The Mojette Erasure Code:

Application to fault tolerant Distributed File System (DFS)

Architecture de codes correcteurs d'erreurs

Journée inter GDR ISIS et SoCSiP

4 Novembre 2014, salle B007, Télécom Bretagne

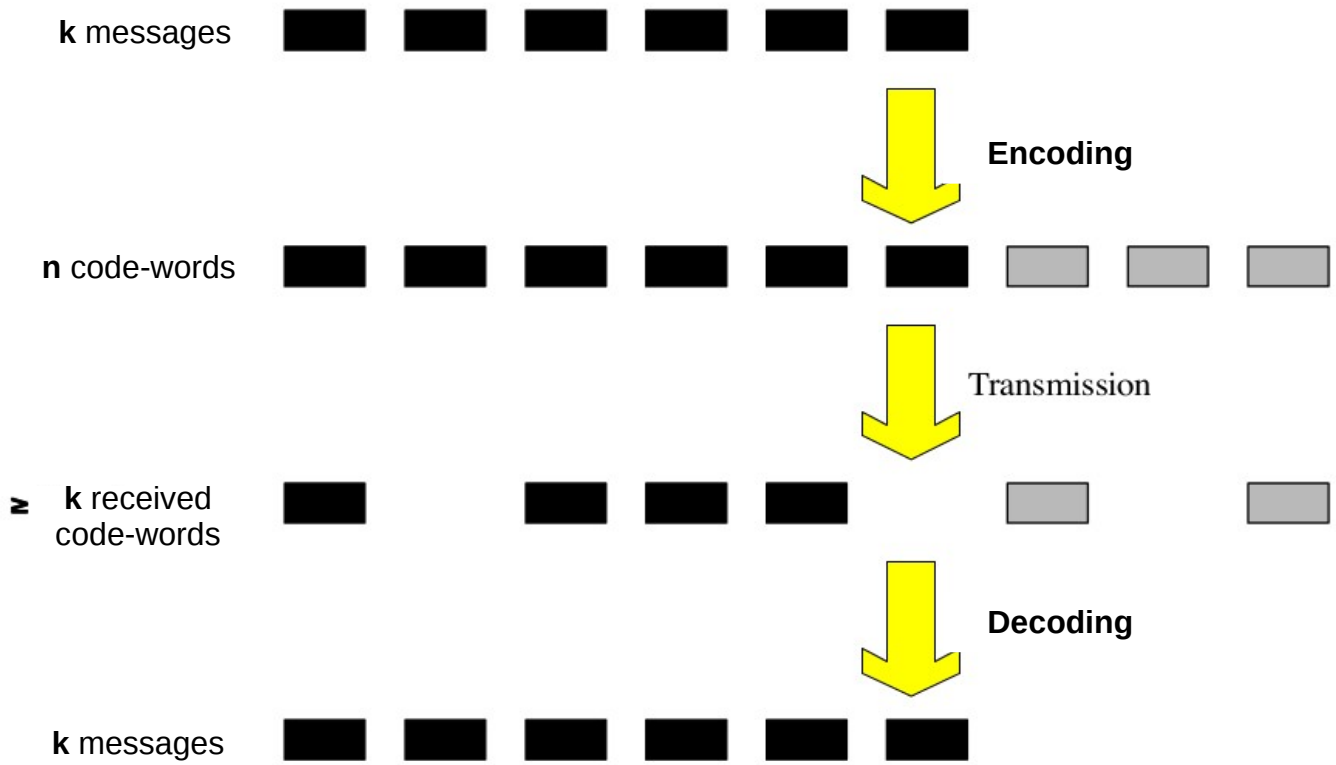
Benoît Parrein, Université de Nantes, IRCCyN Lab, UMR 6597

Joint work with FIZIANS SAS

Outline

- MDS erasure codes
- FEC4Cloud project
- Mojette erasure code
- Performances
- Application to DFS: RozoFS

Erasure codes (MDS property)

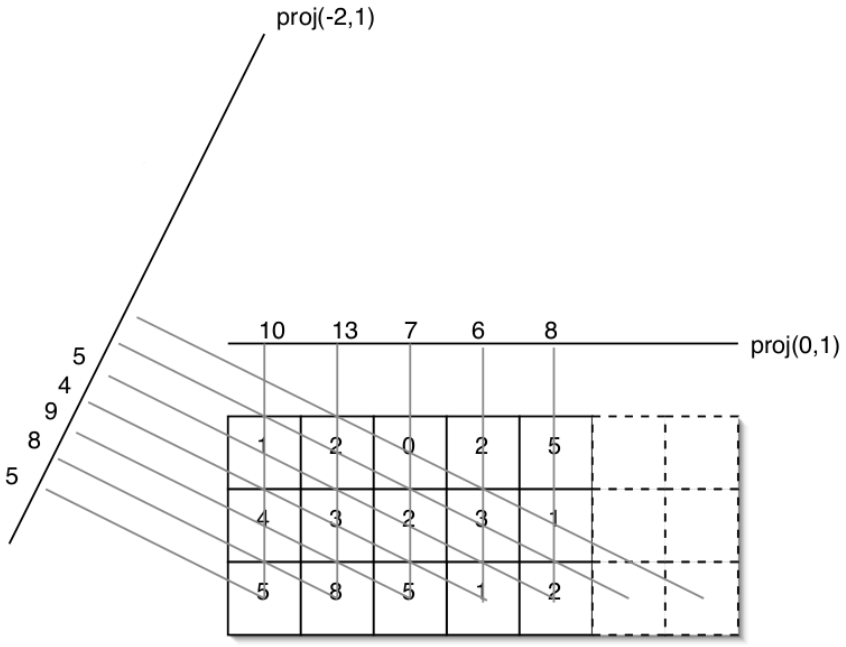


FEC4Cloud Project

- ANR 2012 (appel Emergence)
- Partners: IRCCyN (lead), ISAE, SATT-Ouest Valorisation
- Budget: 256 K€
- Duration: 24 months (**product** oriented)
- Goal: promoting erasure codes within Cloud storage infrastructure

The forward Mojette transform

- based on Radon transform [Guédon, 1995]
- compute 1D projections from a 2D geometrical buffer



Conditions of reconstruction

- Myron Katz criteria (1978)

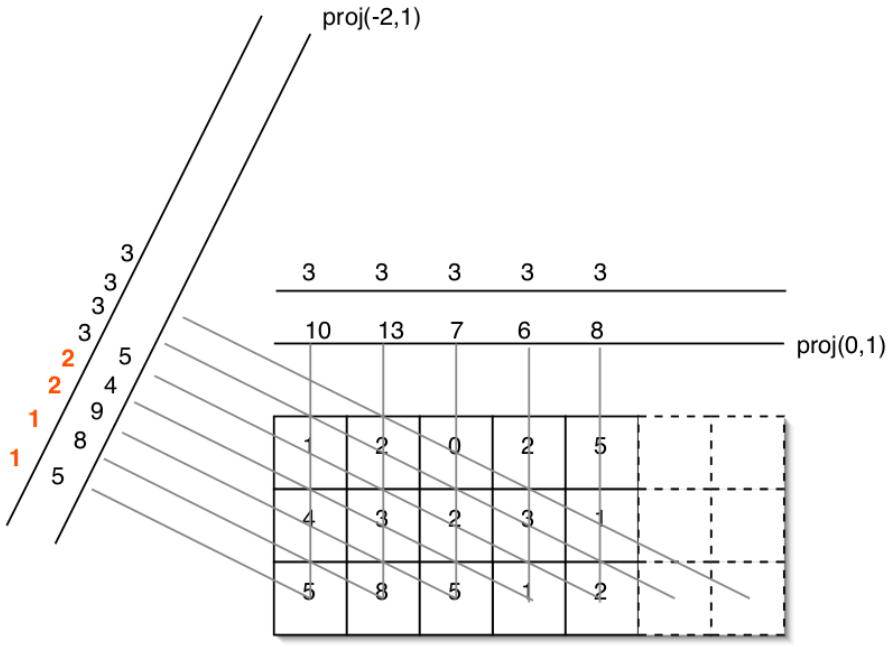
$$\sum_{i}^{N} p_i \geqslant P \quad \text{or} \quad \sum_{i}^{N} q_i \geqslant Q$$

With a rectangular geometrical buffer of PxQ pixels
And a projections set S = {(p_i, q_i)}.

- Mathematical Morphology for non rectangular shape [Normand, 1997]

Ancillary data for inversion

- Number of pixels (ixels) contributing to one bin
- Sum of coordinates (not represented here)



The reverse Mojette transform

- Check Katz Criteria (or mathematical morphology if necessary)
- While 2D geometrical buffer is not completely reconstructed do
  - Find one-to-one correspondence into the projection set
  - Retroprojection at the right location
  - Update the projections (bins and ancillary data)

Properties of Mojette erasure code

- (1+ϵ) MDS
- Systematic and non systematic coding
- Asynchronous reconstruction
- No algebraic constraints (as Galois fields)
- No prime size (as in MDS array or FRT)
- Linear complexity in coding/decoding [O(IN)]
- Soft coding and decoding

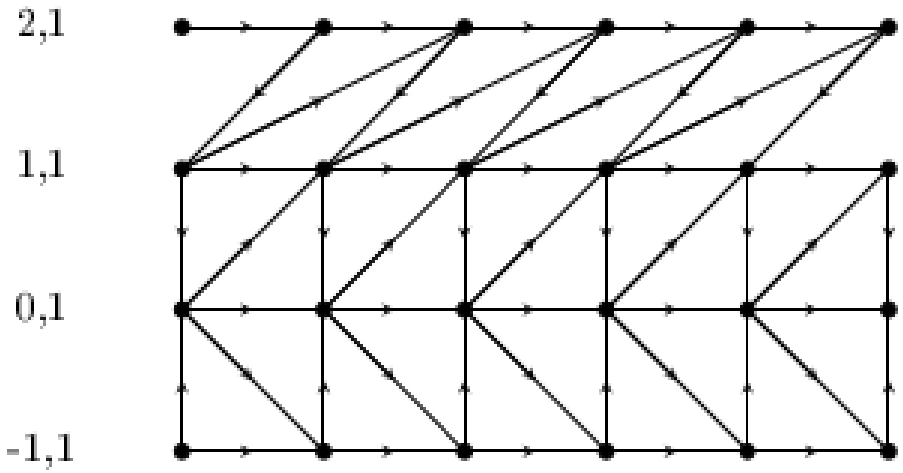
The reverse Mojette transform

- Check Katz Criteria (or mathematical morphology if necessary)
- While 2D geometrical buffer is not completely reconstructed do
  - **Find one-to-one correspondence into the projection set - costly**
  - Retroprojection at the right location
  - **Update the projections (bins and ancillary data) - costly**

Optimizations (1/2)

- Deterministic path of reconstruction [Normand, 2006]

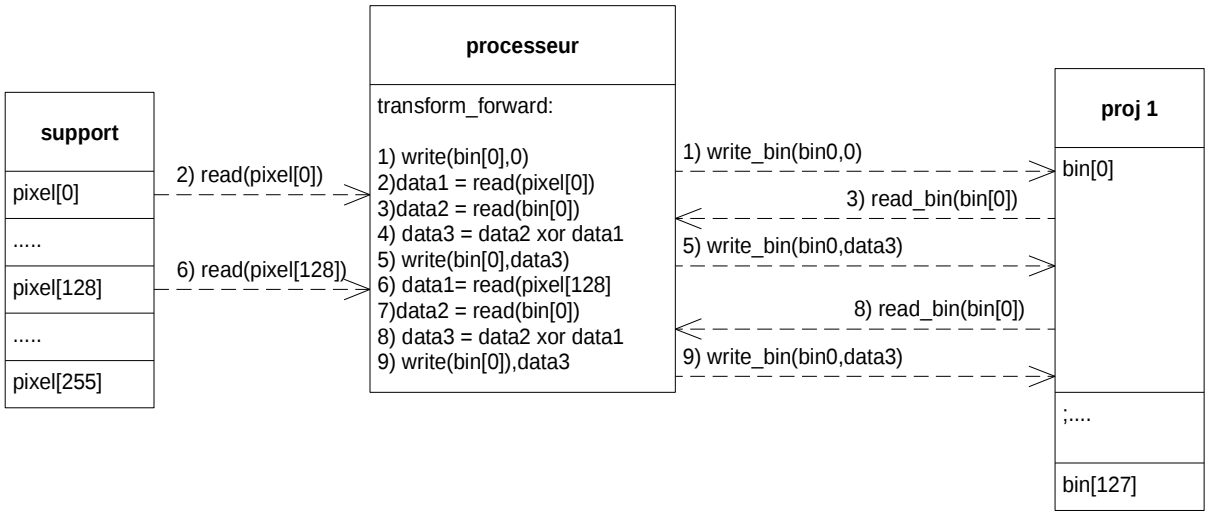
  (if geometrical buffer appears as a stripe)



Example on a 4 lines geometrical buffer
with 4 projections

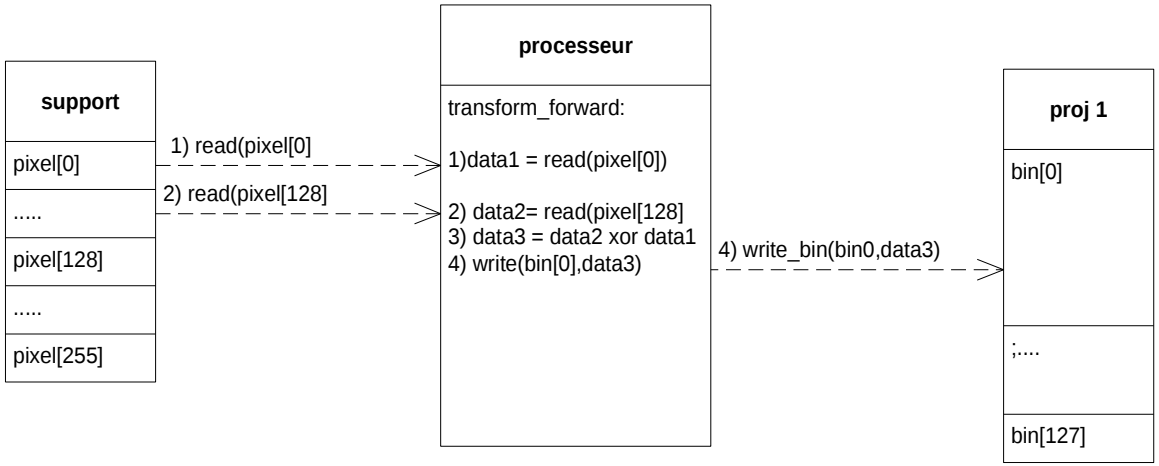
- +drastic reduction in writes [engineers of Fizians, 2013]

Optimizations (2/2)



Classical forward mojette transform

Optimizations (2/2)

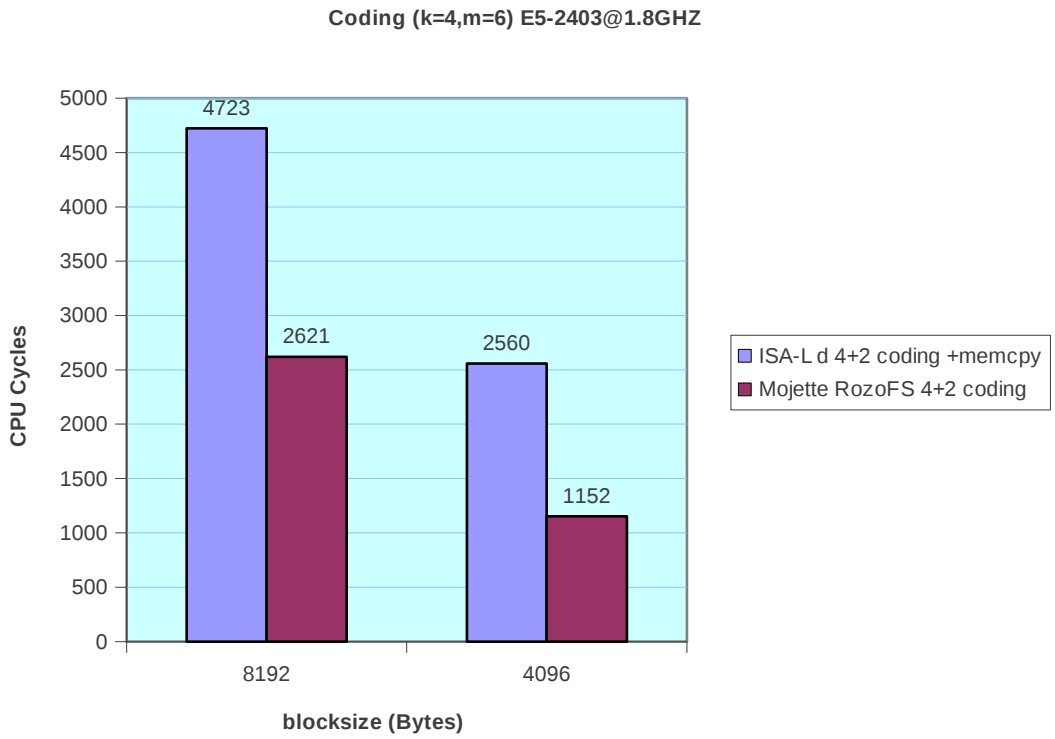


Optimized forward mojette transform [Féron et al., 2014]

Related works (software)

- Reed-Solomon (by Cauchy matrices [Byers, 1995])
- Reed-Solomon (by Vandermonde matrices [Rizzo, 1998] now a RFC5510)
- Cauchy “Good” [Planck, 2008] in Jerasure 1.2
- Intel ISA-L (includes SSE instructions)
- …

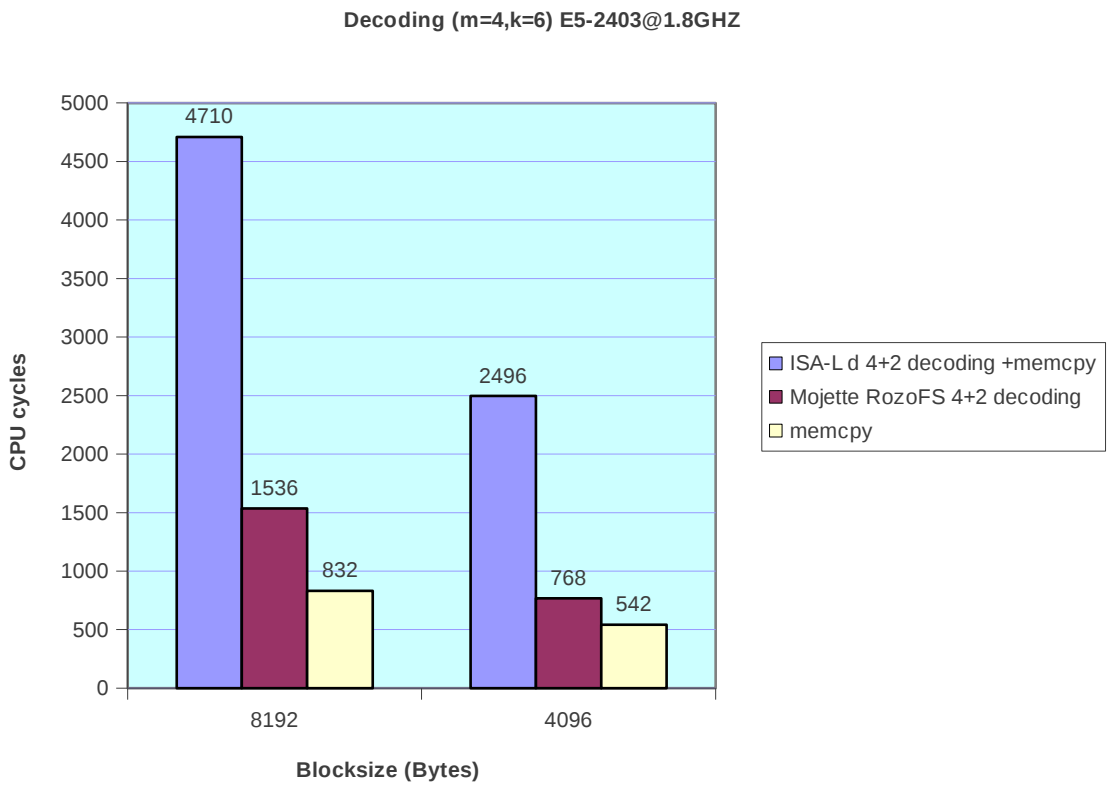
Performances (coding)



that means 5.625 GB/s (resp. 6.40 GB/s with 4KB) for Mojette coding (in purple) and 3.122 GB/s (resp. 2.88 GB/s with 4KB) for RS coding (in blue)

x1.8 (resp. x2.22) faster (for a 3x more coding blocks)

Performances (decoding)



that means 9.6 GB/s (resp. 9.60 GB/s with 4KB) for Mojette coding (in purple) and 3.130 GB/s (resp. 2.953 GB/s with 4KB) for RS coding (in blue)

x3 (resp. x3.25) faster (for a 2x more coding blocks)

UNIVERSITÉ DE NANTES

IRCCyN

Fizians

The Mojette Erasure Code:
Application to fault tolerant Distributed File System (DFS)

Architecture de codes correcteurs d'erreurs

Journée inter GDR ISIS et SoCSiP

4 Novembre 2014, salle B007, Télécom Bretagne

Benoît Parrein, Université de Nantes, IRCCyN Lab, UMR 6597

Joint work with FIZIANS SAS

High availability means...

- 99.999999...% reachable
- Copies and copies and copies... (up to 7 times)
- Hard disks and hard disks and hard disks...
- High consumption of energy
- Privacy problems

  - Erasure codes reduce drastically the size of the infrastructure for the same availability rate (2x) and facilitate privacy policy

Distributed File Systems

- HDFS (Hadoop)
- Facebook file system (f4)…not really I/O centric
- CephFS, GlusterFS,...
- Scality (based on Chord)
- …

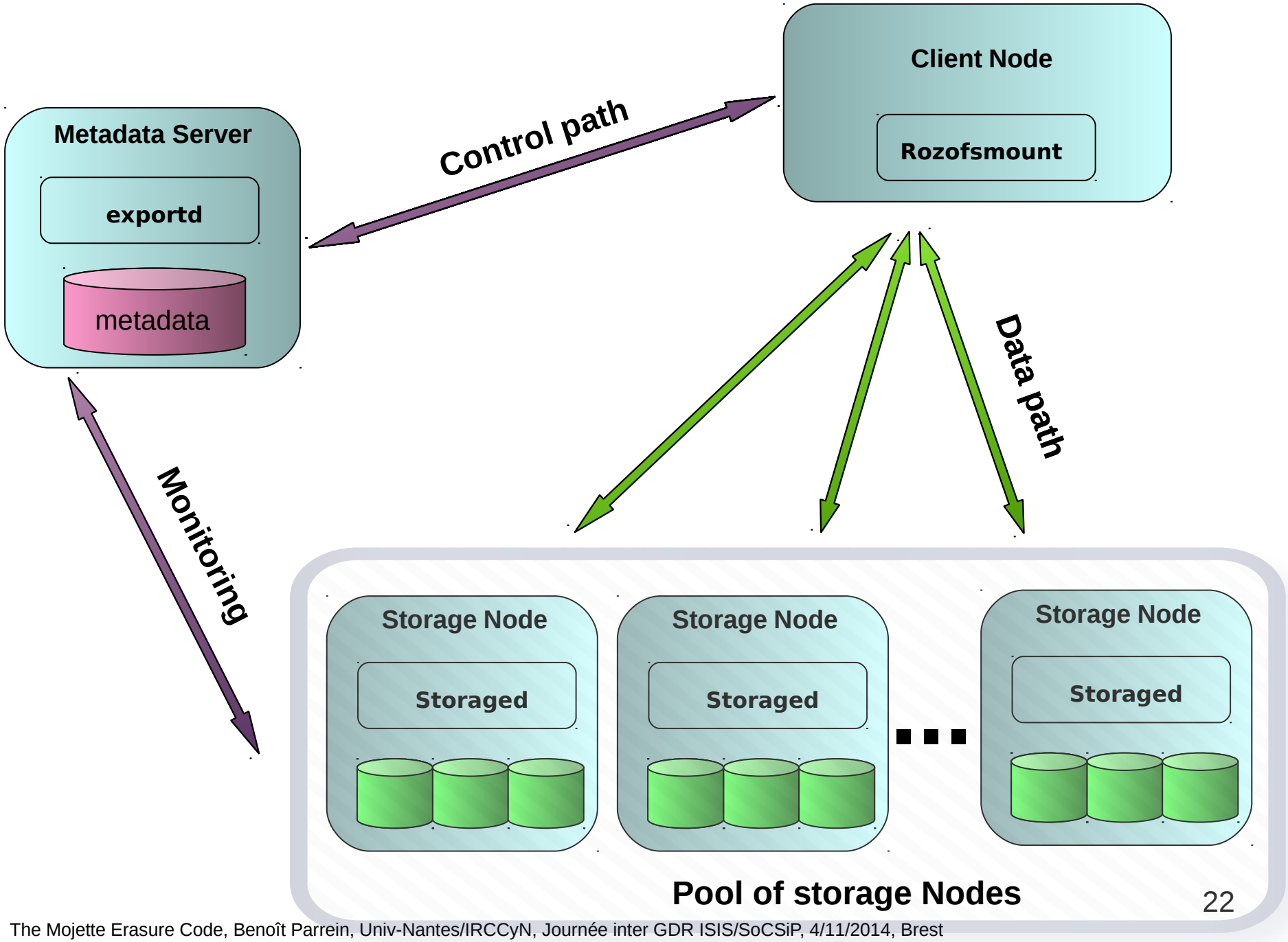
Mix of replicas (hot data) and erasure coding (cold data) :

- none use erasure codes **always**

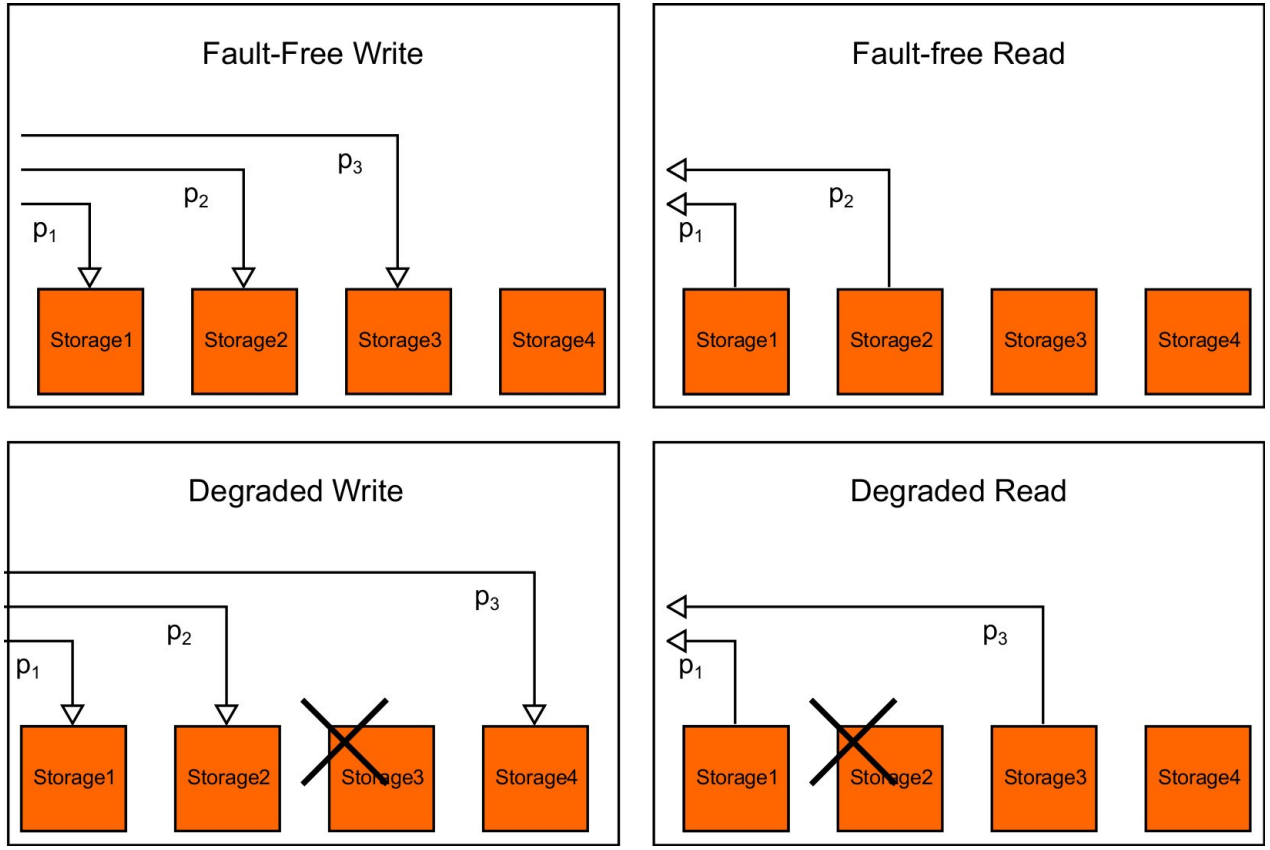

ROZOBOX v1

RozoFS

- I/O Centric Distributed File System
  - POSIX Scale-out storage
  - Commodity hardware
  - Fault tolerance (up to 4 failures)
  - Based on erasure coding (Mojette coding)
  - Dedicated to cold and hot data
- Open source project

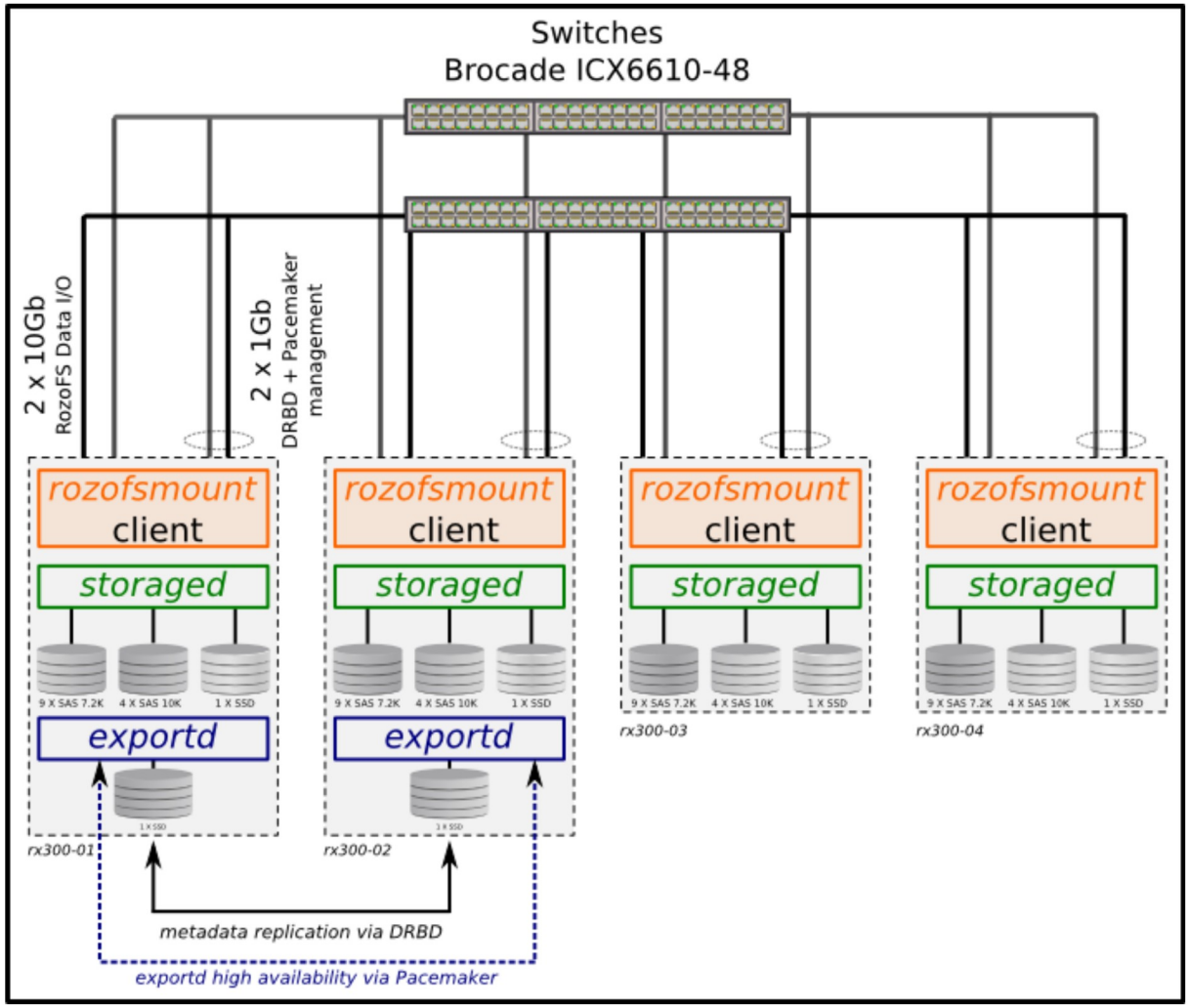

Client Node
Rozofsmount
Metadata Server
exportd
metadata
Control path
Data path
Monitoring
Storage Node
Storaged
Storage Node
Storaged
Storage Node
Storaged
Pool of storage Nodes

Read/Write function (in non-sys coding)



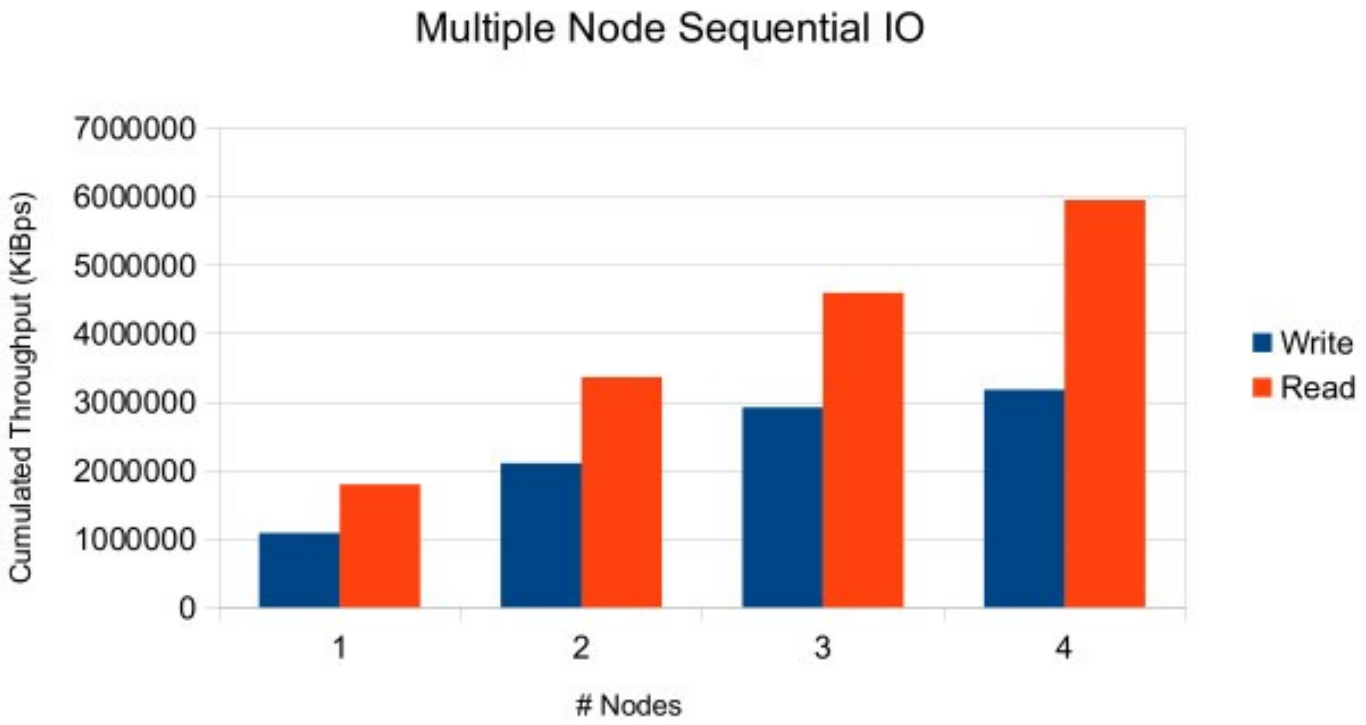
In a layout 0 i.e (2, 3) coding i.e two projections are necessary for reconstruction

Testbed


Switches
Brocade ICX6610-48
2 x 10Gb
RozoFS Data I/O
2 x 1Gb
DRBD + Pacemaker
management
rozofsmount
client
storaged
9 X SAS 7.2K
4 X SAS 10K
1 X SSD
exportd
rx300-01
rx300-02
rx300-03
rx300-04
metadata replication via DRBD
exportd high availability via Pacemaker

Performances

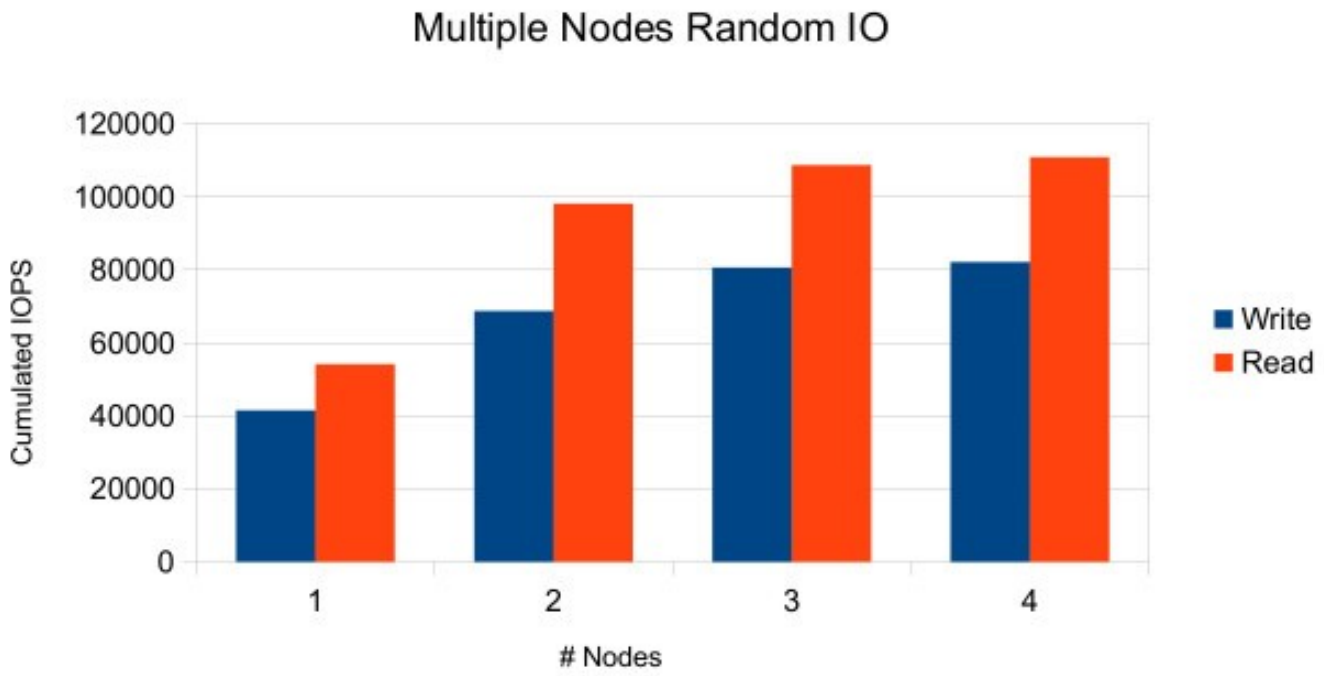
- Sequential access: layout 0, 4K blocks



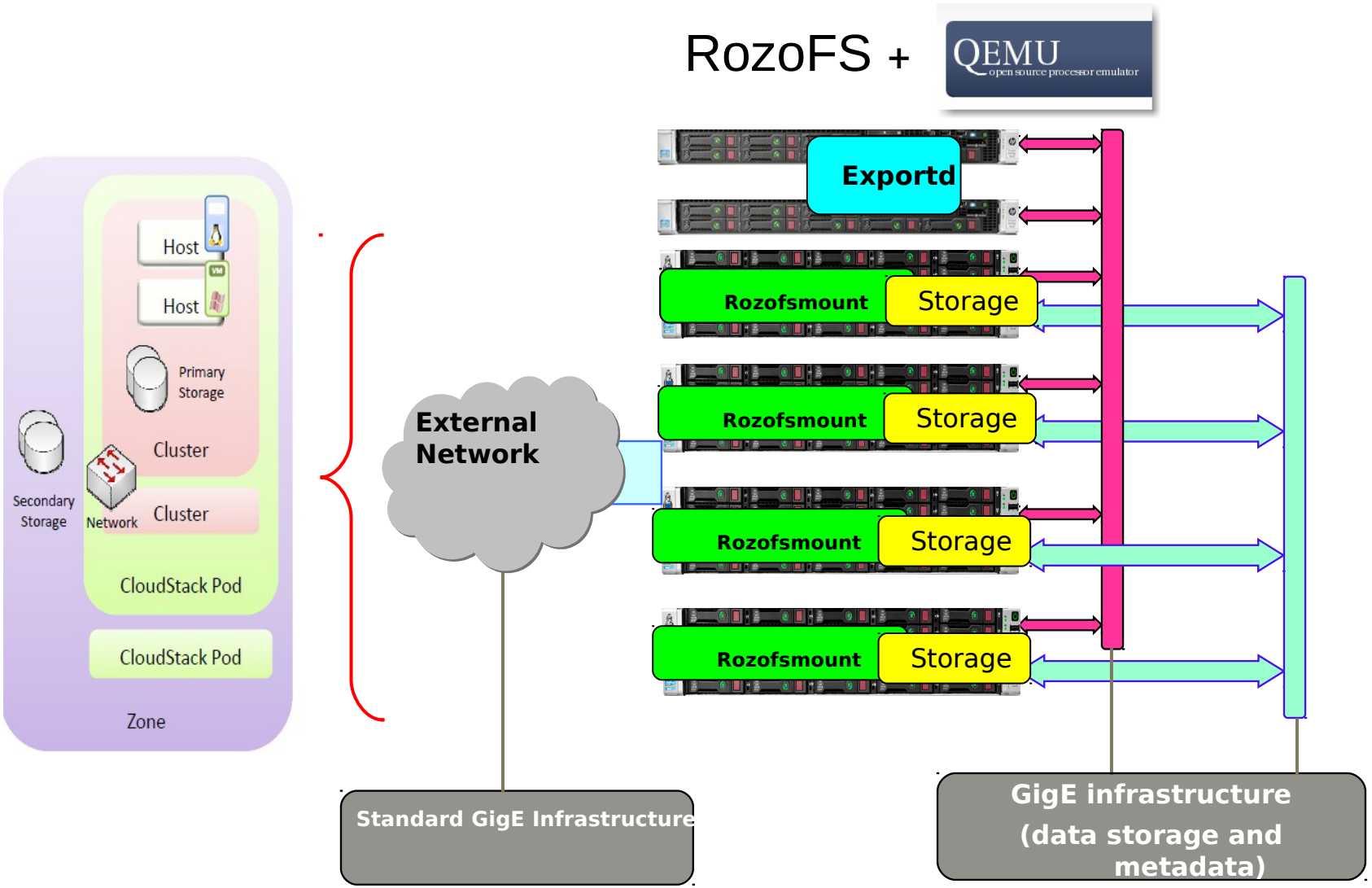
...6 GB/s in read
...3 GB/s in write

Performances

- Random access: layout 0, 4K blocks



...100K IOPS in read
...80K IOPS in write


RozoFS +
QEMU
open source processor emulator
Exportd
Rozofsmount
Storage
Rozofsmount
Storage
Rozofsmount
Storage
Rozofsmount
Storage
External Network
Host
Host
Primary Storage
Cluster
Secondary Storage
Network
Cluster
CloudStack Pod
CloudStack Pod
Zone
Standard GigE Infrastructure
GigE infrastructure (data storage and metadata)

Credits

Pierre Evenou

Jeanpierre Guédon

RozoFS

https://github.com/rozofs

Sylvain David

Alex Van Kempen

FEC4CLOUD

Quentin Lebourgeois

Jean-Pierre Monchanin

Didier Féron

Louis Legouriellec

UNIVERSITÉ DE NANTES

IRCCyN

UMR 6597

OUEST VALORISATION

Ressources d'innovation

Dimitri Pertin

Nicolas Normand

Christophe de la Guérrande

Bastien Confais

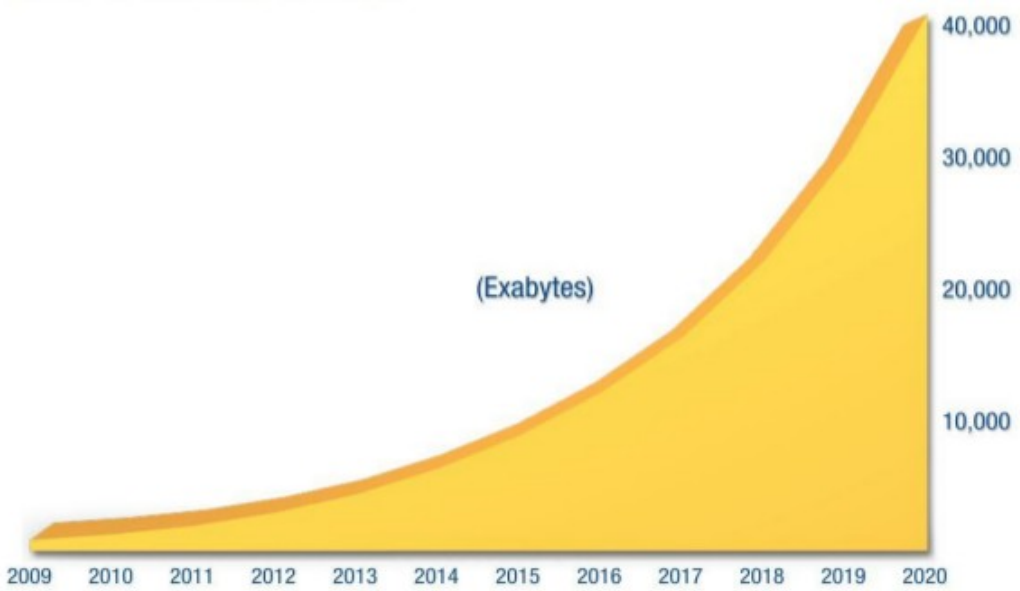
Olivier Blin

Vielen Danke!

Backup slides

The storage in the world

- 40 Exabytes (10^{18} bytes) stored in 2020
- 15 EB (37%) in the Cloud(s)
- 7,5 EB (50%) video, images, ...



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Server type Fujistu RX300-S8 (R3008S0035FR)
CPU model name 2 x Intel Xeon CPU E5-2650 v2 @ 2.60GHz (8 cores & 16 threads/core)
Memory (GB) 64 GB
RAID card RAID Controller SAS 6Gbit/s 1GB (D3116C)
Virtual DRIVE 0 - Seagate Constellaton.2, SAS 6Gb/s, 1TB, 2.5", 7200 RPM (ST91000640SS)
- 11 drives
- RAID 5
Virtual DRIVE 1 - Seagate Pulsar.2, SAS 6Gb/s, 100GB, 2.5", MLC (ST100FM0002)
- 1 drive
- RAID 0
Virtual DRIVE 2 - WD Xe, SAS 6Gb/s, 900GB, 2.5", 10000 RPM (WD9001BKHG)
- 4 drives
- RAID 0
Ethernet
controllers - Intel 82599EB 10-Gigabit SFI/SFP+ - 2*10Gb
- Intel I350 Gigabit Network - 2*1Gb
- Intel I350 Gigabit Network - 4*1Gb


ROZOBOX

Conclusions

- RozoFS is an I/O centric distributed file system based on a erasure code (always)
- Performances: 100K IOPS, throughput of 6 Gbps...
- RozoFS follows up the infrastructure
- Apps: on line video editing, virtualisation (QEMU), database...
- participate to the convergence of cold and hot data
- Next: privacy (to check), grid5000 experiments (to come), deduplication (to attach)