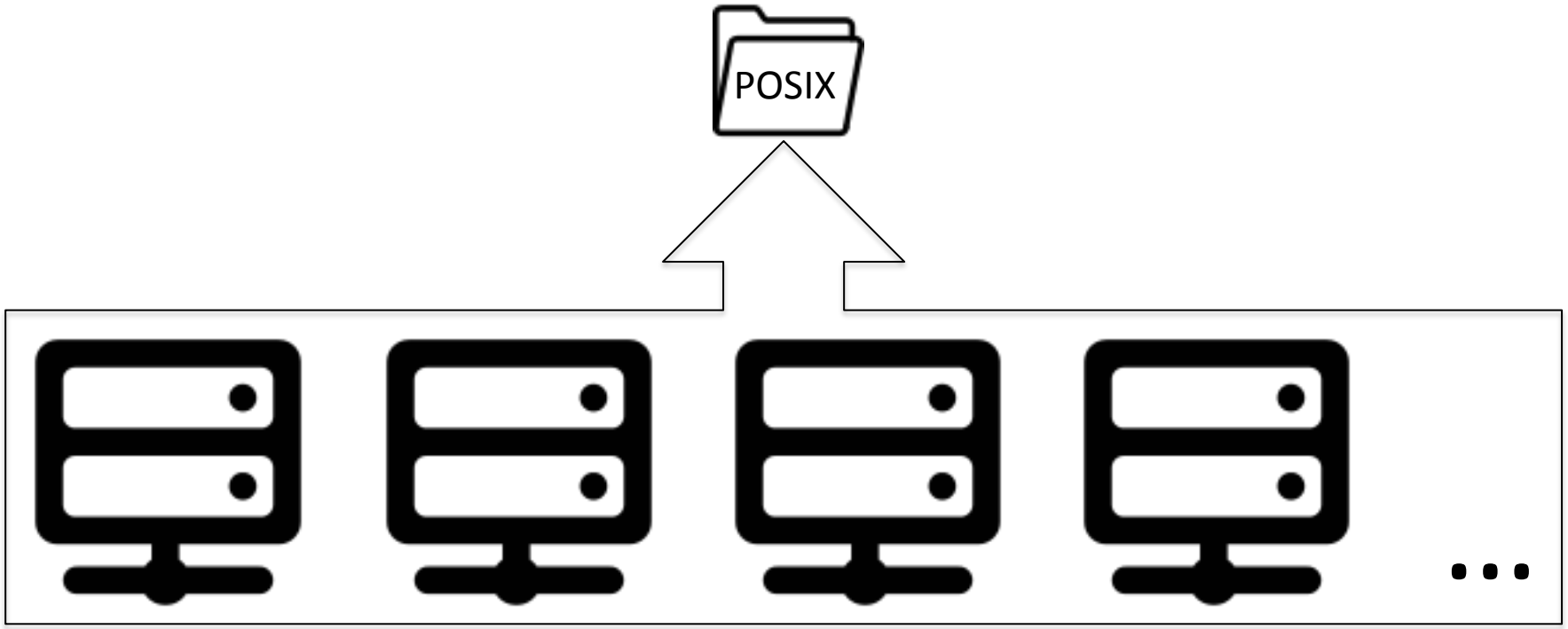
Principe et architecture

Un Scale-out

The more you add the more you get

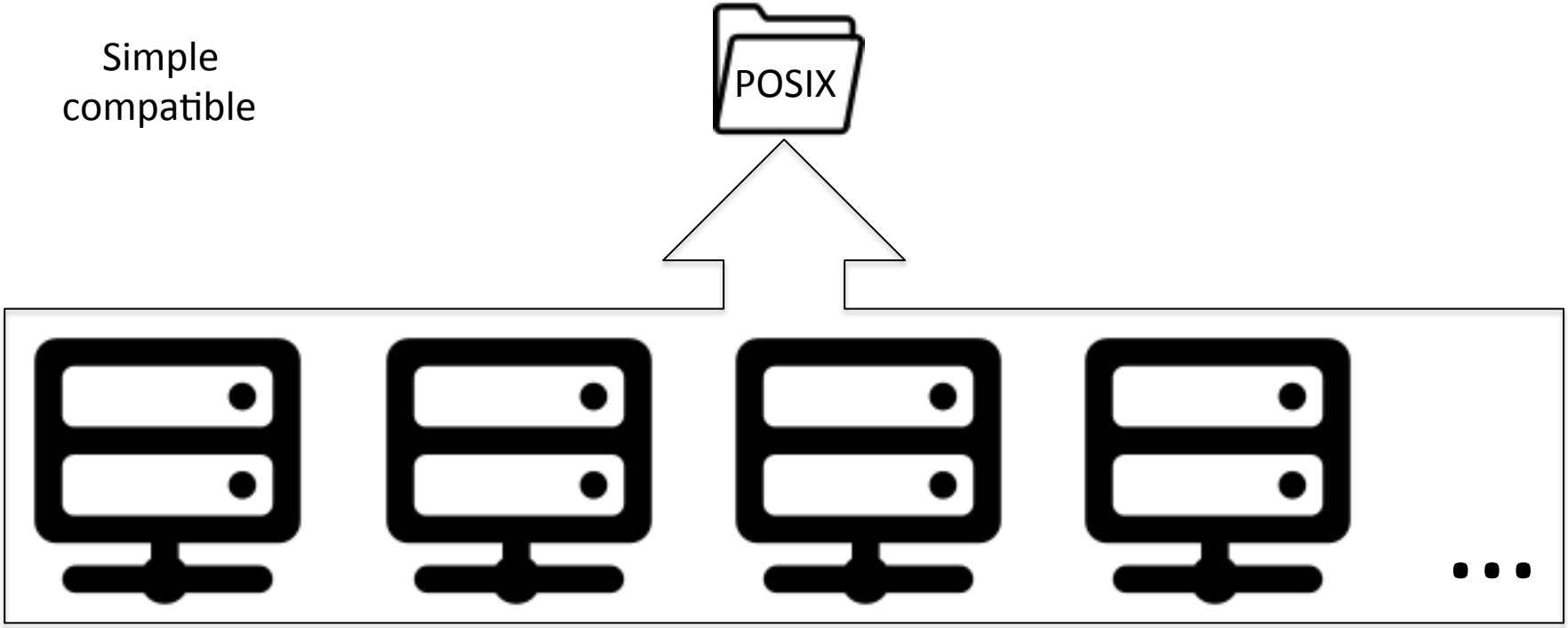


Un Scale-out NAS

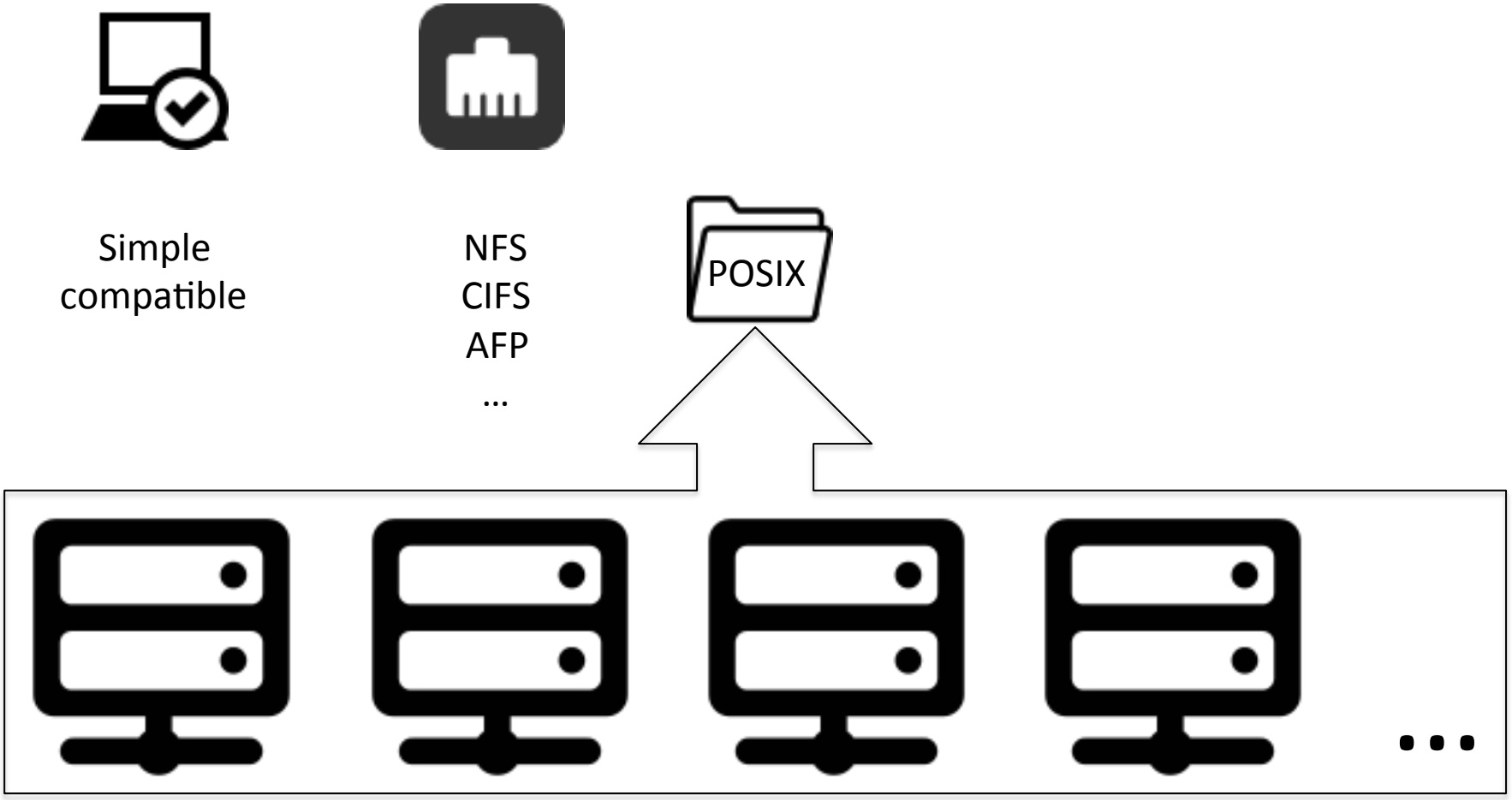


Un Scale-out NAS

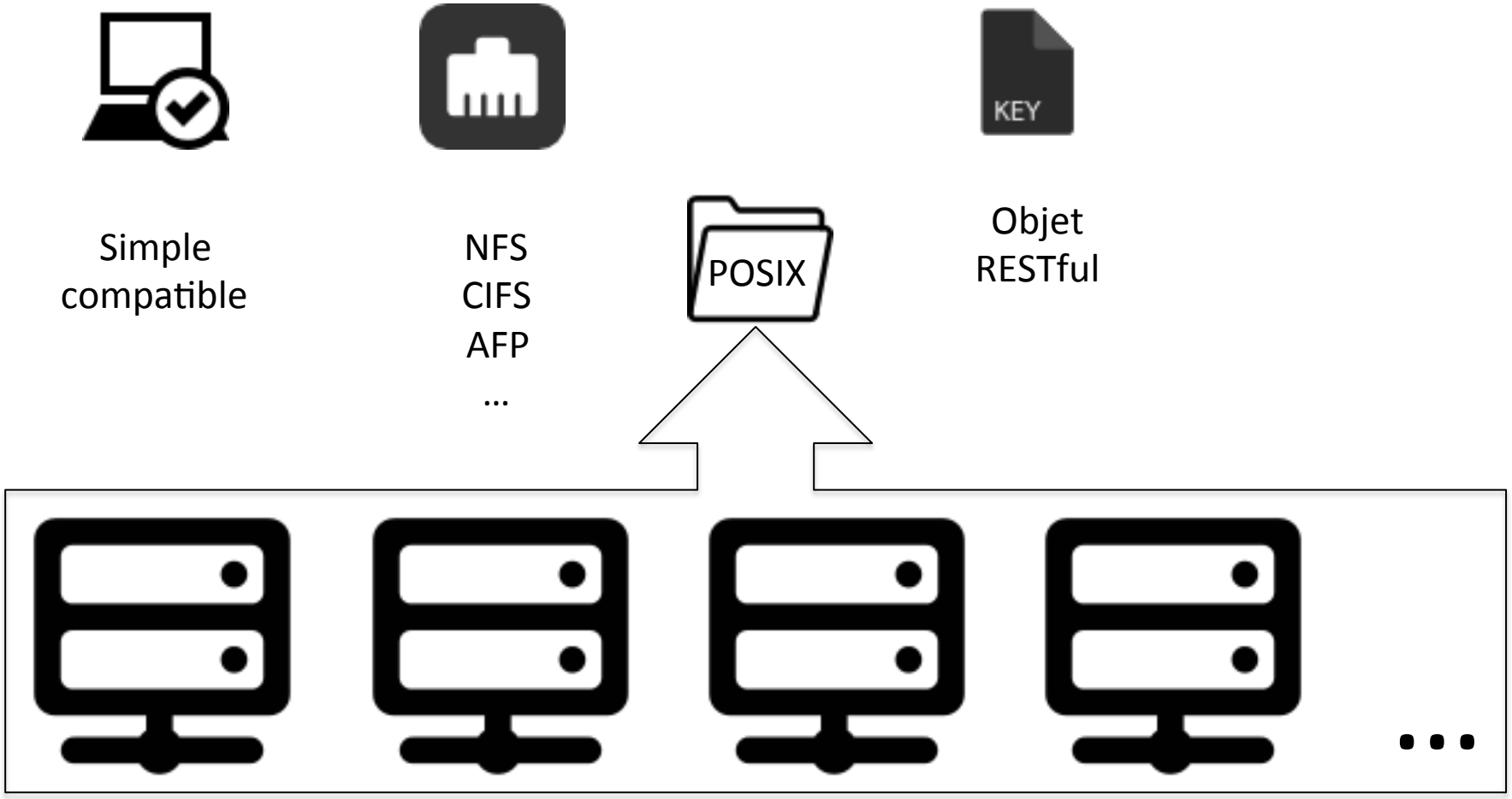
Simple
compatible



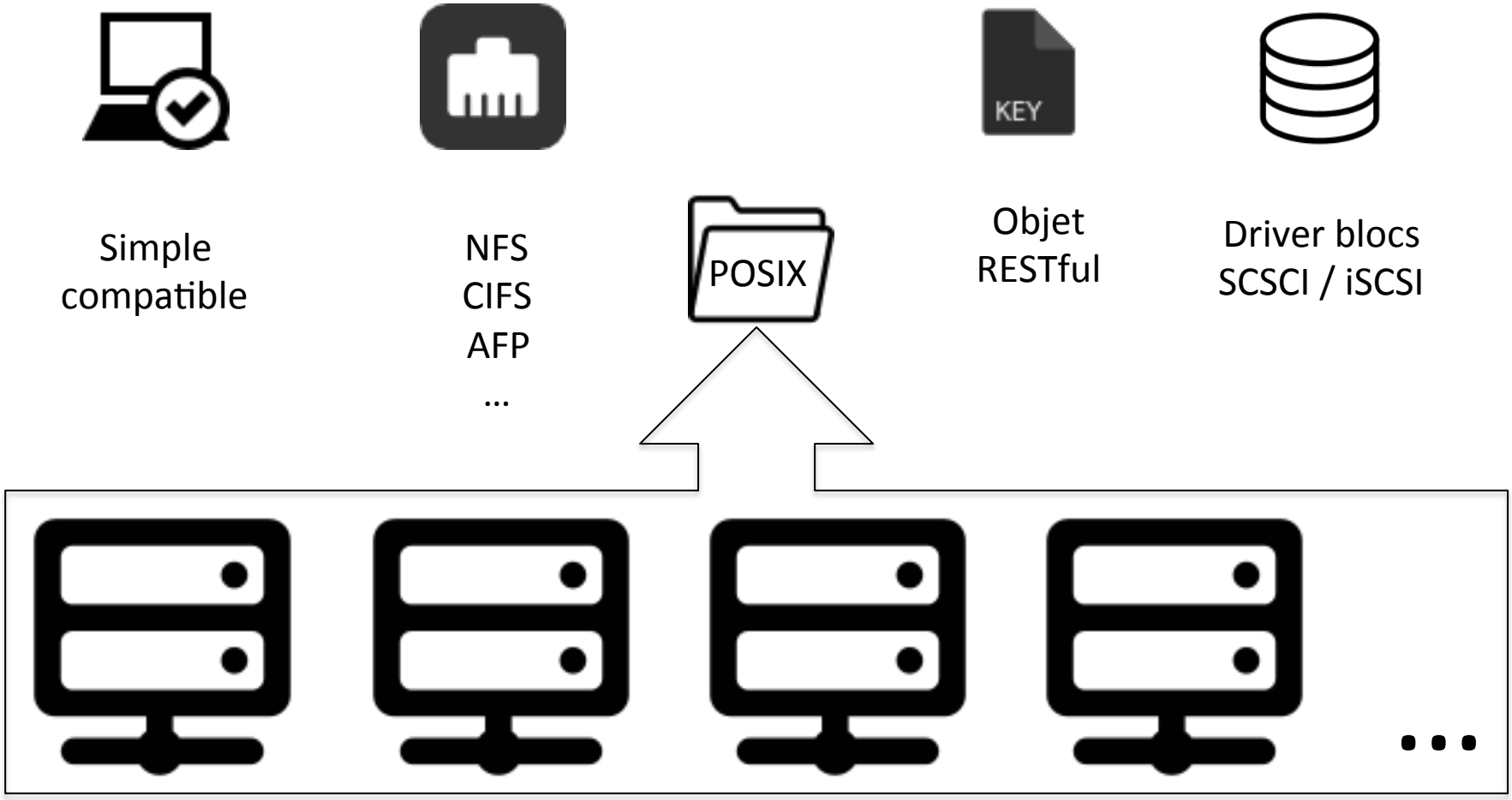
Un Scale-out NAS



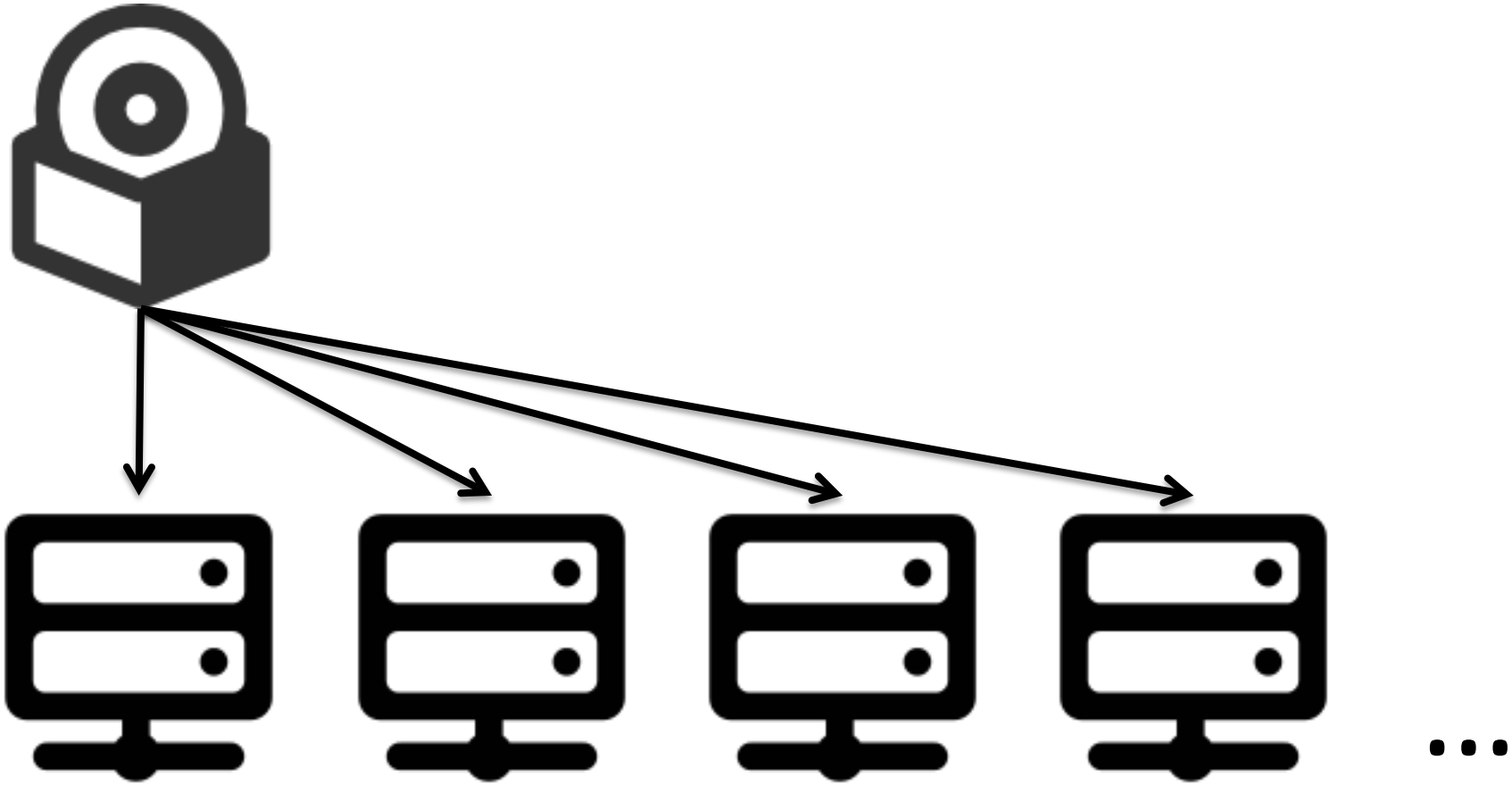
Un Scale-out NAS



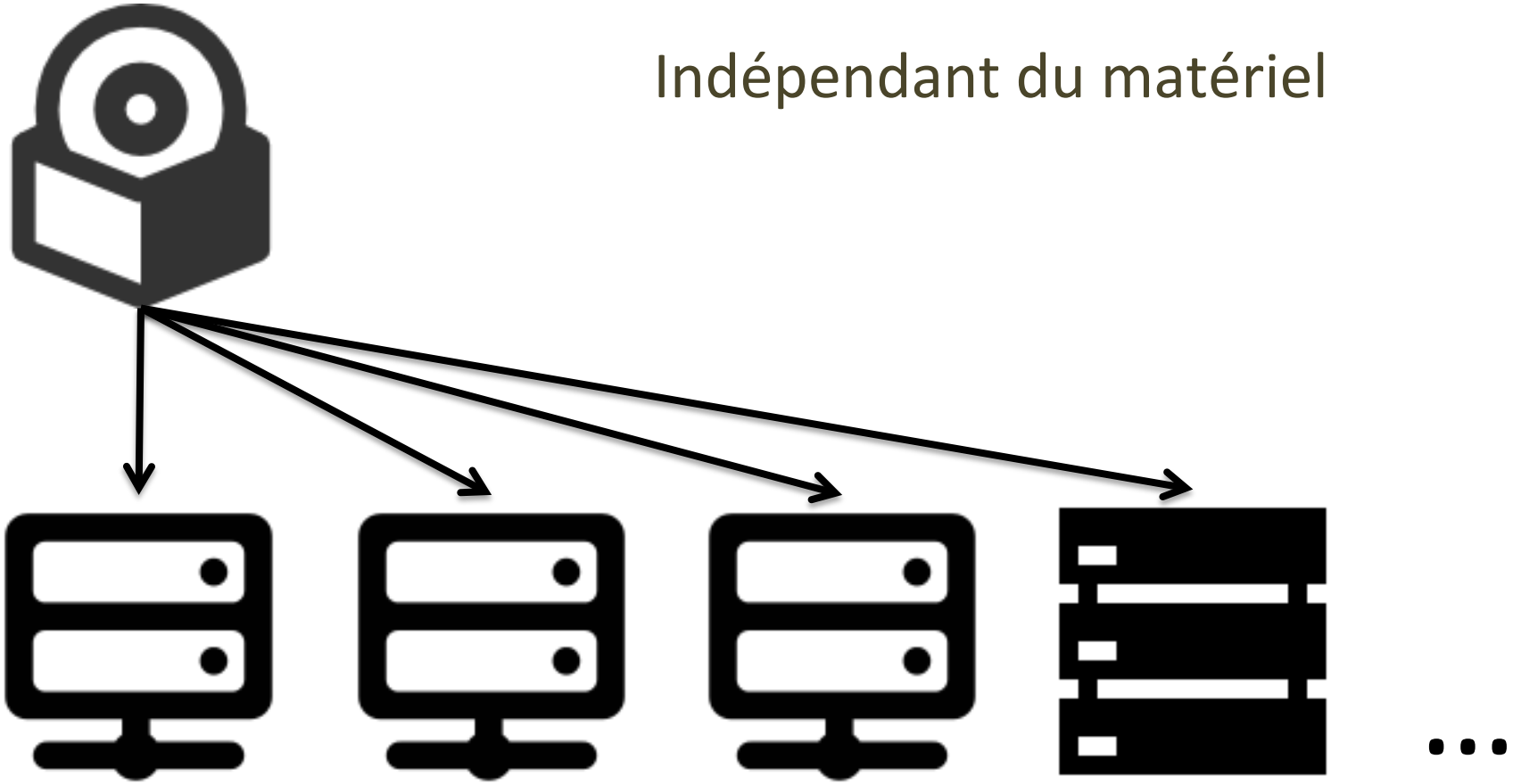
Un Scale-out NAS



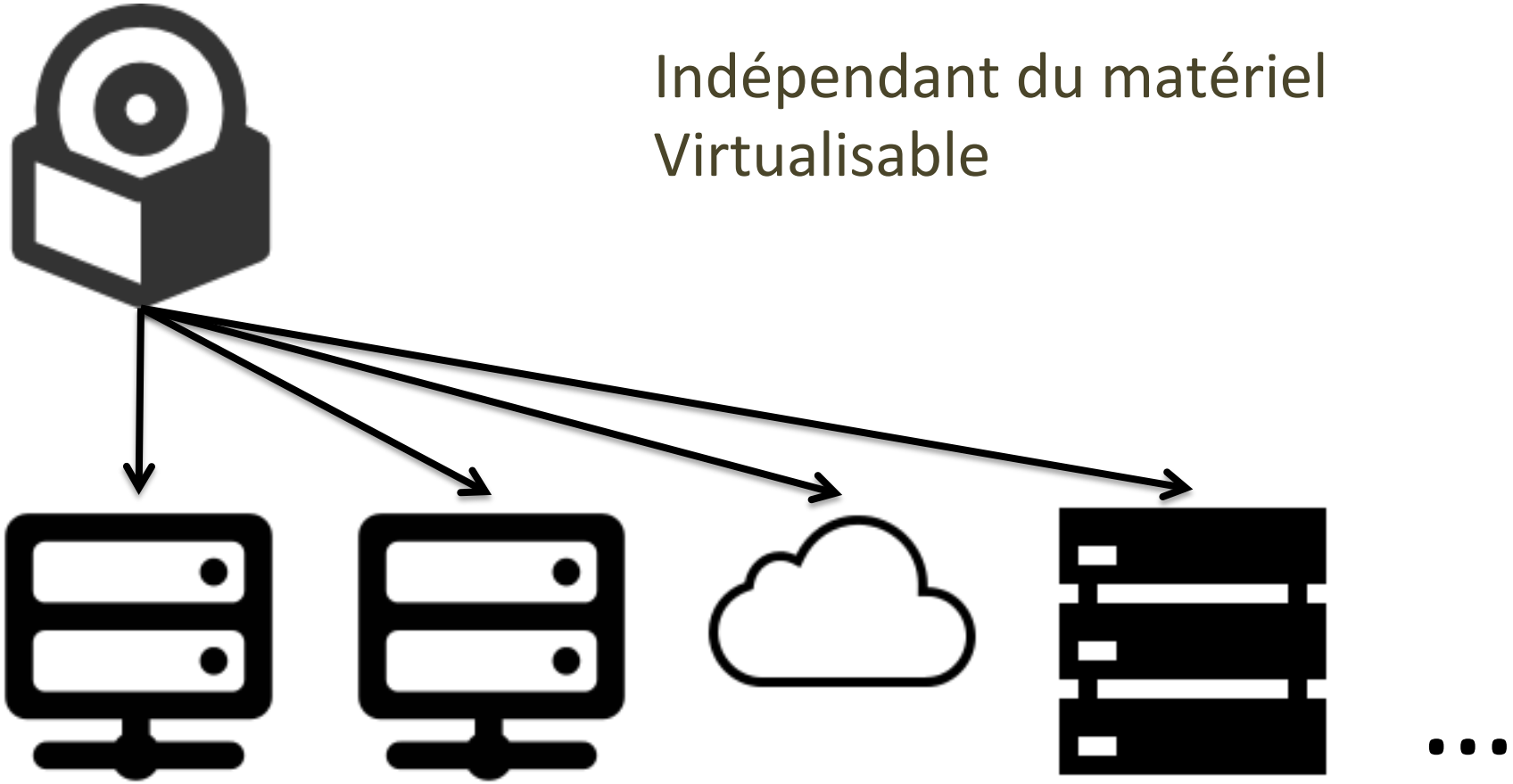
Un Scale-out NAS logiciel



Un Scale-out NAS logiciel



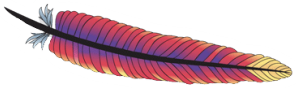
Un Scale-out NAS logiciel

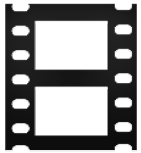


Indépendant du matériel
Virtualisable

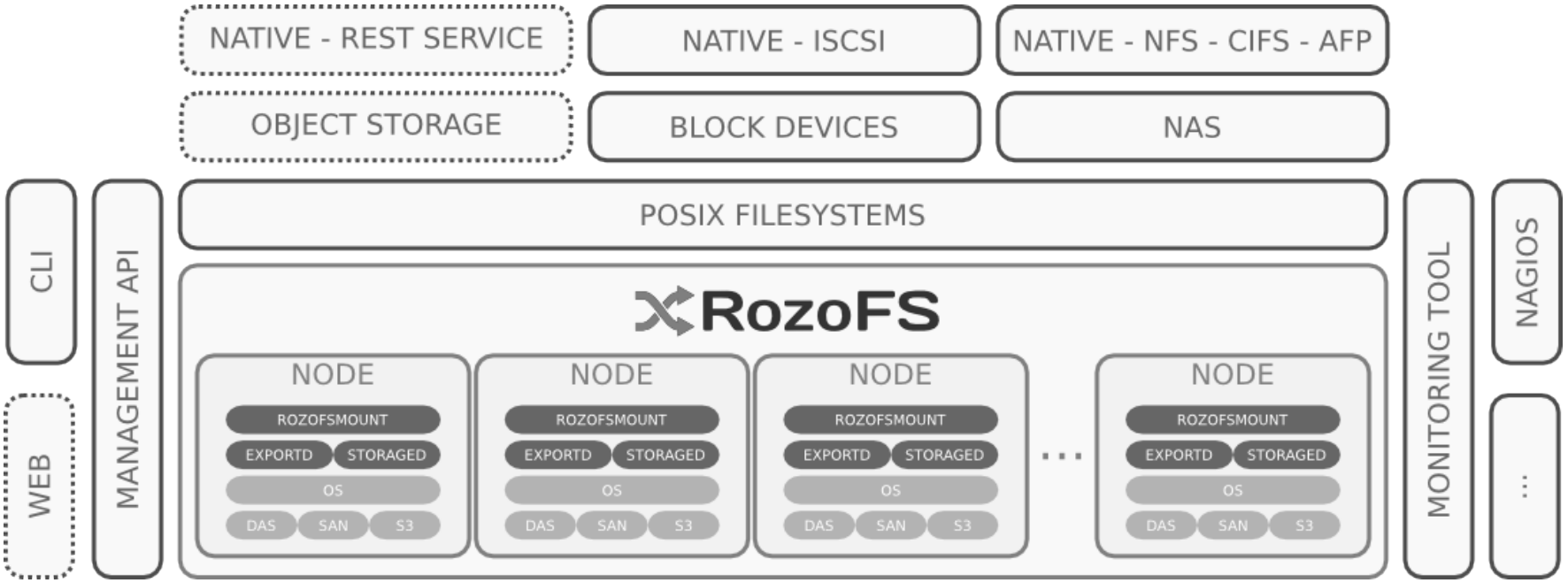
Un Scale-out NAS logiciel

Indépendant du matériel
Virtualisable
Co localisable





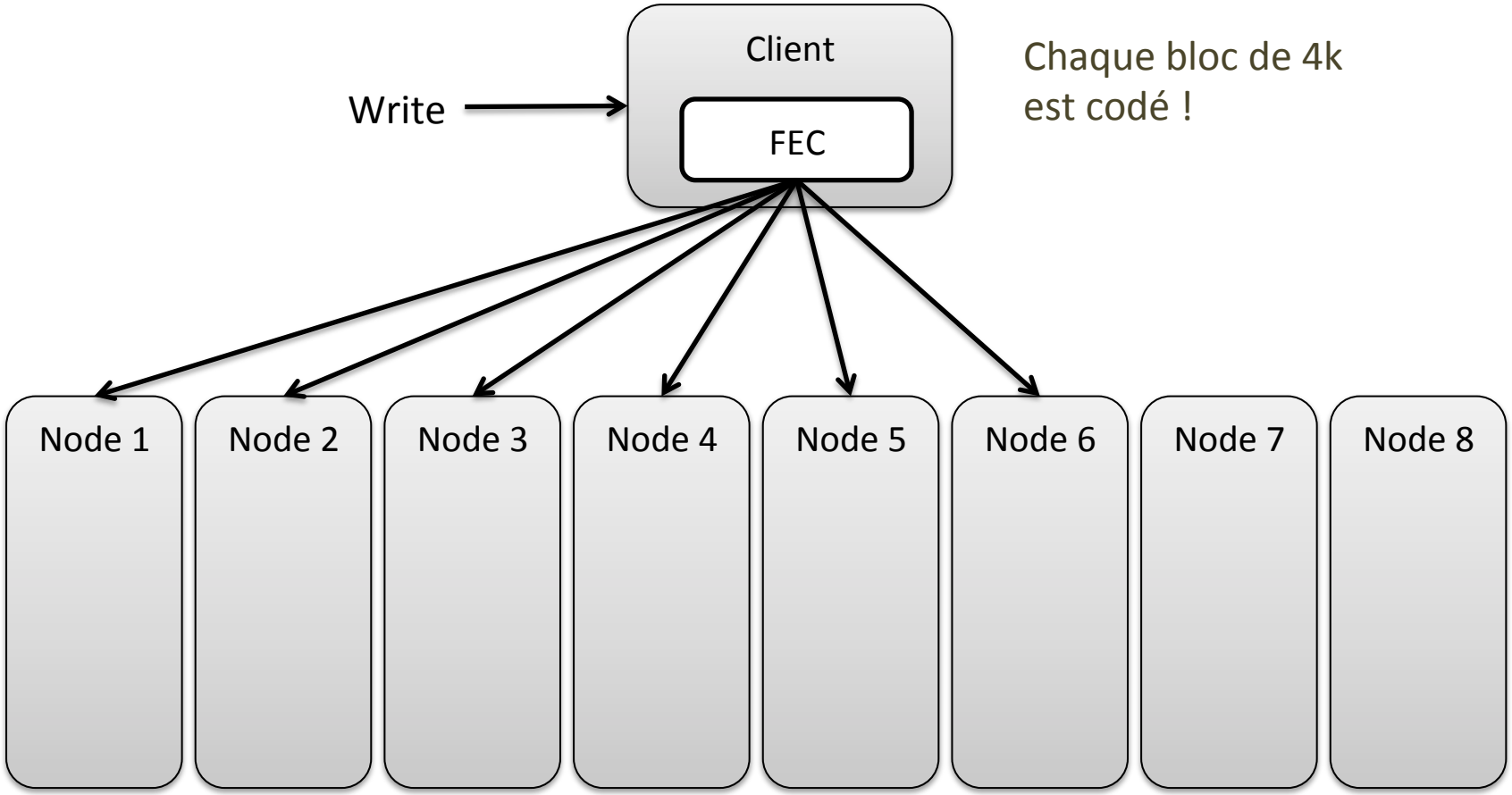
...



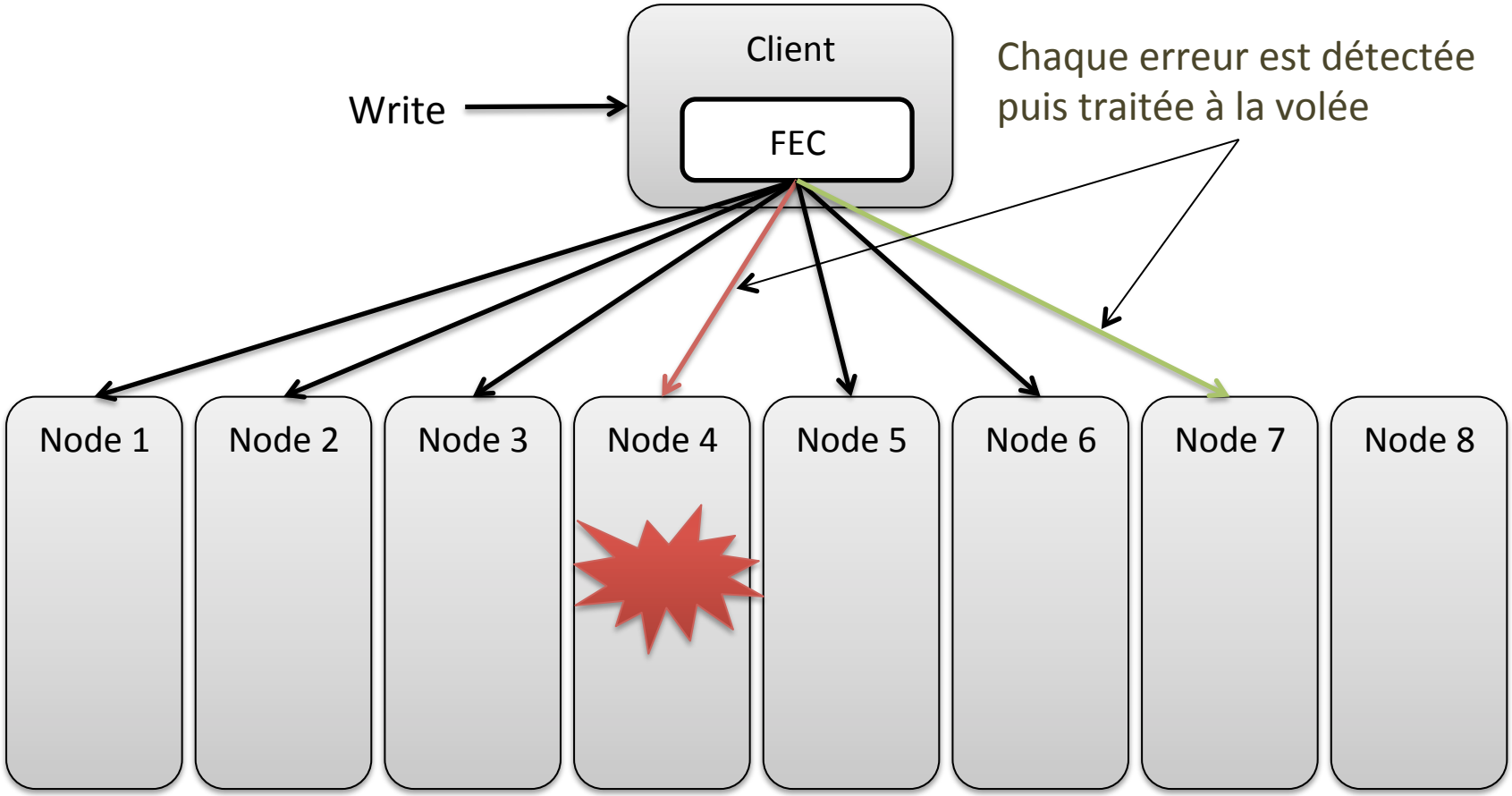
ROZOFS

Protection des données

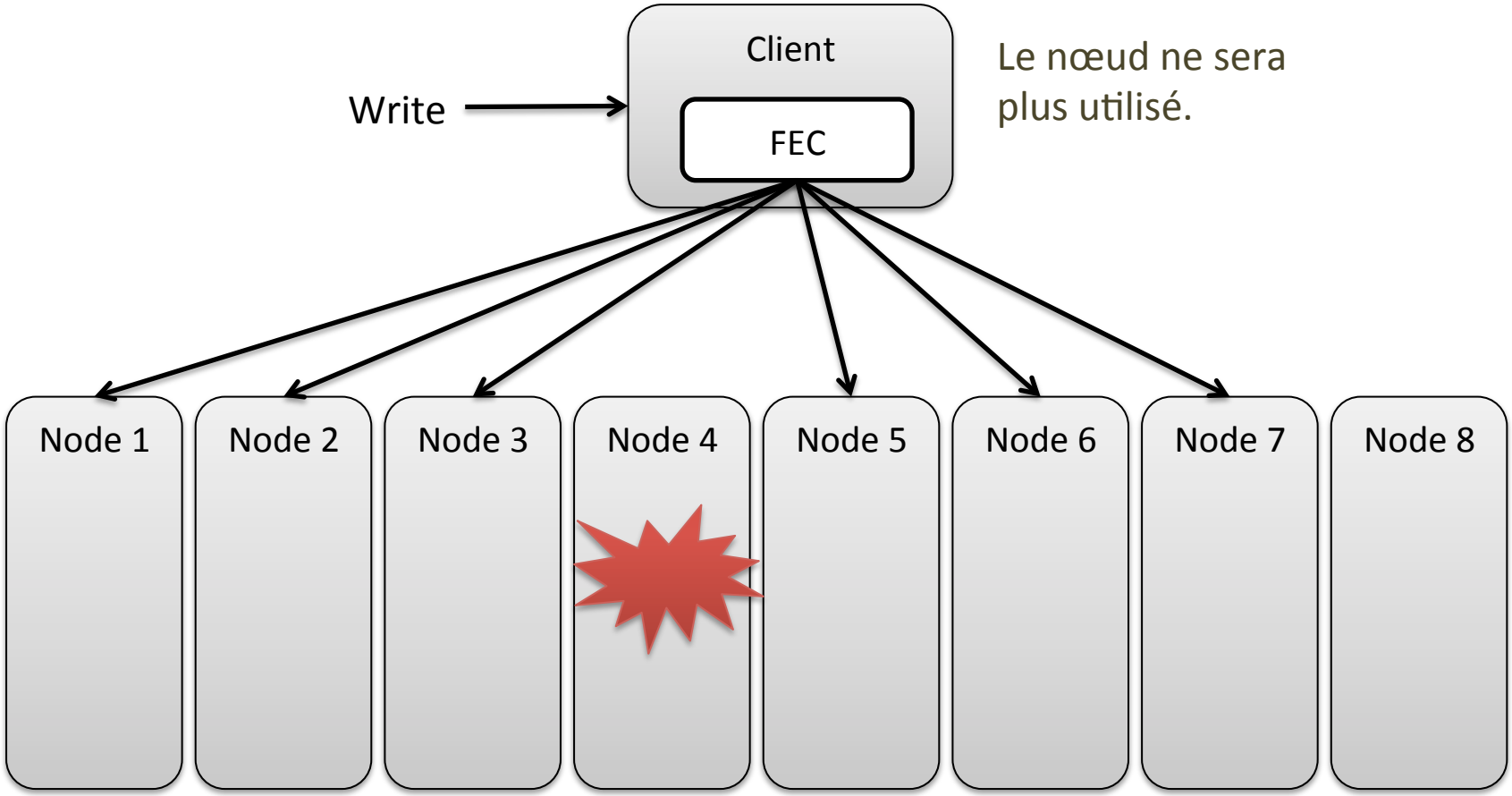
FEC dans le data path



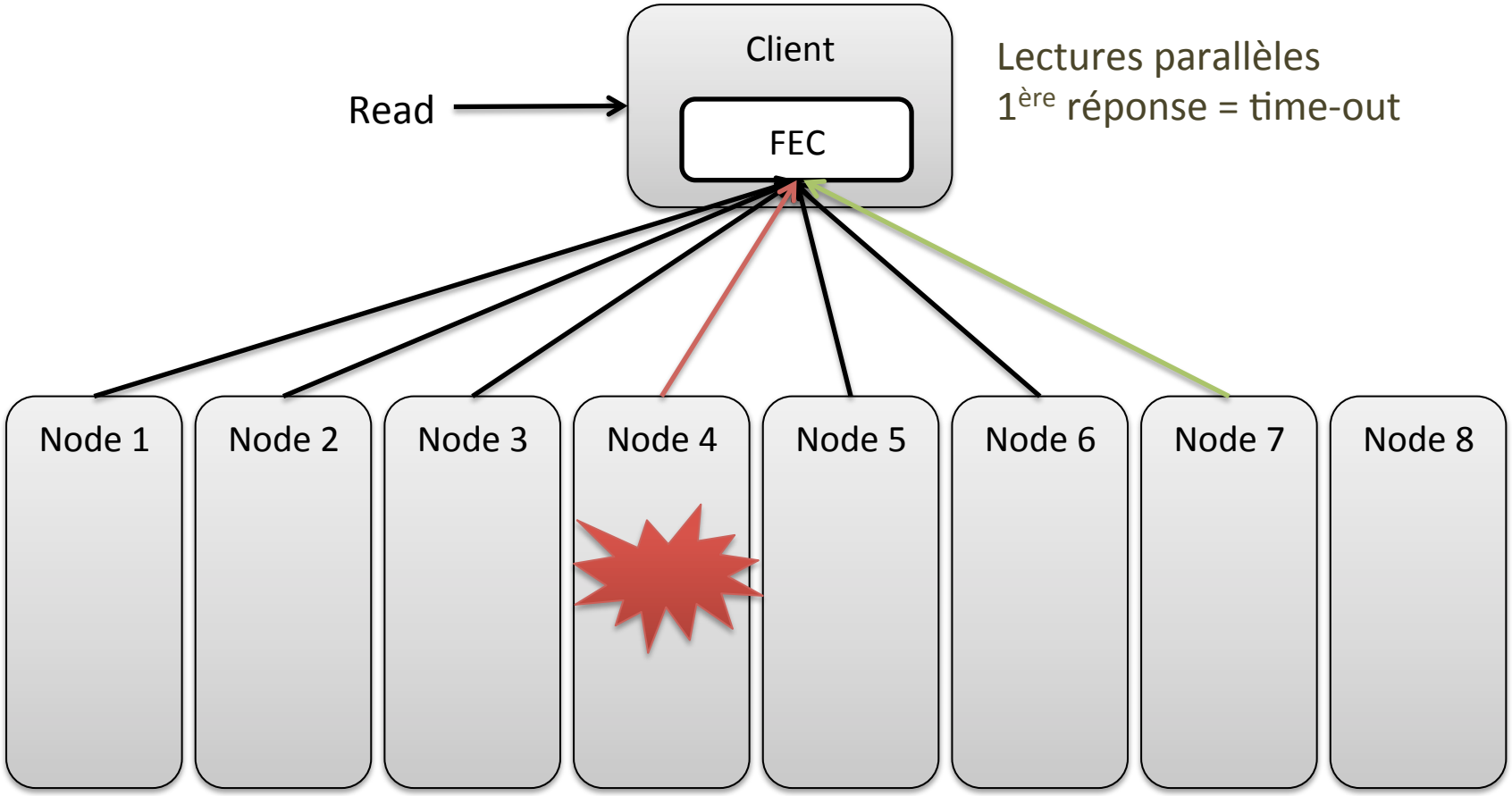
FEC dans le data path



FEC dans le data path



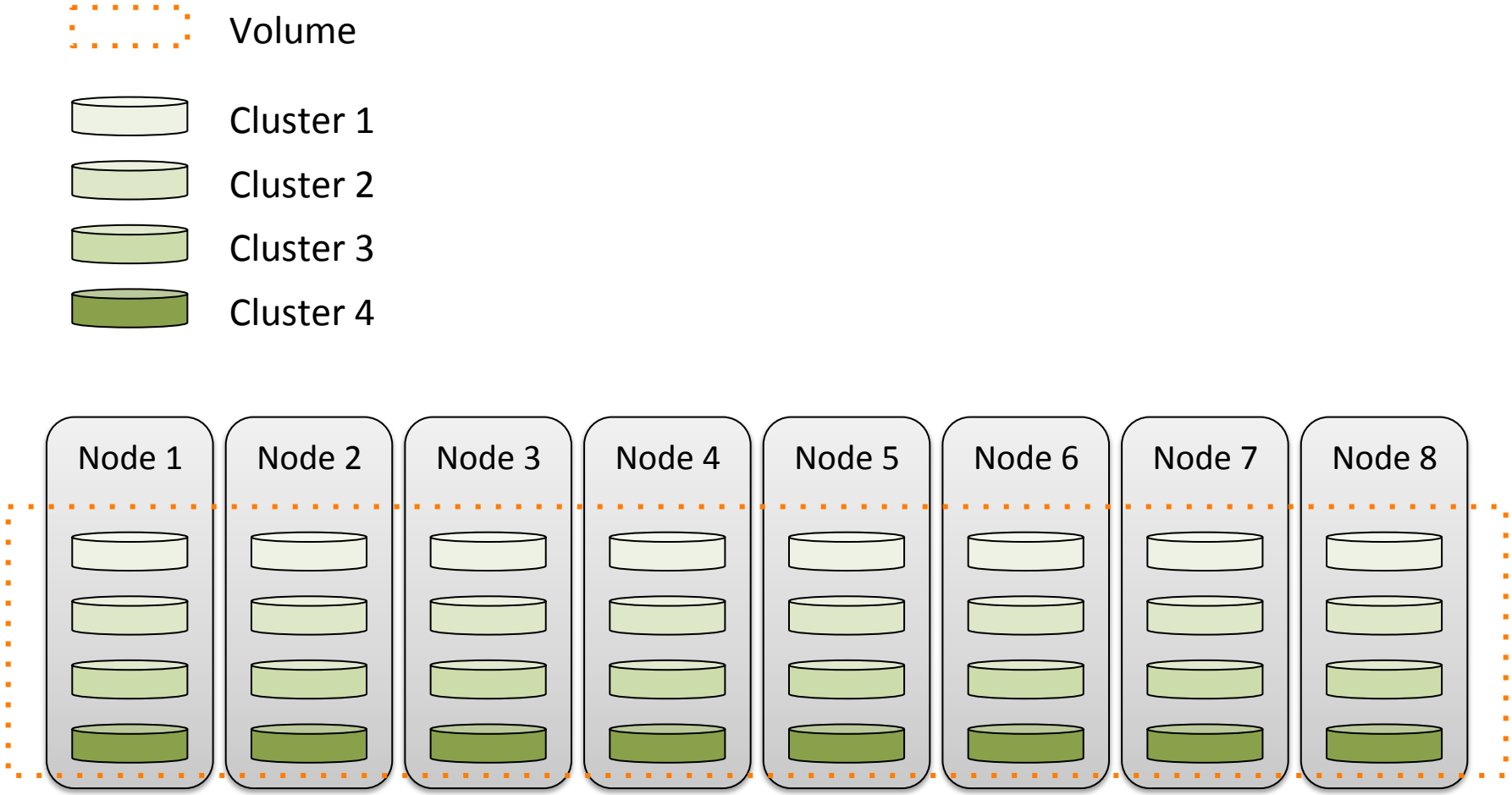
FEC dans le data path



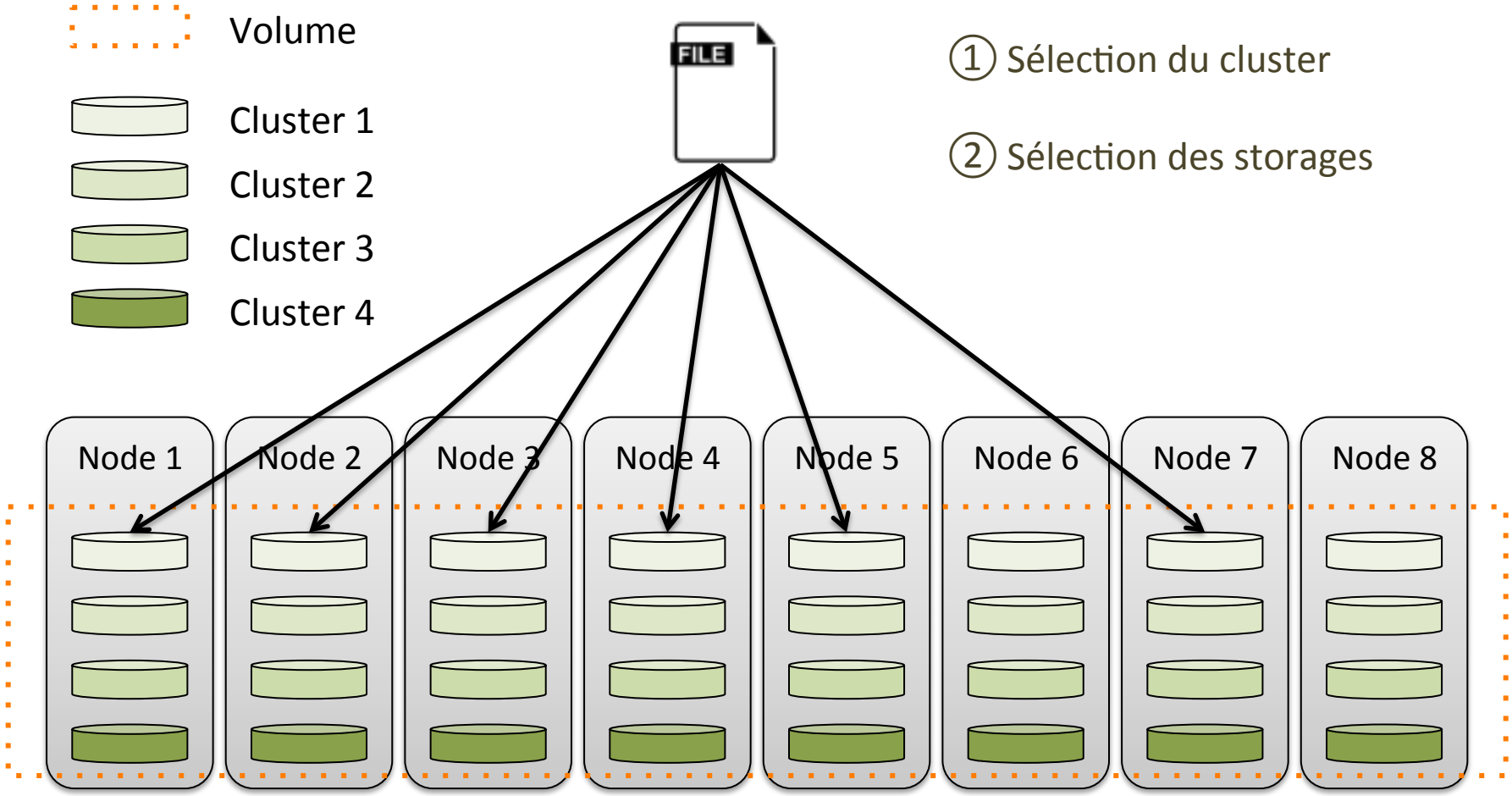
FEC dans le data path

	Replication	RS-based	Mojette (RozoFS)
Durabilité	99,9999	99,9999	99,9999
Codage	-	Write/Pannes	Write/Read
Storage overhead (> 4M)	200%	50%	50%
Storage overhead (< 1M)	200%	500%	50%
Read overhead (panne)	0%	500%	0%
Détection de panne	2000 ms	2000 ms	200 ms

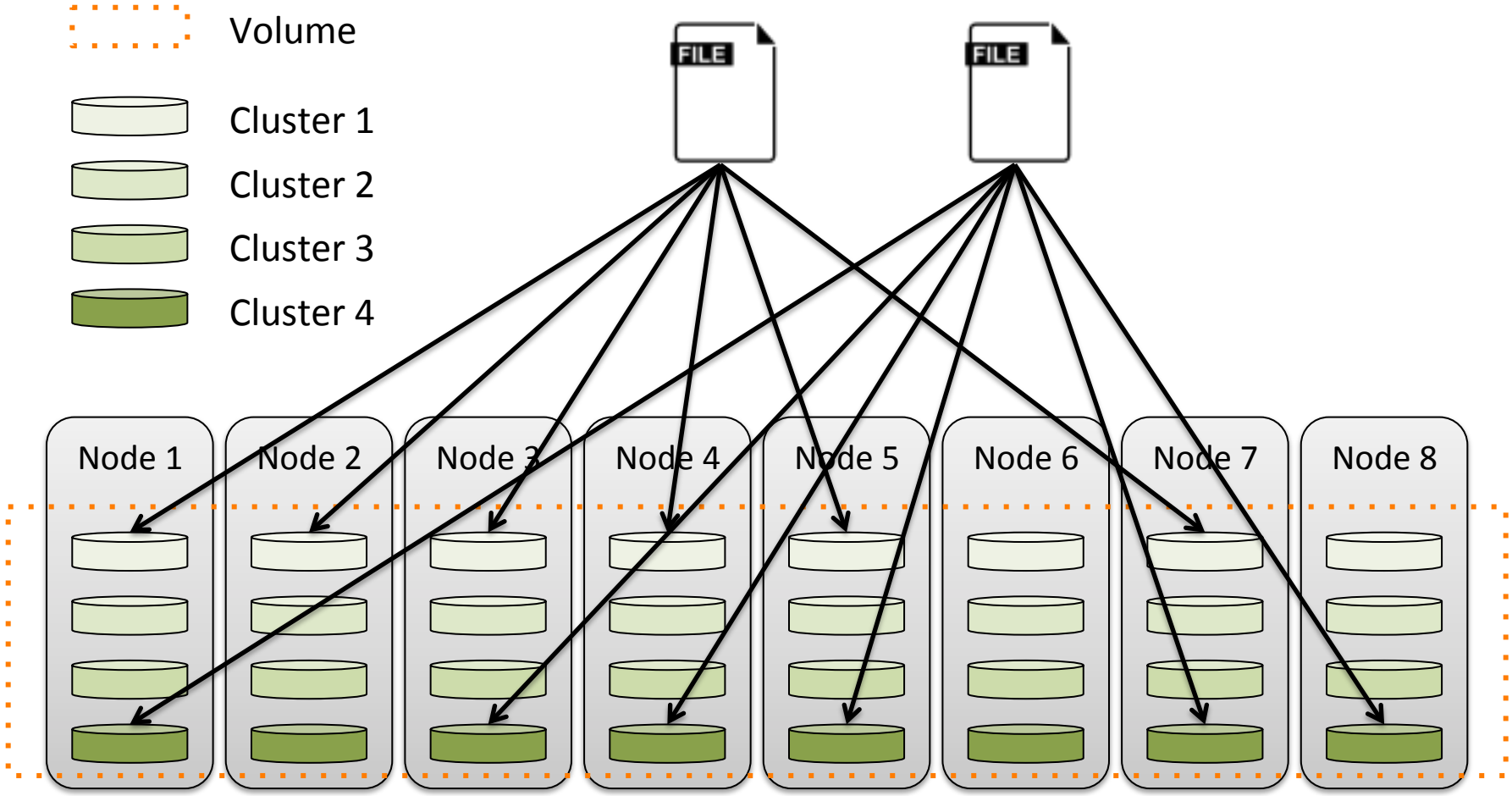
FEC et répartition



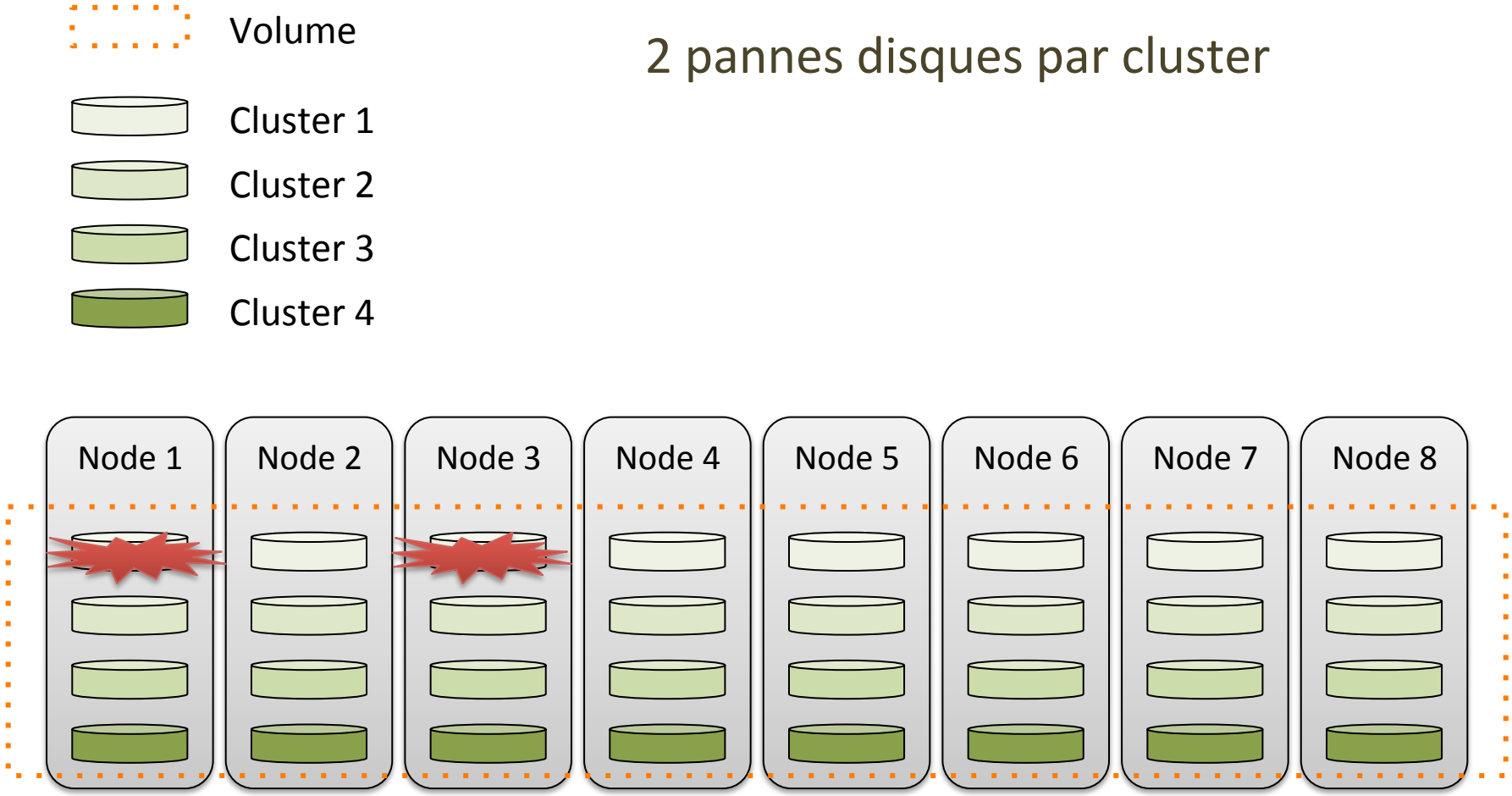
FEC et répartition (e.g 6, 4)



FEC et répartition (e.g 6, 4)



FEC et répartition (e.g 6, 4)



FEC et répartition (e.g 6, 4)

Volume

Cluster 1

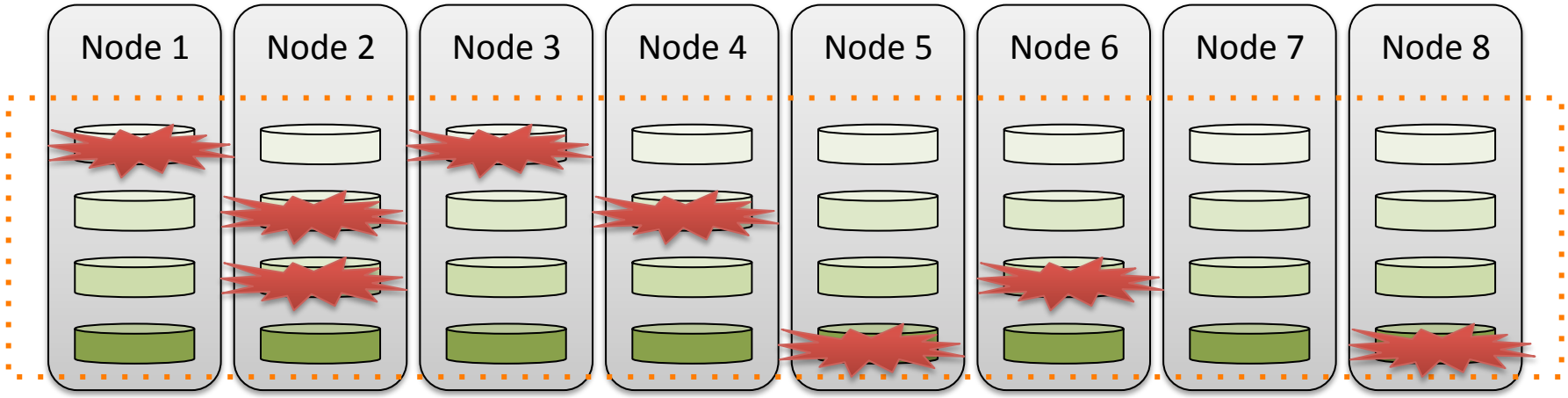
Cluster 2

Cluster 3

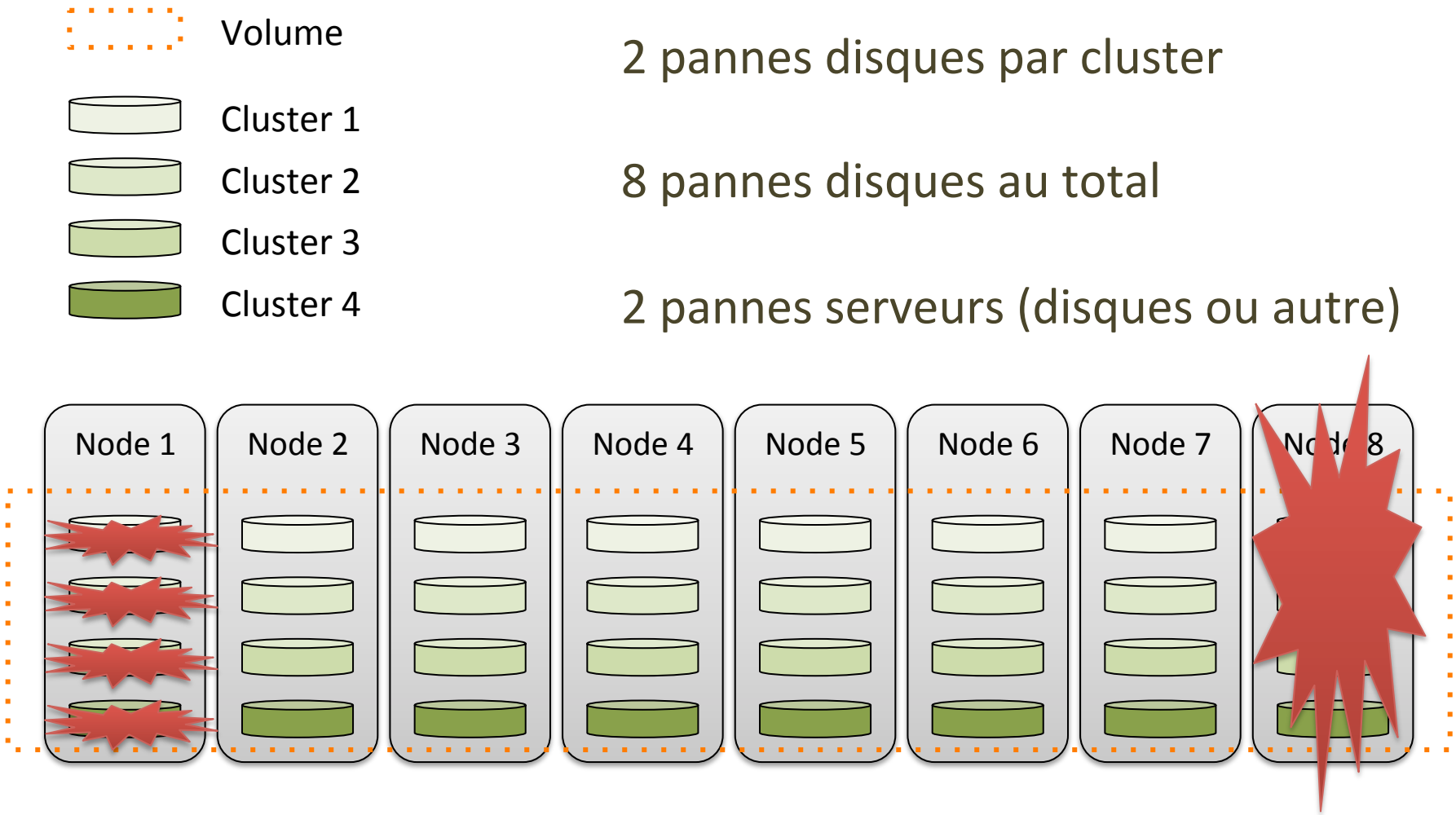
Cluster 4

2 pannes disques par cluster

8 pannes disques au total



FEC et répartition (e.g 6, 4)

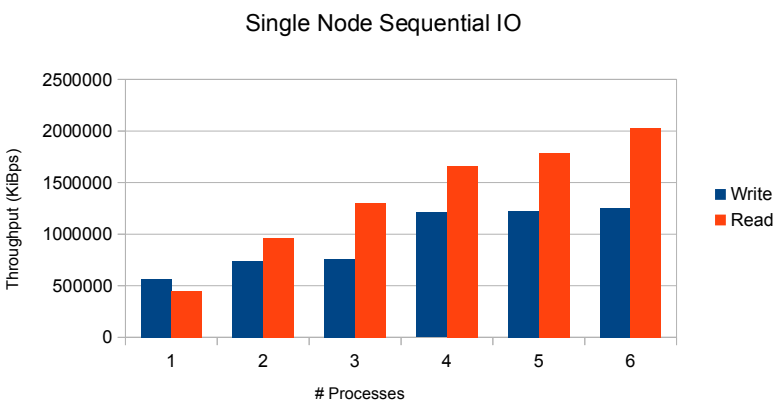


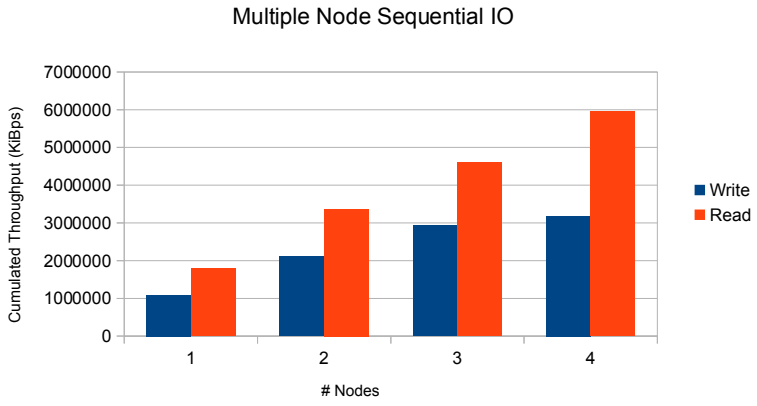
ROZOFS

Performances

Sequential I/O

Fichiers de 1 GiB - Accès séquentiels par blocs de 64 KiB



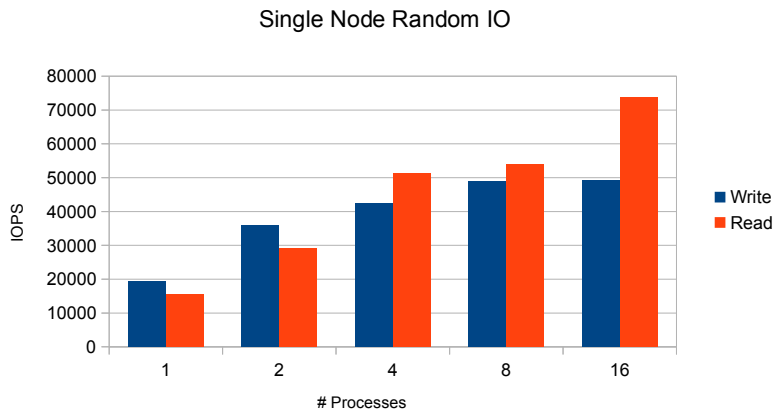


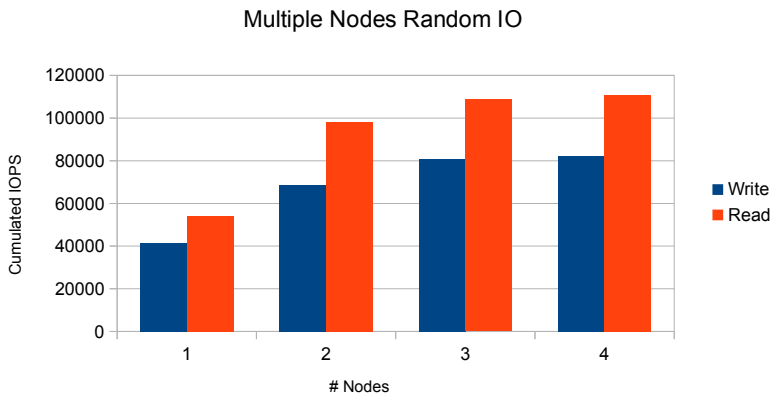
Write : 3 GiBps

Read : 6 GiBps

Random I/O

Fichiers de 100 MiB - Accès aléatoires par blocs de 4 KiB





Write : 82 000 IOPS

Read : 110 000 IOPS

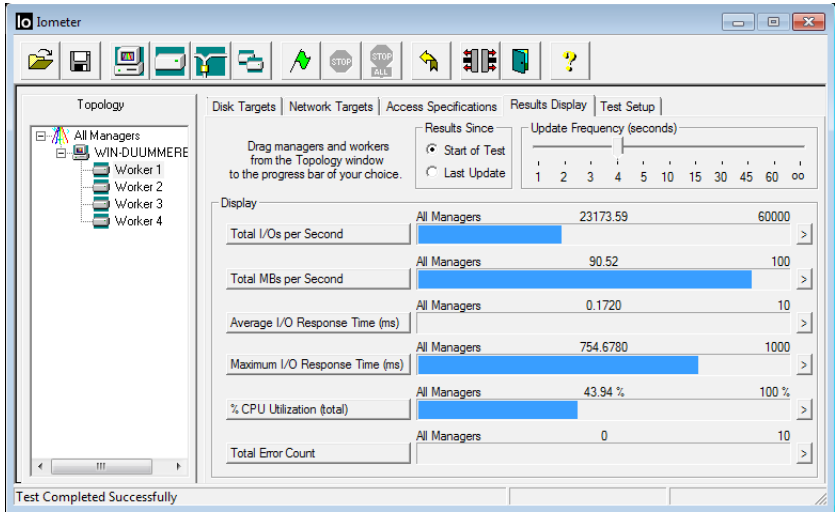
ROZOFS

Use case : virtualisation

Windows 7 / VMware WS 10

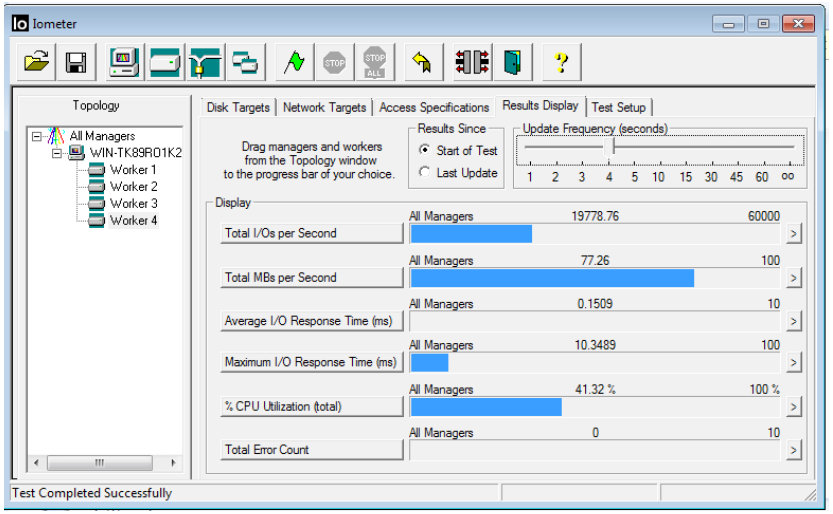
Accès aléatoires 4kiB

RozoFS*



231 000 IOPS

Ext4 local



197 000 IOPS

*4 nœuds – volume SAS 10K RPM

Projet CARMIN

BUSINESS CASE

Business Case

Projet CARMIN de Polytechnique, IHES, IHP, CMLS.

Cahier des charges:

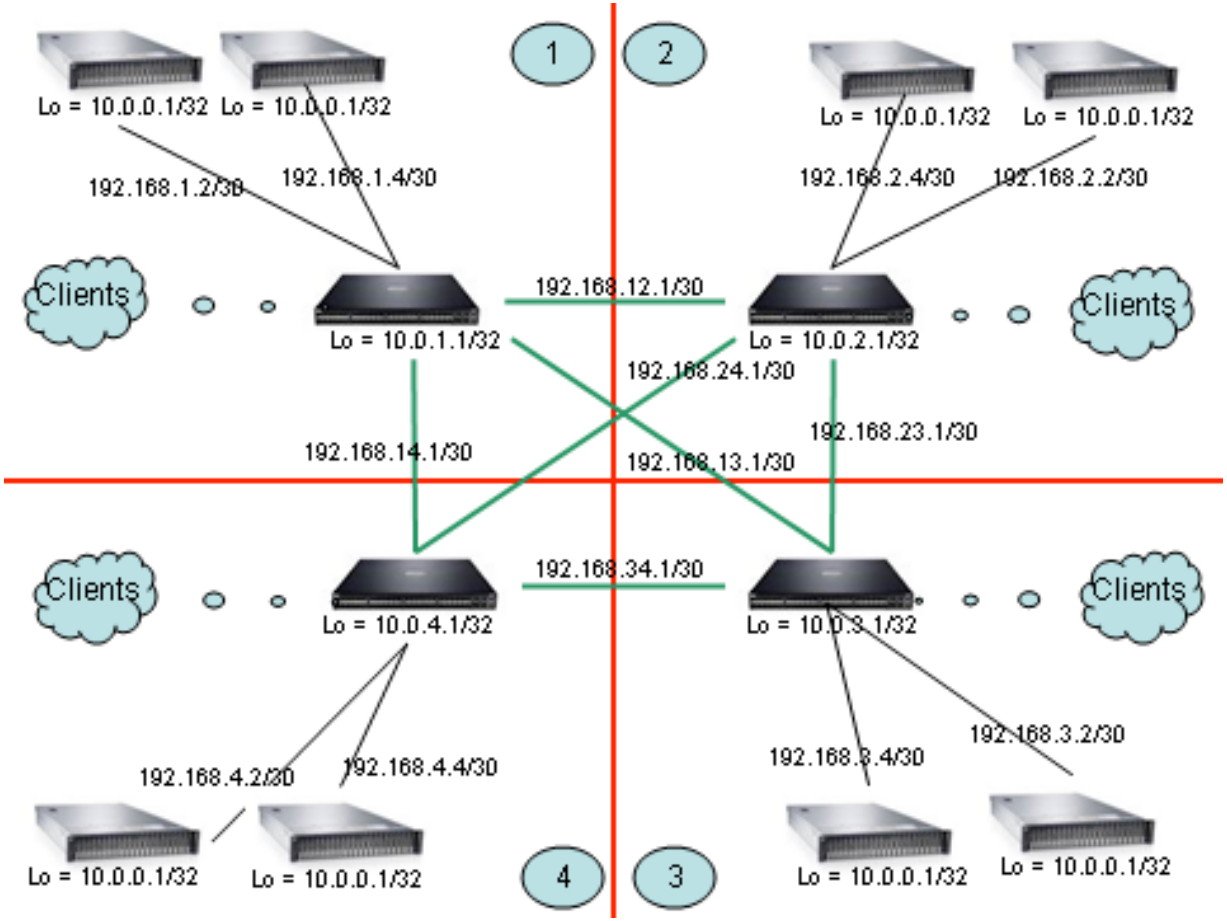
- Multi sites.
- Temps de reconstruction réduit.
- Fichiers entre 1 Go et 2 Go. (vidéo).
- Répartie 50 % écriture / 50 % lecture.
- Raccordement de type : 10 Gb/s et 40 Gb/s pour le backbone.
- Robuste avec une tolérance aux pannes, sans perte des données.

Business Case

Solution proposée :

- Sur 4 sites.
- 2 nœuds par site soit 8 serveurs de stockage.
- Temps de reconstruction d'un site < 24H (Vs 10 jours).
- Tolérance aux pannes, 2 serveurs ou 1 site. Sans interruption de service.
- Sans SPOF.
- Serveur Dell R720xd.

Schéma Trafic Model Projet CARMIN



Les points différenciant.

- Solution unique.
- IOPS Intensif , constant même en cas de panne,
- Solution compatible avec les technologies de stockage rapide,
- Hautement polyvalente (types et usages des données),
  - ✓ Virtualisation et Cloud Computing . Ready Openstack et Cloudstack.
  - ✓ Multimédia, Transactionnel, HPC.
  - ✓ Archivage, backup Big Data.

Les plus de la solution.

- Solution Open Source,
- Pas besoin de modifier les applications clientes,
- Tiering : Mix SATA/SAS/SSD clusters,
- Thin Provisionning : Système de fichiers avec quotas,
- Protocoles : Natif (RozoFS) / NFS / CIFS / AFP / HTTP / FTP / ISCSI
- Compatible : NAS, SAN, DAS, RAID.

RozoFS élimine les limitations du code à effacement avec ces ruptures technologiques et bouscule les standards du marché.

QUESTIONS ?

RozoFS

Erasure Code Based Scale-out NAS

Merci de votre attention

www.rozofs.com

https://github.com/rozofs/rozofs

Pierre Evenou
CEO
pierre.evenou@rozofs.com

Christophe de La Guerrande
Sales
cdlg@rozofs.com