#rozofs

Dimitri Pertin @denaitre

RozoFS: The Scalable Distributed File System based on Erasure Coding

available on https://github.com/rozofs/rozofs

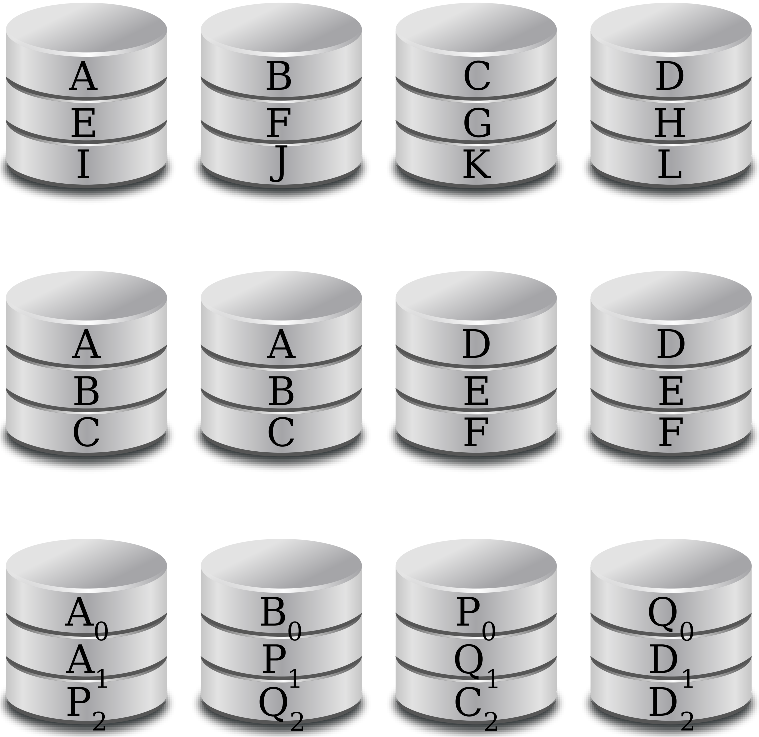
Distributed Storage Systems

Distributed Storage Systems

Goal: Improve storage protection and/or performance

RAID controllers for local data distribution over disks

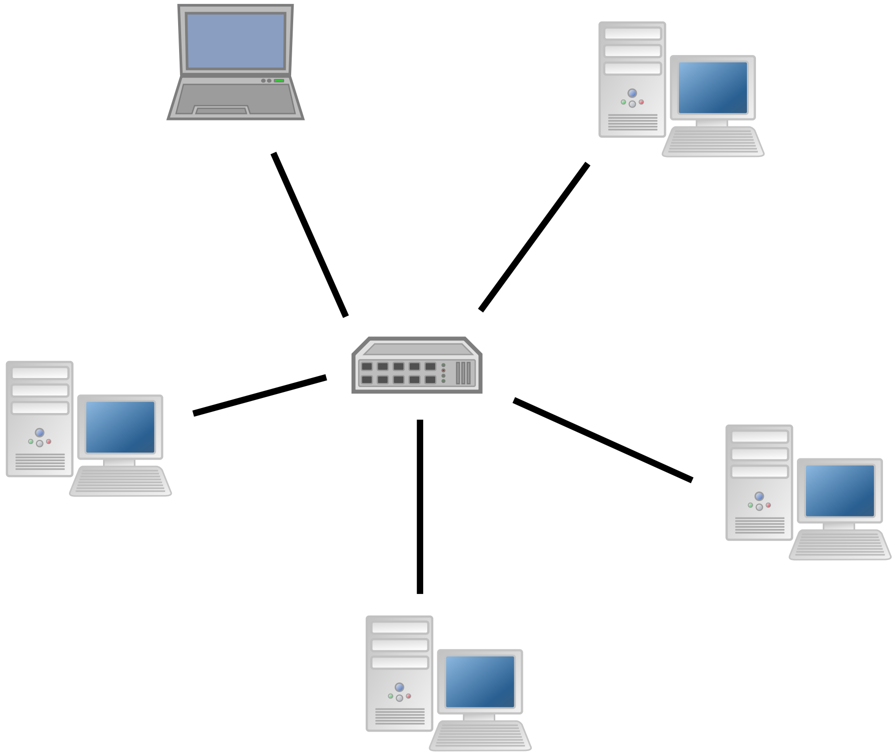
- **RAID-0** improve performance, no protection;
- **RAID-1** improve protection, bad performance;
- **RAID-6** trade-off between protection and performance.



Distributed Storage Systems

Distributed storage systems for network data distribution

New client node joins the storage network:



RozoFS File System

A Unique Namespace relying on several storage nodes

A POSIX Distributed File System can be simultaneously mounted by multiple clients and provides:

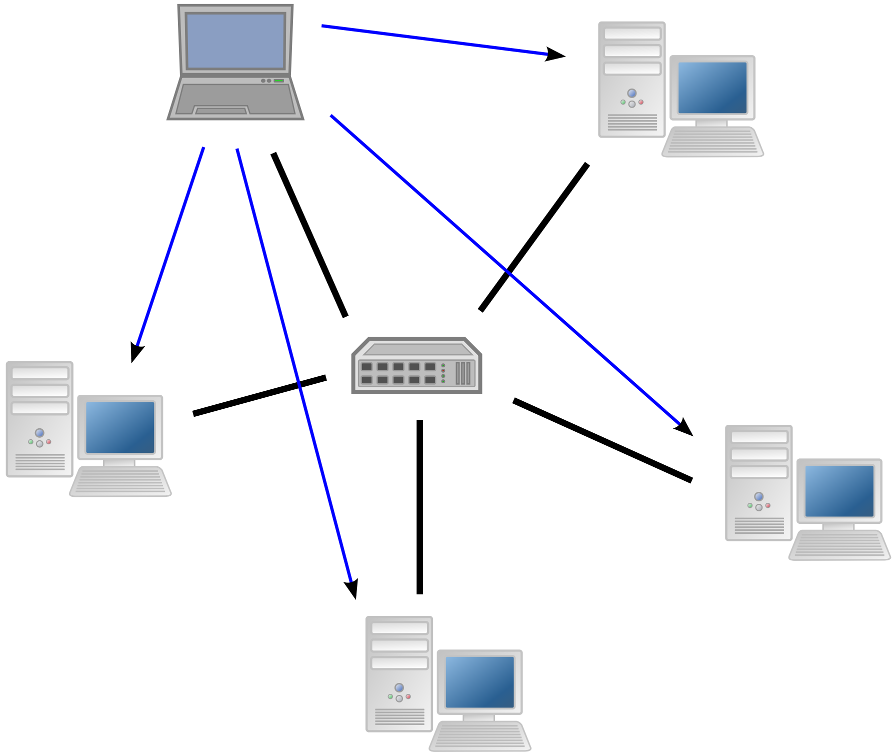
- Scalability;
- Flexibility and heterogeneity;
- Access/Location transparency;
- Data protection by an erasure code.

Fault Tolerance

Fault Tolerance

Distributed storage systems for network data distribution

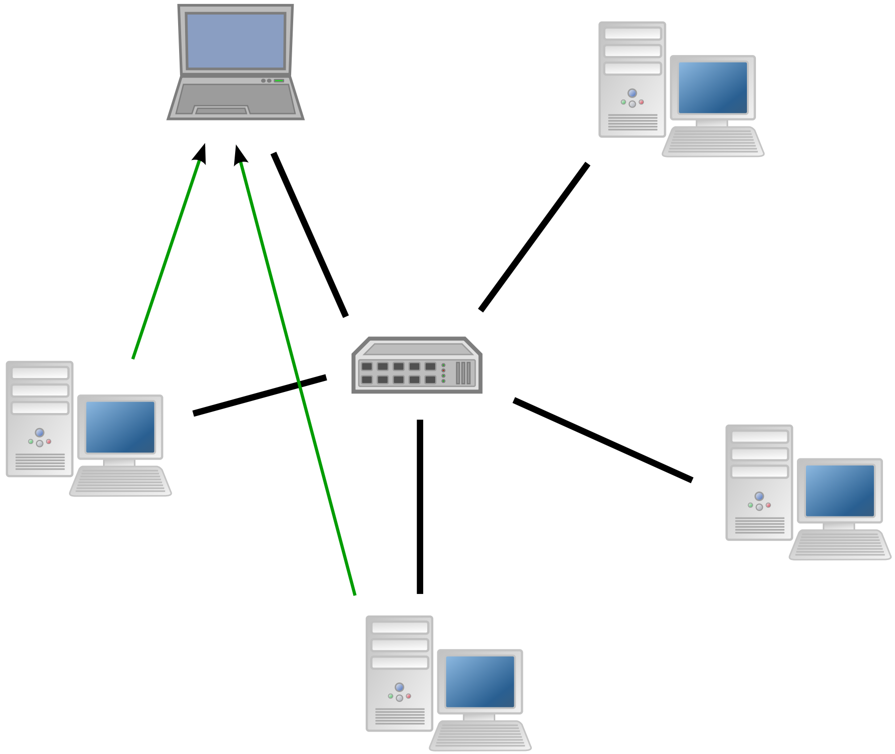
Write redundant information over nodes:



Fault Tolerance

Distributed storage systems for network data distribution

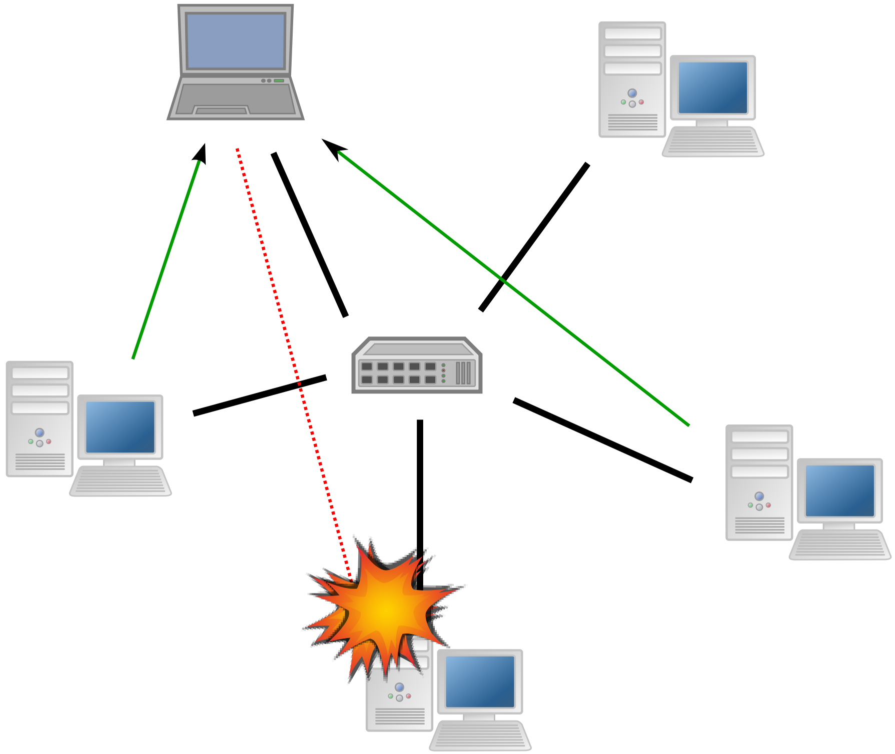
Read a subset is sufficient:



Fault Tolerance

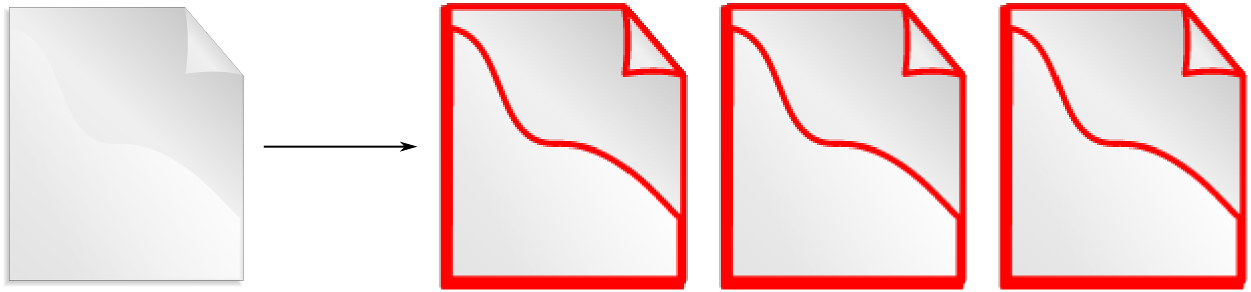
Distributed storage systems for network data distribution

Face node/link/matrix failures:



Fault Tolerance

Data Replication (3 copies)

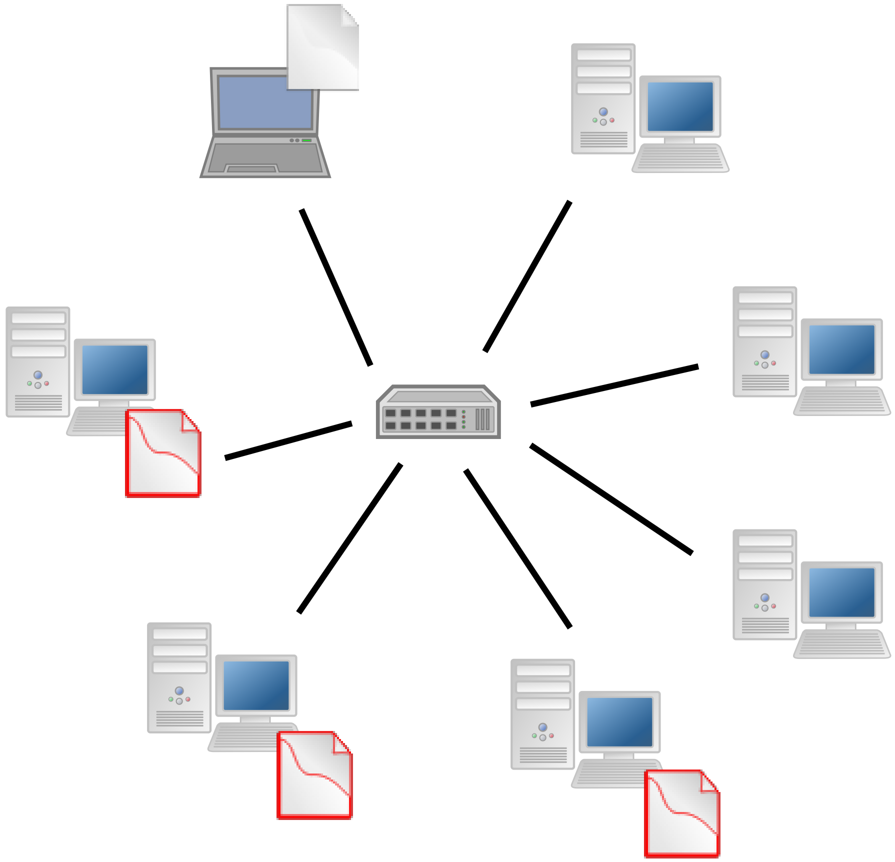


Remarks:

- Does not need any computation;
- But is very expensive;
- Three copies cost 3 times the original amount of information.

Fault Tolerance

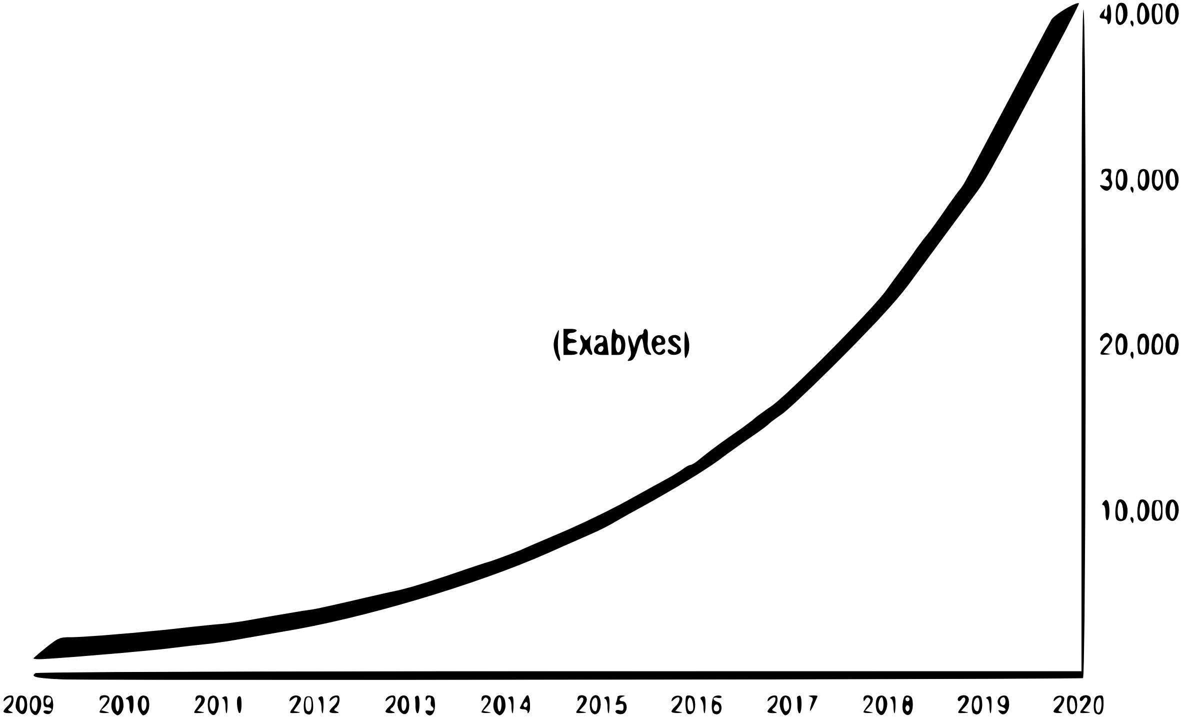
Data Replication (3 copies)



Problem ?

Distributed Storage Systems

What is the problem ?



The Digital Universe in 2020, J. Grantz and D. Reinsel (2012).

Distributed Storage Systems

What is the problem ?

Data protection plays a major role in storage consumption:

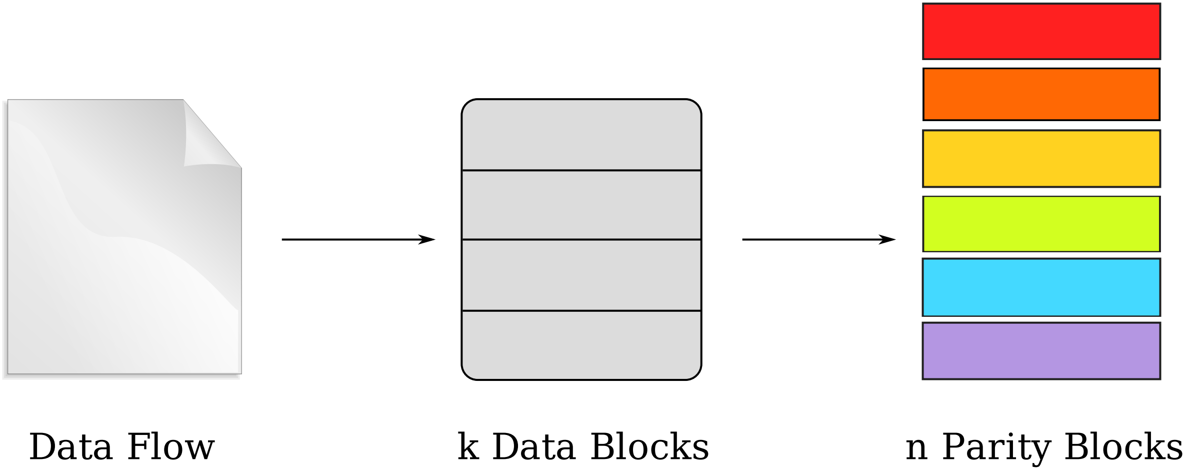
> *The amount of information indivuals create themselves - writing documents, taking pictures, downloading music, etc. - is* ***far less than the amount of information being created about them*** *in the digital universe.*
>
> *The proportion of data in the digital universe that requires protection is* ***growing faster than the digital itself****, from less than a third in 2010 to more than 40% in 2020.*

The Digital Universe in 2020, J. Grantz and D. Reinsel (2012).

Erasure Coding

Data Protection by Erasure Coding

(6,4) Erasure Encoding

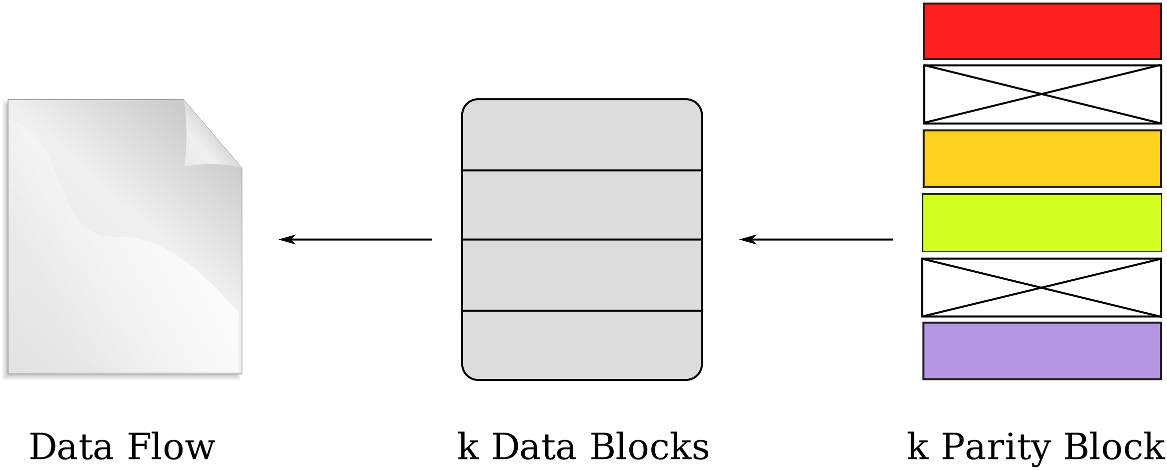


Remarks

- Optimal (MDS) codes decode from any subset of k parity blocks out of n;
- The system can face $n - k = 2$ failures;
- The storage overhead is $\frac{n}{k} = 1.5$

Data Protection by Erasure Coding

(6,4) Erasure Decoding

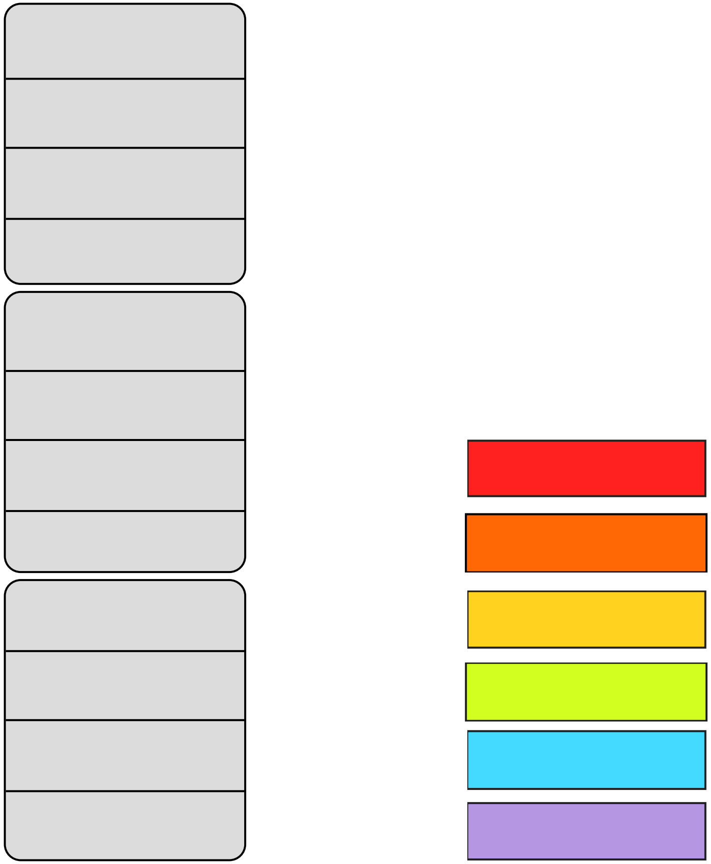


Remarks

- Optimal (MDS) codes decode from any subset of k parity blocks out of n;
- The system can face $n - k = 2$ failures;
- The storage overhead is $\frac{n}{k} = 1.5$

Data Protection by Erasure Coding

Comparison ?



Data Replication by 3

(6,4) Erasure Code

The Mojette Transform

The Mojette Transform

Presentation

- The Mojette Transform is a linear operation based on discrete geometry;
- Computes redundant information from user's data;
- The algorithm relies only on additions.

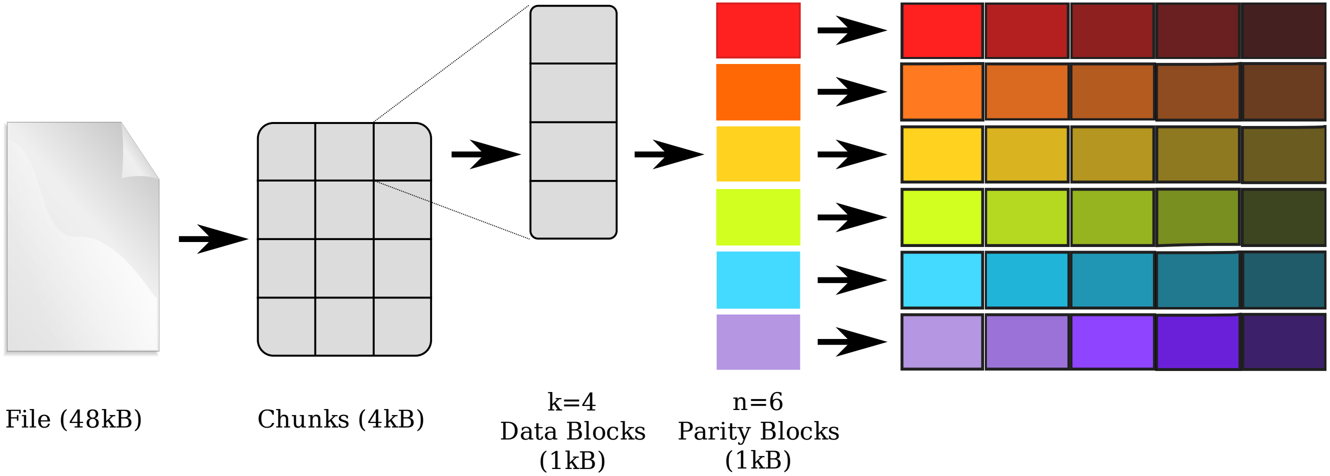
Performances

- Implementation uses fast XOR;
- Encoding and decoding computations are transparent.

The Mojette Transform, Theory and Applications, J. Guédon (2009).

The Mojette Transform

Protection in Storage Systems



- The MT is applied on 4 data blocks to produce a set of 6 parity blocks;
- Parity blocks are distributed over storage nodes;
- Any subset of $k = 4$ parity blocks out of the $n = 6$ is sufficient to decode.

Architecture of RozoFS

Architecture of RozoFS

Metadata Server: exportd service

Stores metadata (data about user data)

- POSIX information (e.g. size, permissions, timestamps, etc.)
- RozoFS related information (e.g. data localisation)

Knows the position of data blocks

- answers data location in reading
- answers where to store projections in writing

Architecture of RozoFS

Storage Servers: storaged daemon

Hold a storaged daemon that manages

- data storing
- data retrieval
- data accessibility

Data can be stored on:

- local file system (ext4, xfs, etc.) or remote Amazon bucket
- native or other protocol (CIFS, AFP, etc.)

Architecture of RozoFS

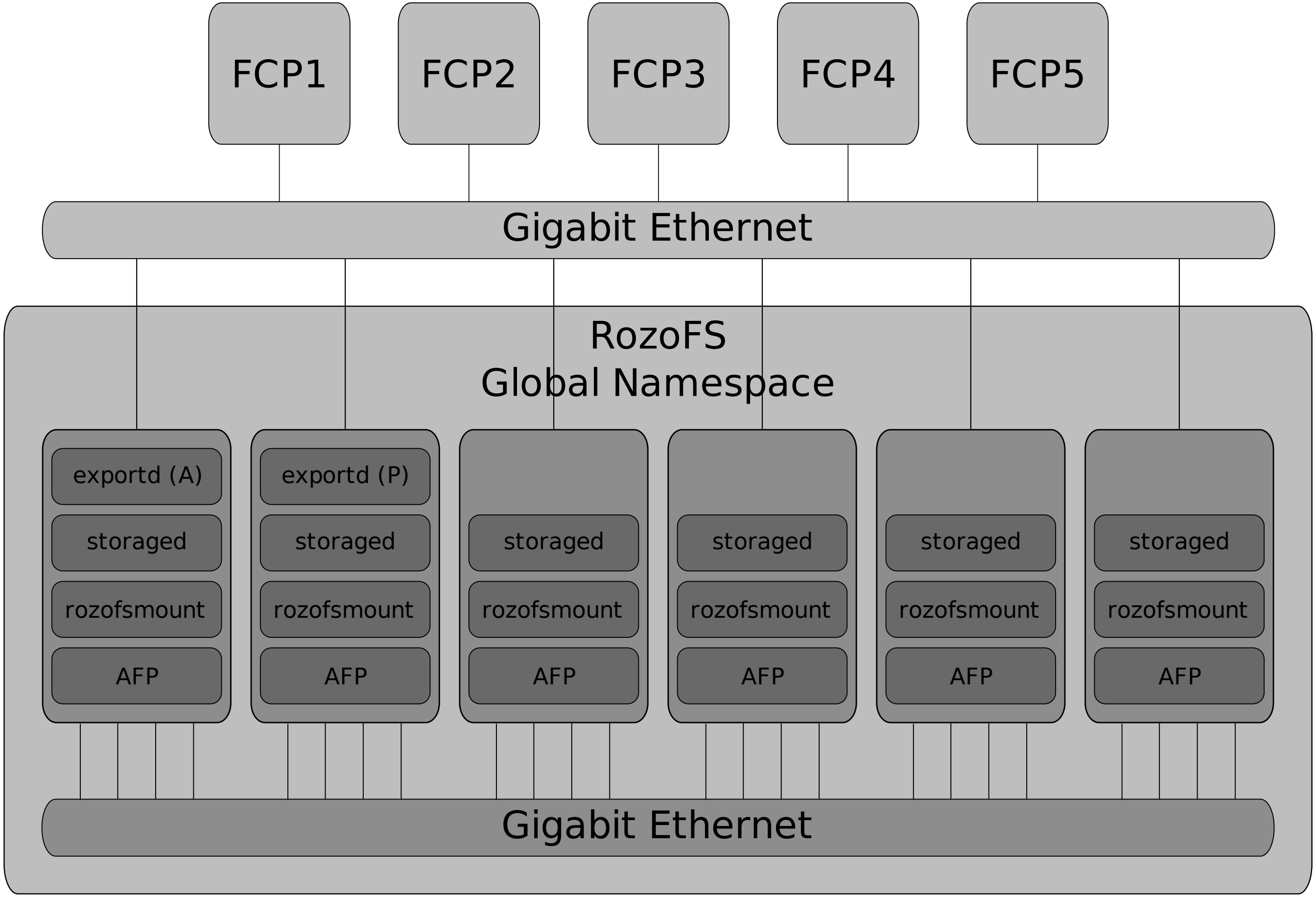
Clients

Rely on FUSE (rozofsmount)

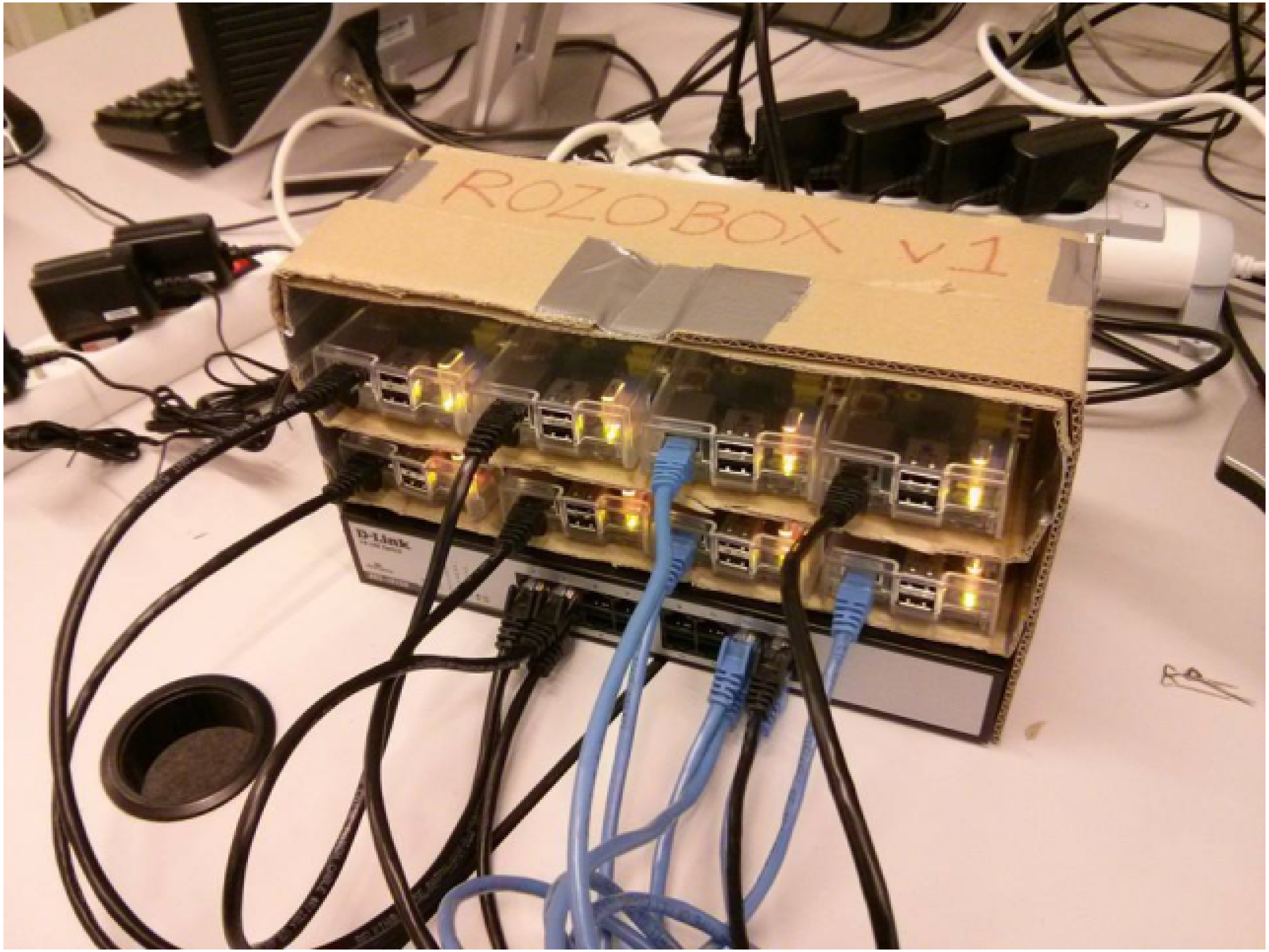
- mounts locally RozoFS
- translates transparently user actions for the network system

Manage encoding (write) and decoding (read)

Production Use Example



Academic Use Example



Thanks!

Contribute:

https://github.com/rozofs/rozofs

Contact me at:

@denaitre or dimitri.pertin@univ-nantes.fr

Have a look at

ANR FEC4Cloud project

Slideshow created by remark.