Ceph中国社区

IT大咖说

CEPHALOCON APAC 2018

THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

Email Storage

with

Ceph

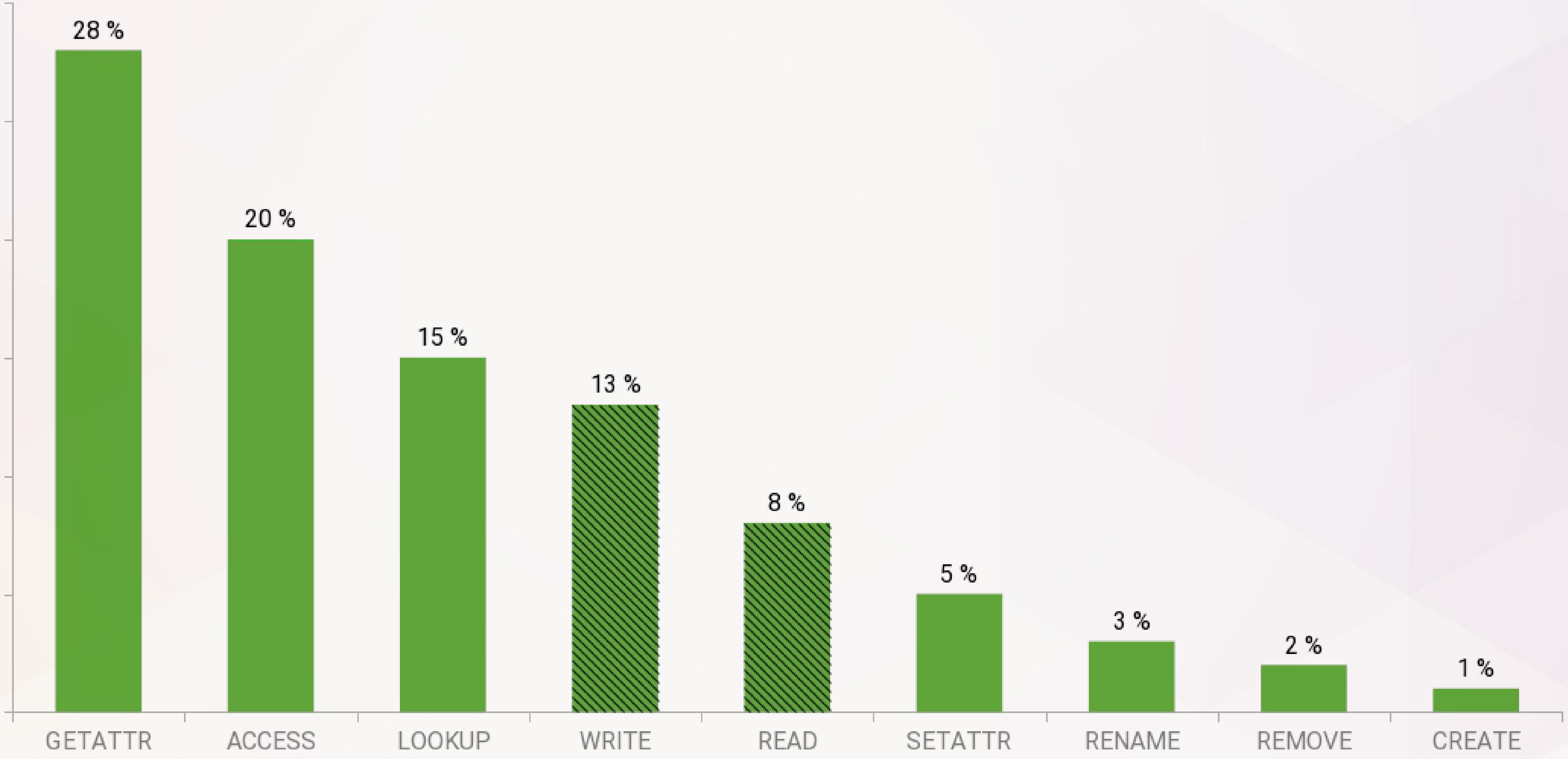
Danny Al-Gaaf
Deutsche Telekom AG

TelekomMail platform

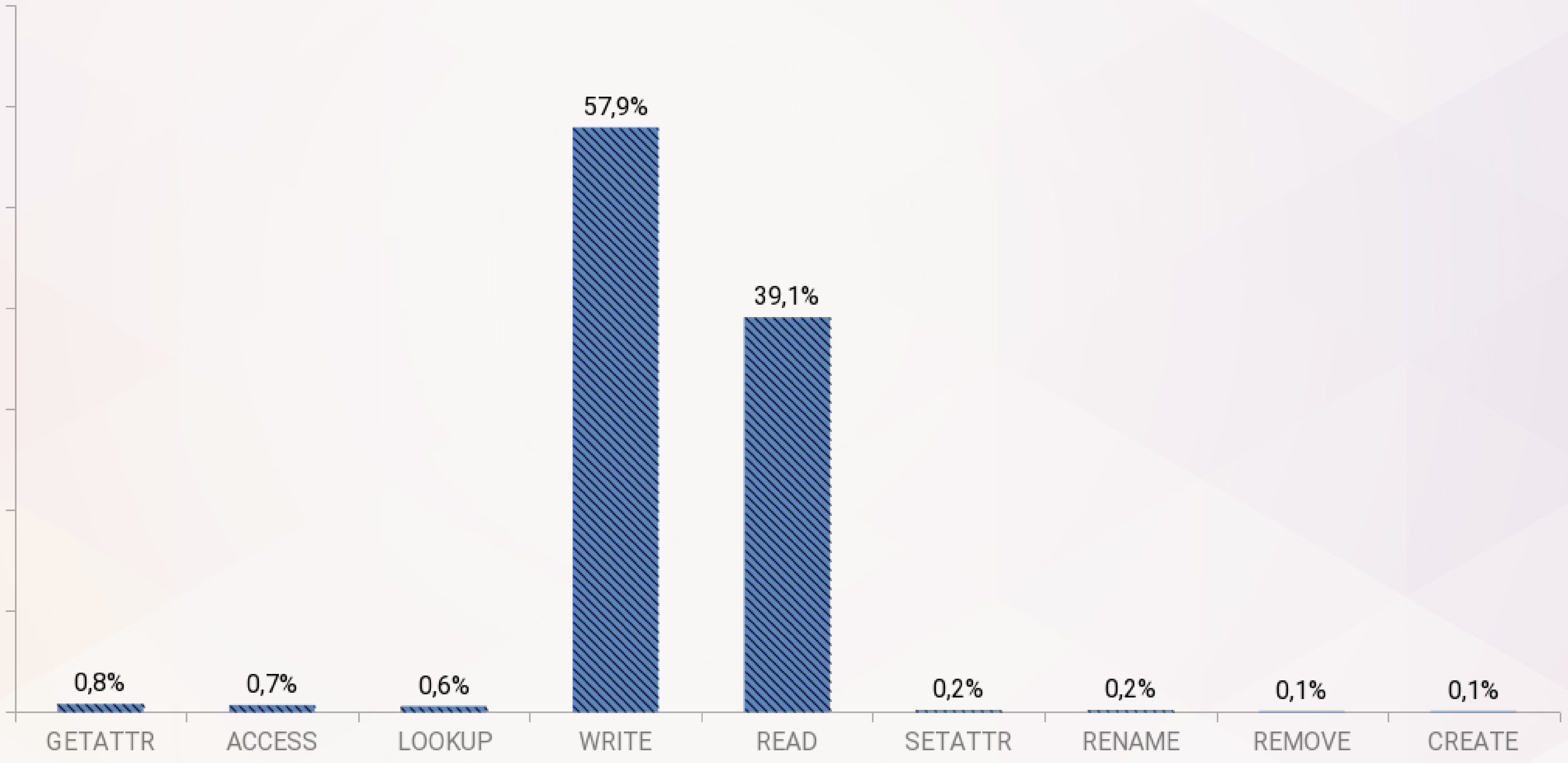
TelekomMail

- DT's mail platform for customers
- dovecot
- Network-Attached Storage (NAS)
- NFS (sharded)
- ~39 million accounts
- ~1.3 petabyte net storage
  - ~6.7 billion emails
  - ~1.2 billion index/cache/metadata files
  - ~42% usable raw space

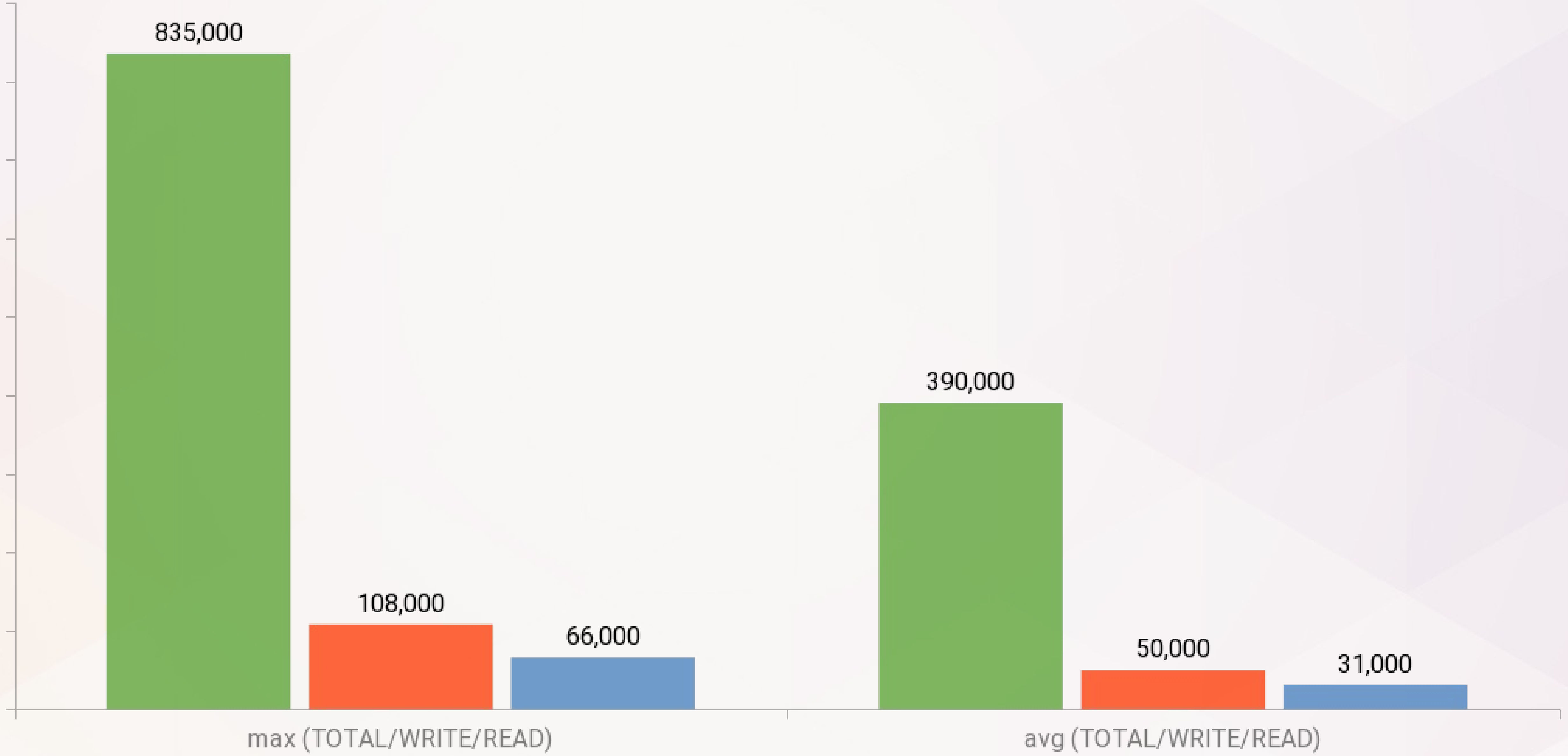
NFS Operations



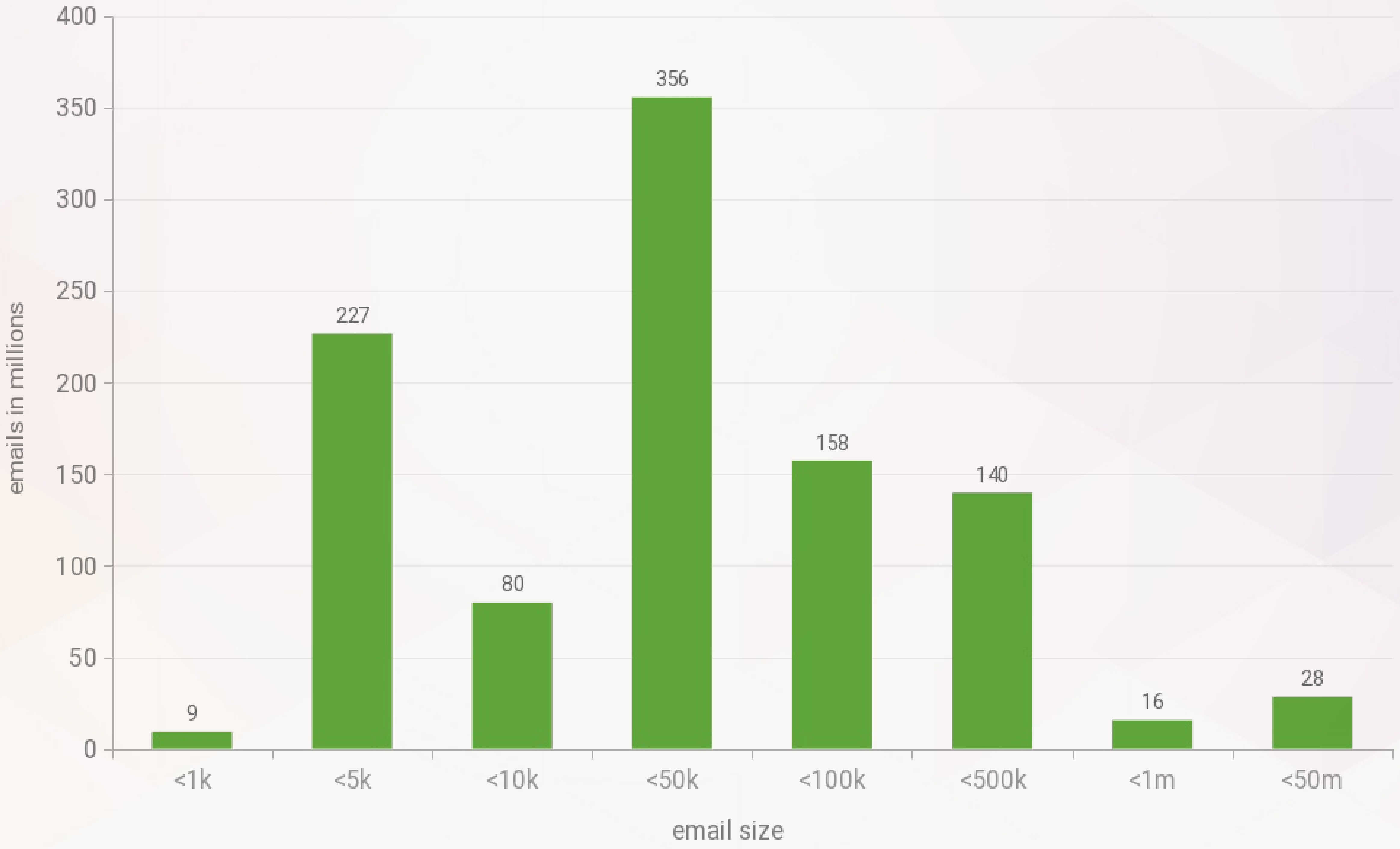
NFS Traffic


0,8%
0,7%
0,6%
57,9%
39,1%
0,2%
0,2%
0,1%
0,1%
GETATTR
ACCESS
LOOKUP
WRITE
READ
SETATTR
RENAME
REMOVE
CREATE

NFS relevant IOs



Email Distribution



How are emails stored?

Emails are written once, read many (WORM)

mailbox, maildir

Usage depends on:

- protocol (IMAP vs POP3)
- user frontend (mailer vs webmailer)

usually separated metadata, caches and indexes

- lost of metadata/indexes is critical

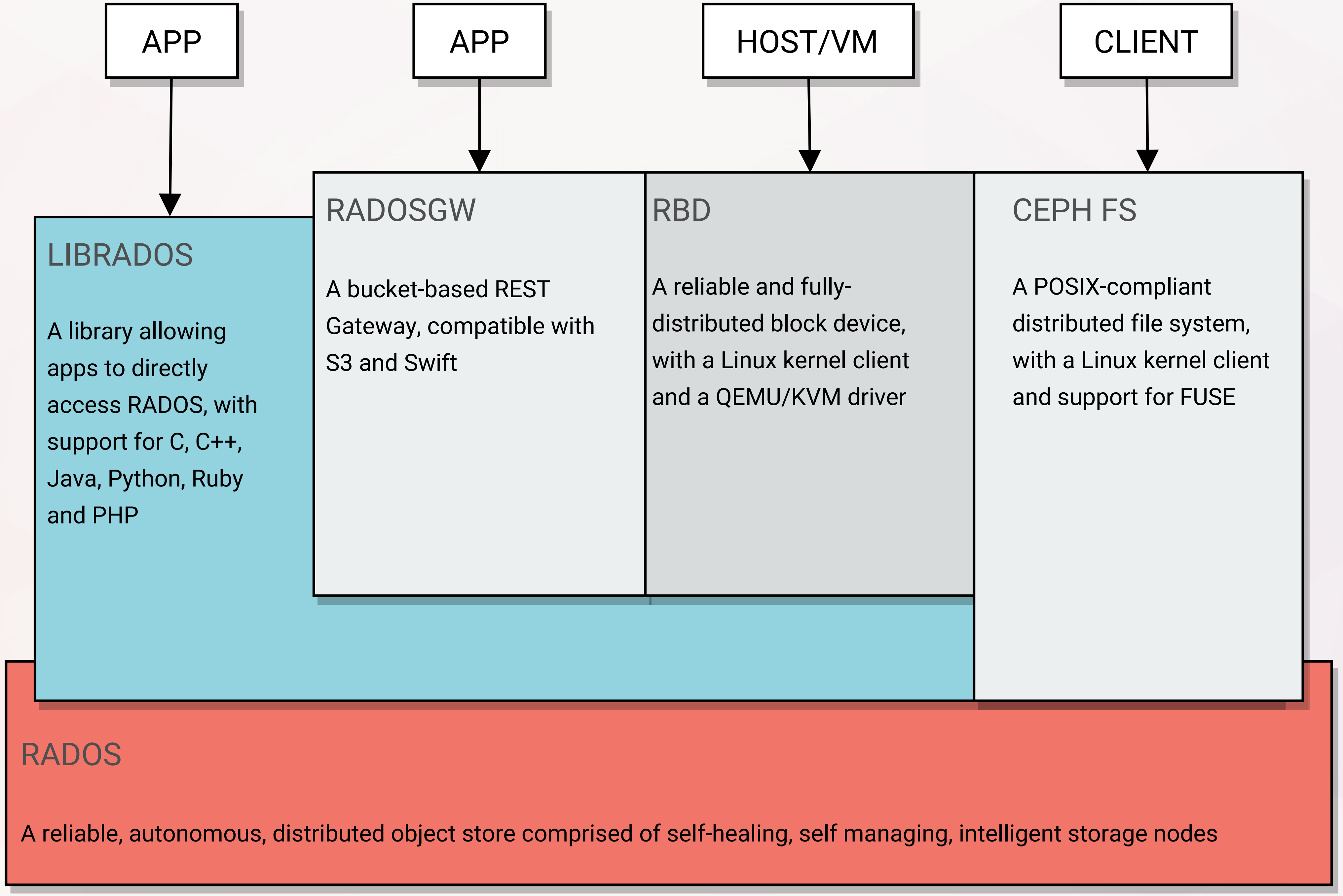
without attachments easy to compress

Motivation

- faster and automatic self healing
- less IO overhead
- prevent vendor lock-in
- commodity hardware
- open source where feasible
- reduce Total Cost of Ownership

Ceph

ceph



APP

APP

HOST/VM

CLIENT

LIBRADOS

A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby and PHP

RADOSGW

A bucket-based REST Gateway, compatible with S3 and Swift

RBD

A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

RADOS

A reliable, autonomous, distributed object store comprised of self-healing, self managing, intelligent storage nodes

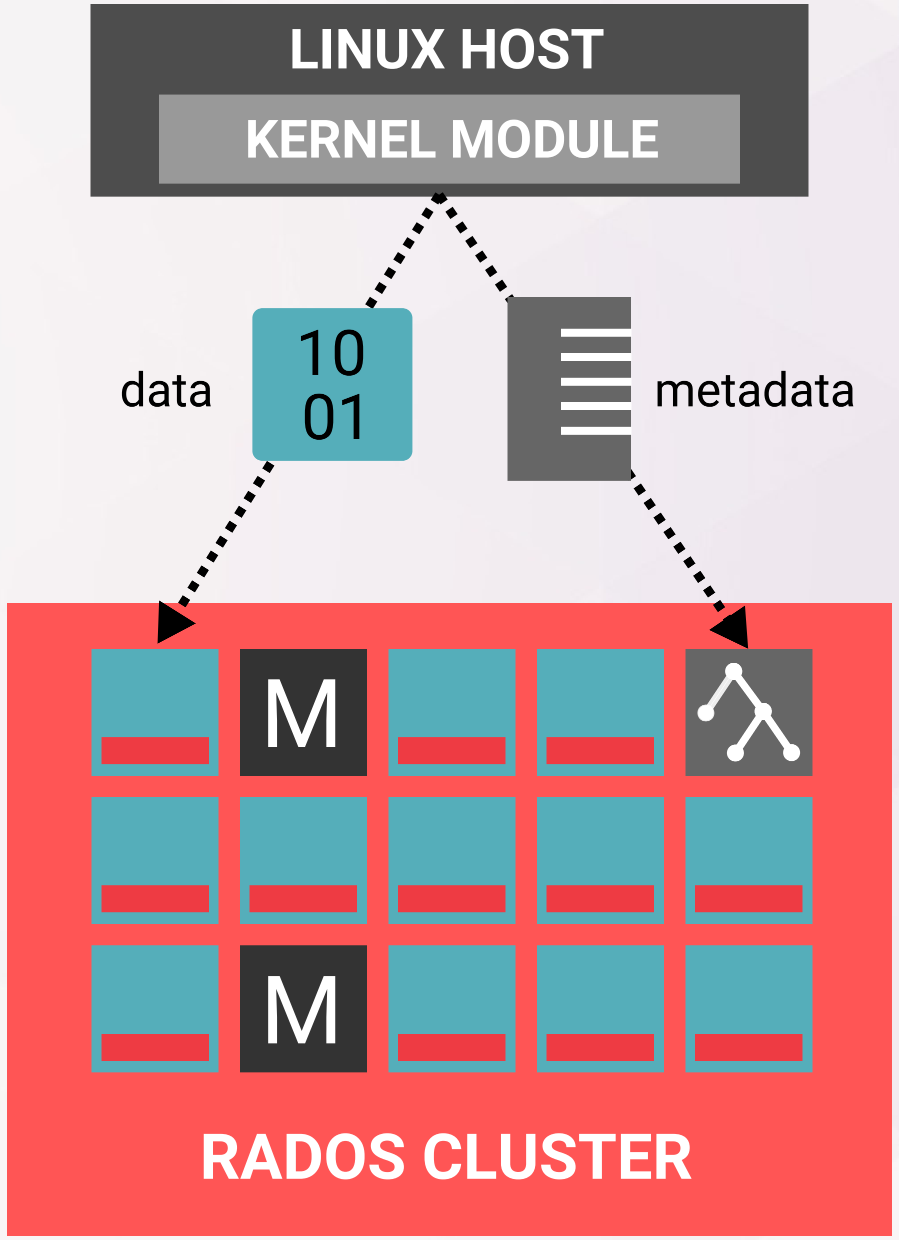
Where to store in Ceph?

CephFS

- same issues as NFS
- mail storage on POSIX layer adds complexity
- no option for emails
- usable for metadata/caches/indexes

Security

- requires direct access to storage network
- only for dedicated platform



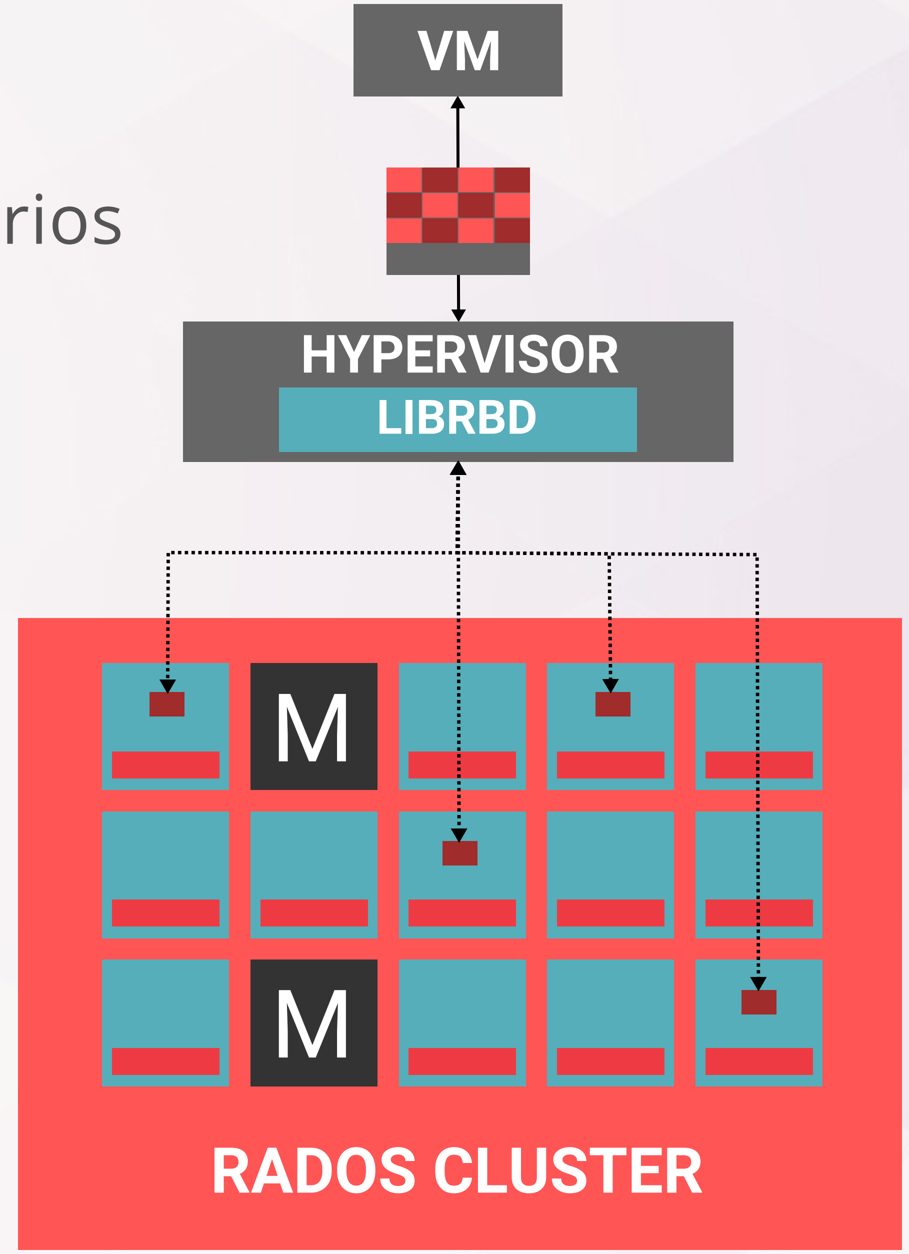
Where to store in Ceph?

RBD

- needs sharding and large RBDs
- needs account migration and RBD/fs extend scenarios
- still includes POSIX layer as NFS
- no sharing between clients
- impracticable

Security

- no direct access to storage network required
- secure through hypervisor abstraction (libvirt)



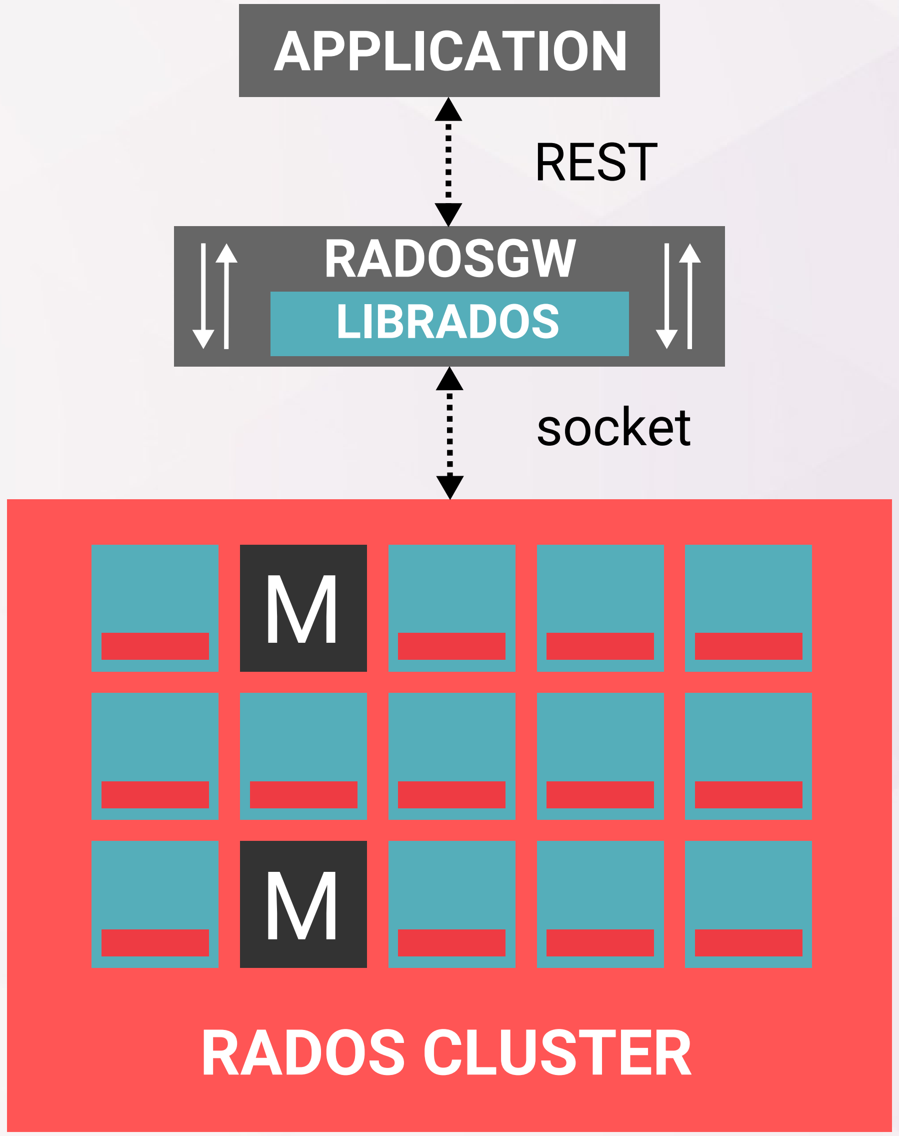
Where to store in Ceph?

RadosGW

- can store emails as objects
- extra network hops
- potential bottleneck
- very likely not fast enough

Security

- no direct access to Ceph storage network required
- connection to RadosGW can be secured (WAF)



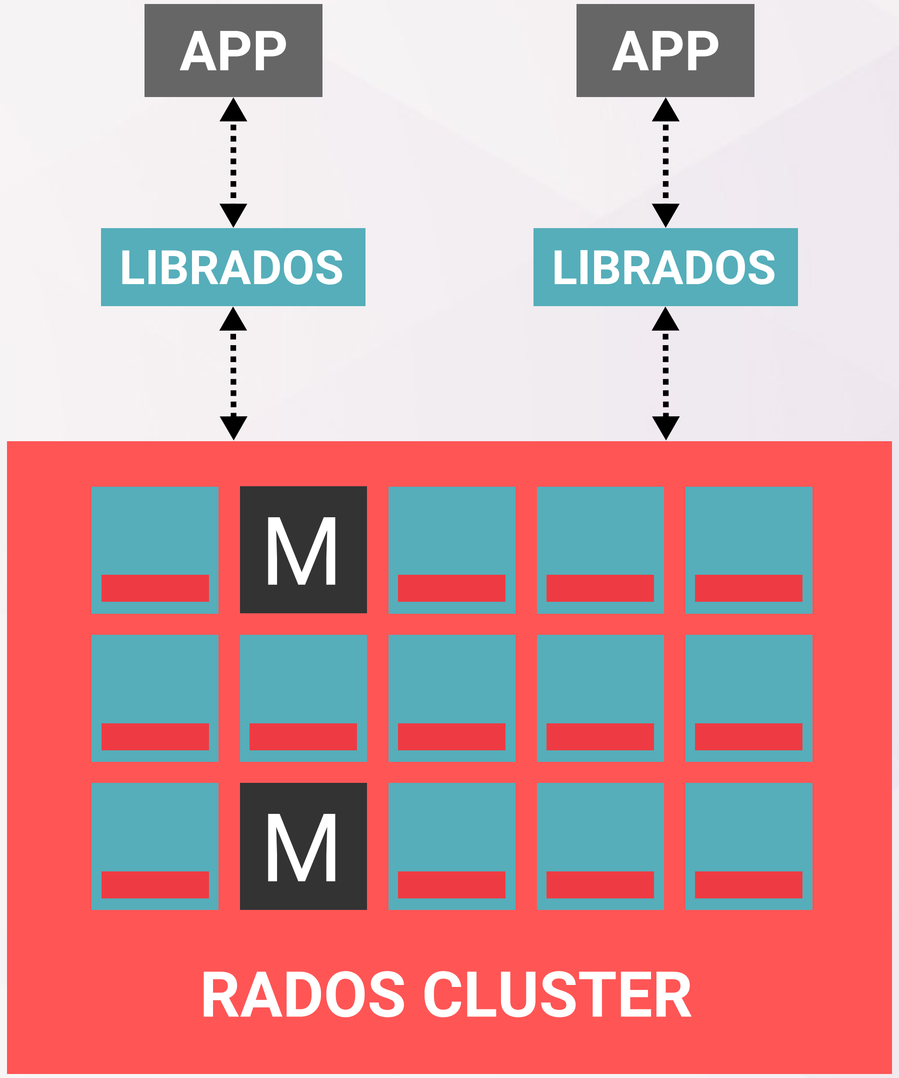
Where to store in Ceph?

Librados

- direct access to RADOS
- parallel I/O
- not optimized for emails
- how to handle metadata/caches/indexes?

Security

- requires direct access to storage network
- only for dedicated platform



Dovecot and Ceph

Dovecot

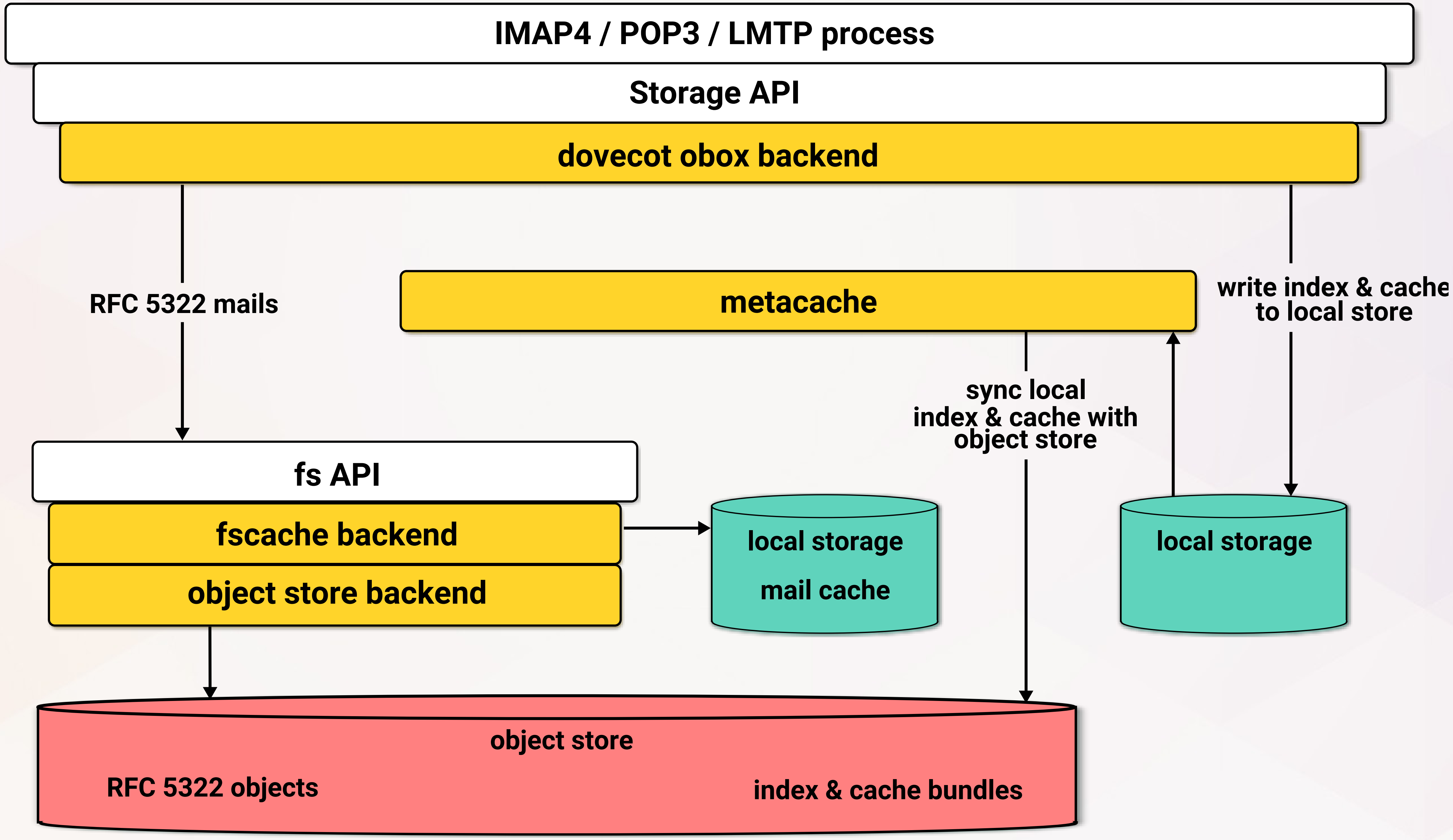
Open source project (LGPL 2.1, MIT)

72% market share (openemailsurvey.org, 02/2017)

Objectstore plugin available (obox)

- supports only REST APIs like S3/Swift
- not open source
- requires Dovecot Pro licence
    - large impact on TCO

Dovecot Pro obox Plugin



DT's approach

- no open source solution on the market
- closed source is no option
- develop / sponsor a solution
- open source it
- partner with:
  - Wido den Hollander (42on.com)
  - Tallence AG for development
  - SUSE for Ceph

Ceph plugin for Dovecot

First Step: hybrid approach

Emails

- RADOS Cluster

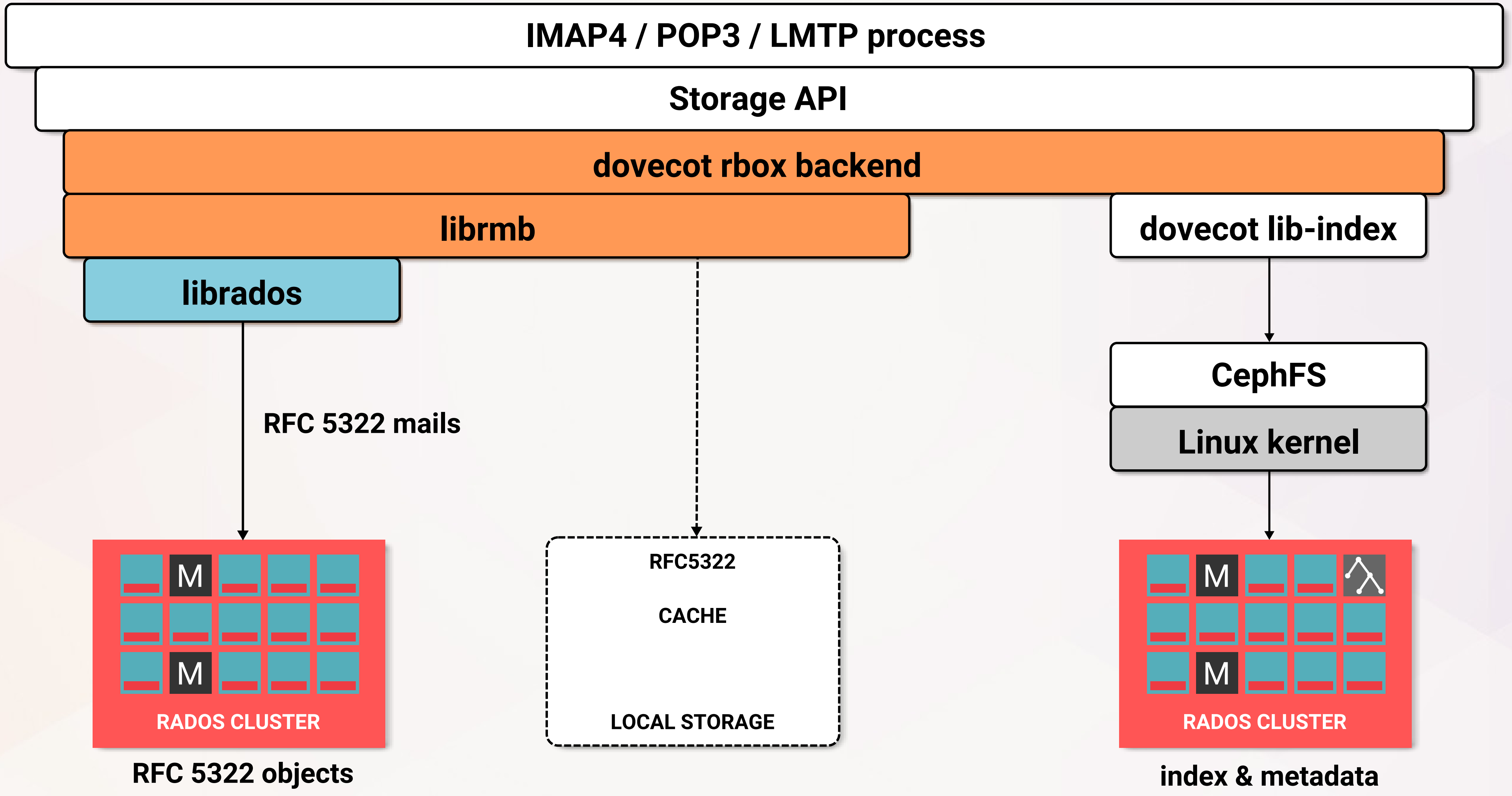
Metadata and indexes

- CephFS

Generic email abstraction on top of librados

- Split code into libraries
- Integrate into corresponding upstream projects

Librados mailbox (librmb)



librmb - Mail Object Format

Mails are immutable regarding the RFC-5322 content

RFC-5322 content stored in RADOS directly

Immutable attributes used by Dovecot stored in RADOS xattr

- rbox format version
- GUID
- Received and save date
- POP3 UIDL and POP3 order
- Mailbox GUID
- Physical and virtual size
- Mail UID

writable attributes are stored in Dovecot index files

Dump email details from RADOS

```
$> rmb -p mail_storage -N t1 ls M=ad54230e65b49a59381100009c60b9f7

mailbox_count: 1

MAILBOX: M(mailbox_guid)=ad54230e65b49a59381100009c60b9f7
         mail_total=2, mails_displayed=2
         mailbox_size=5539 bytes

         MAIL:   U(uid)=4
                 oid = a2d69f2868b49a596a1d00009c60b9f7
                 R(receive_time)=Tue Jan 14 00:18:11 2003
                 S(save_time)=Mon Aug 21 12:22:32 2017
                 Z(phy_size)=2919 V(v_size) = 2919 stat_size=2919
                 M(mailbox_guid)=ad54230e65b49a59381100009c60b9f7
                 G(mail_guid)=a3d69f2868b49a596a1d00009c60b9f7
                 I(rbox_version): 0.1
[..]
```

RADOS Dictionary Plugin

make use of Ceph omap key/value store

RADOS namespaces

- `shared/<key>`
- `priv/<key>`

used by Dovecot to store metadata, quota, ...

It's open source!

License: LGPLv2.1

Language: C++

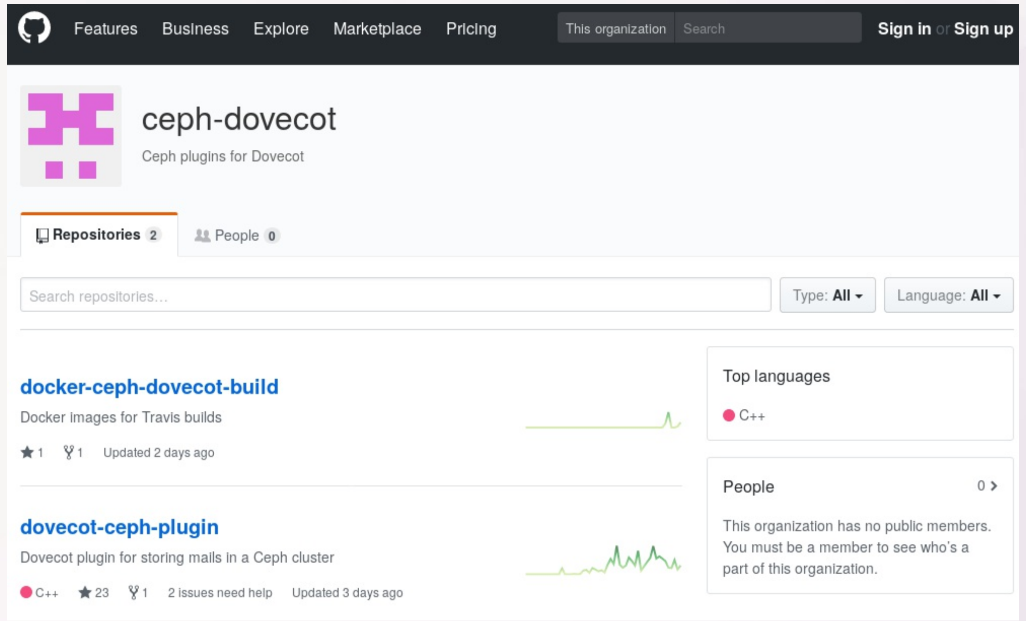
Location: github.com/ceph-dovecot/

Supported Dovecot versions:

- 2.2 >= 2.2.21
- 2.3

still under development

initial SLES12-SP3 and openSUSE RPMs: https://goo.gl/FymRhu



ceph-dovecot

Ceph plugins for Dovecot

docker-ceph-dovecot-build

Docker images for Travis builds

★ 1 · 1 · Updated 2 days ago

dovecot-ceph-plugin

Dovecot plugin for storing mails in a Ceph cluster

● C++ · ★ 23 · 1 · 2 issues need help · Updated 3 days ago

Top languages

● C++

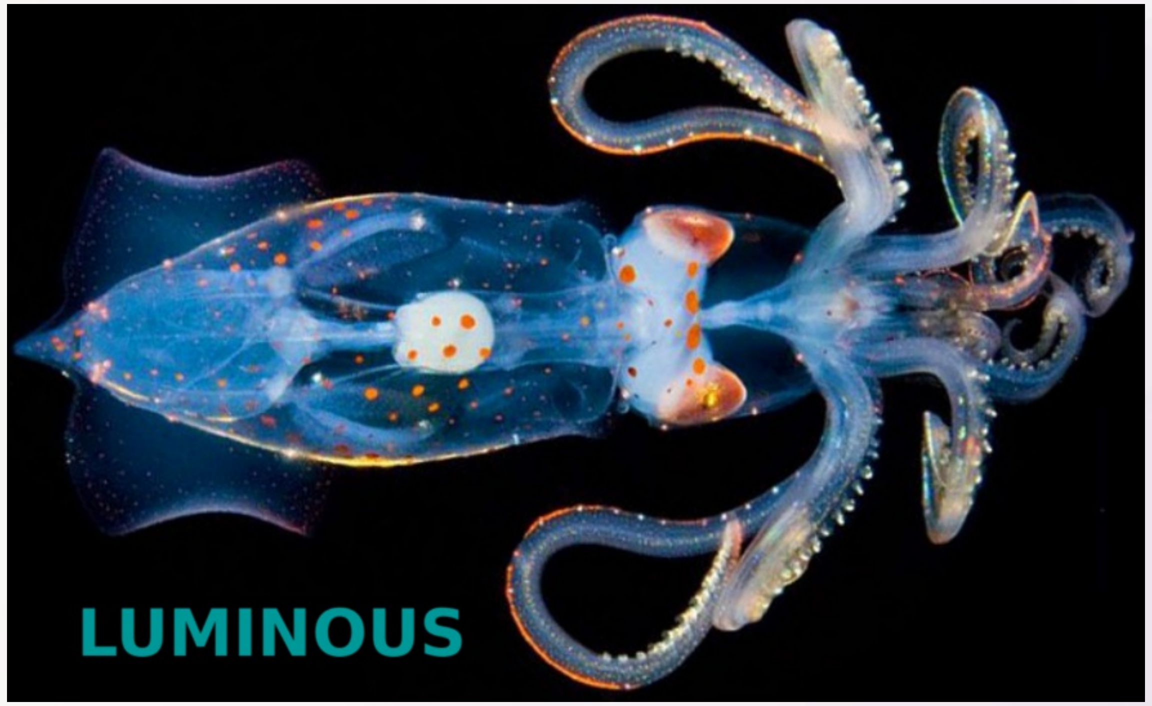
People 0 ›

This organization has no public members. You must be a member to see who's a part of this organization.

Which Ceph Release?

Required Features:

- Bluestore
  - write performance is critical
  - should be at least 2x faster than filestore
- CephFS
  - Stable release
  - Multi-MDS
- Erasure coding
  - Cost reduction
- Resiliency, reliability and fault tolerance



Enterprise products used:

- SES 5, SLES 12-SP3

ceph

Hardware

Commodity x86_64 server

- HPE ProLiant DL380 Gen9/10
- Dual Socket
  - Intel Xeon® E5 V4
- 2x Intel® X710-DA2 Dual-port 10G
  - 40G in total
- 2x boot SATA SSDs
- HBA, no seprate RAID controller

Commodity x86_64 server

CephFS Nodes (MDS, OSDs)

- **CPU:** 2x E5-2643v4 @ 3.4 GHz, 6 Cores, turbo 3.7GHz
- **RAM:** 256 GByte, DDR4, ECC
- **OSDs:** 8x 1.6 TB SSD, 3 DWPD, SAS, RR/RW 125k/92k iops

Rados Nodes (OSDs)

- **CPU:** 2x E5-2640v4 @ 2.4 GHz, 10 Cores, turbo 3.4GHz
- **RAM:** 128 GByte, DDR4, ECC
- **SSD:** 2x 400 GByte, 3 DWPD, SAS, RR/RW 108k/49k iops (for BlueStore database)
- **HDD:** 10x 4 TByte, 7.2K, 128 MB cache, SAS

Why this specific HW?

Community recommendations?

- OSD: 1x 64-bit AMD-64, 1GB RAM/1TB of storage, 2x 1GBit NICs
- MDS: 1x 64-bit AMD-64 quad-core, 1 GB RAM minimum per MDS, 2x 1GBit NICs

NUMA, high clocked CPUs and large RAM overkill?

- It's PoC hardware! Better save than sorry!
- Vendor did not offer single CPU nodes for number of drives
- MDS performance is mostly CPU clock bound and partly single threaded
  - High clocked CPUs for fast single thread performance
- Large RAM: better caching!

Placement

Issues

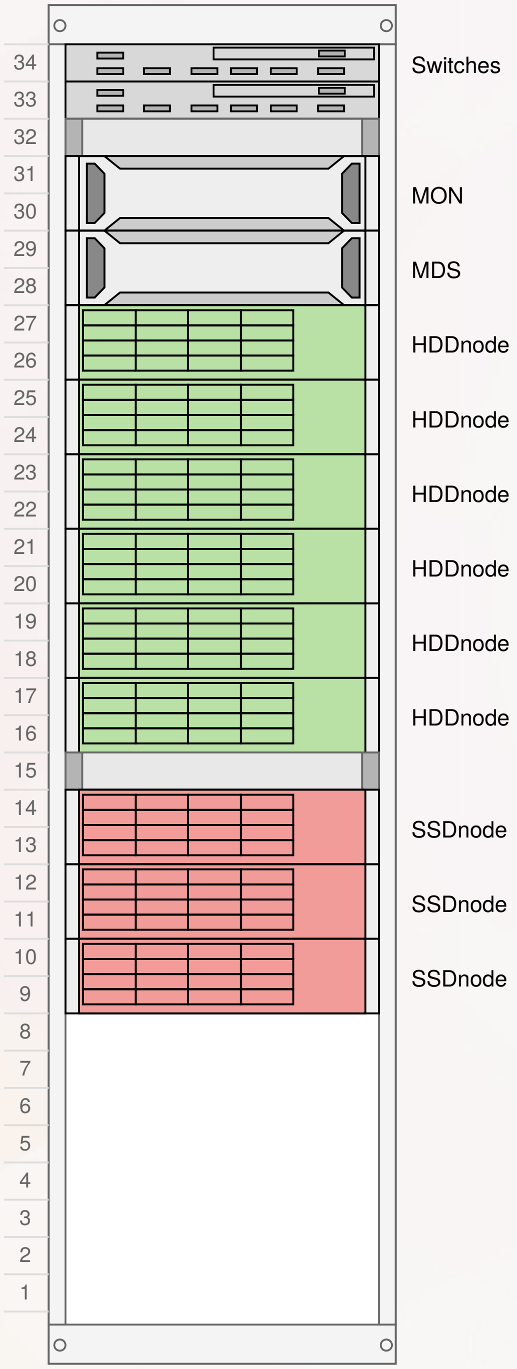
Datacenter

- Usually two independent fire compartments (FCs)
- May additional virtual FCs
- How to place 3 copies?

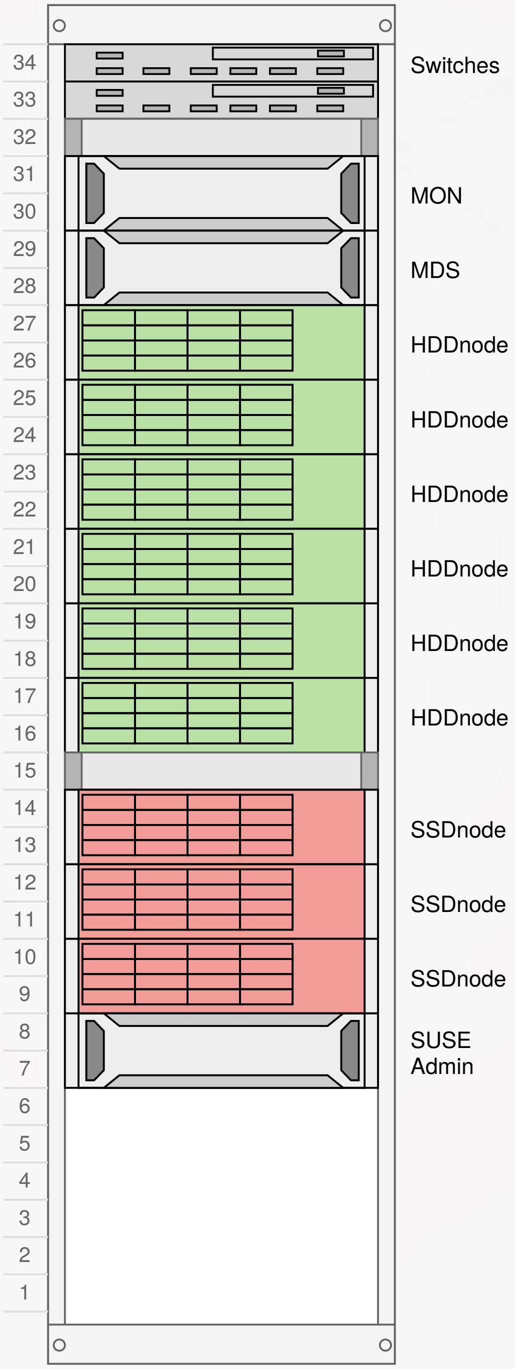
Requirements

- Lost of customer data MUST be prevented
- Data replication at least 3 times (or equivalent)
- Any server, switch or rack can fail
- One FC can fail

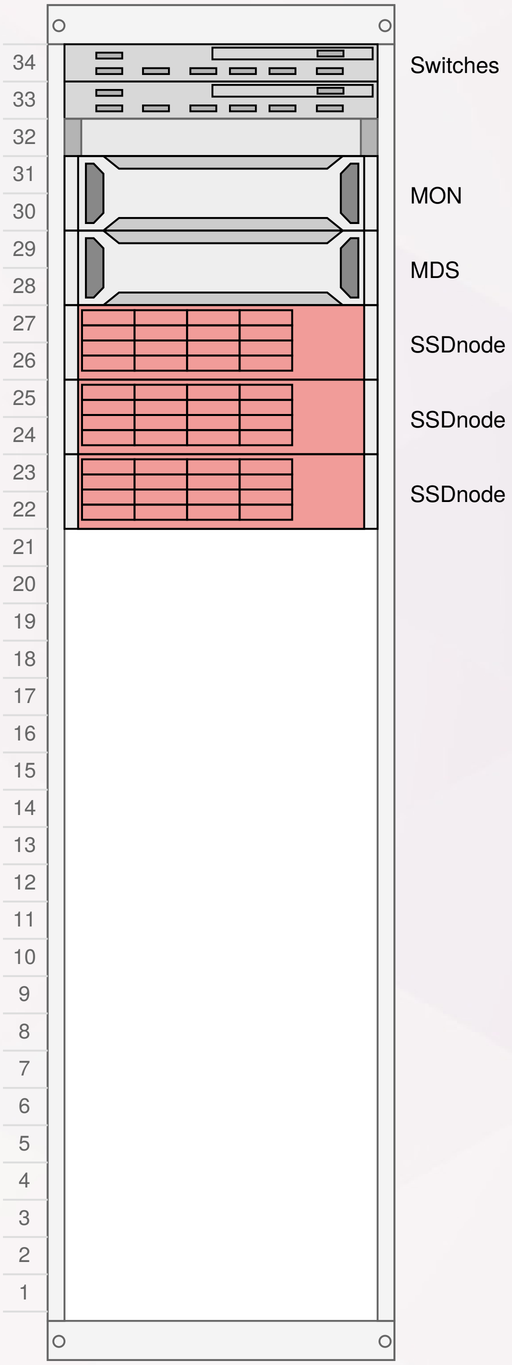
Fire Compartment A

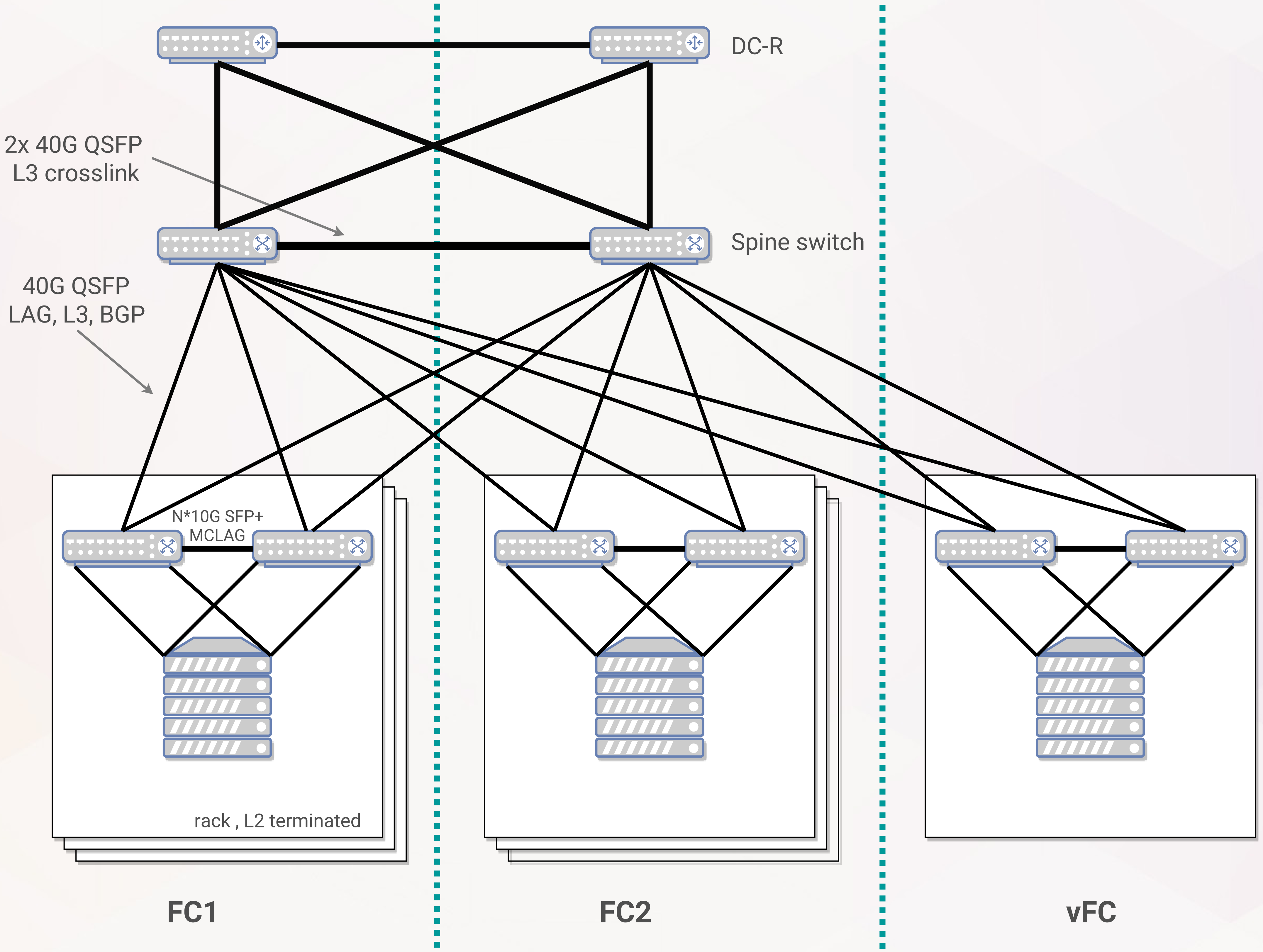


Fire Compartment B



Fire Compartment C
(third room)




DC-R
2x 40G QSFP
L3 crosslink
Spine switch
40G QSFP
LAG, L3, BGP
N*10G SFP+
MCLAG
rack , L2 terminated
FC1
FC2
vFC

Status and Next Steps

Testing

5-node clusters (SUSE and DT's labs)

- Functional IMAP testing against upstream dovecot
  - successful!
  - fixed issues on the way
- Functional testing against DT's installation
  - in progress

Proof-of-Concept

Hardware

- 9 SSD nodes for CephFS
- 12 HDD nodes
- 3 MDS / 3 MON

2 FCs + 1 vFC

Testing

- run load tests
- run failure scenarios against Ceph
- improve and tune Ceph setup
- verify and optimize hardware

Topics to solve

Erasure Coding

- select plugin and profile
- EC performance with small writes

Compression

- BlueStore inline compression
- implement support in librmb

OMAP performance

Cluster network bandwidth

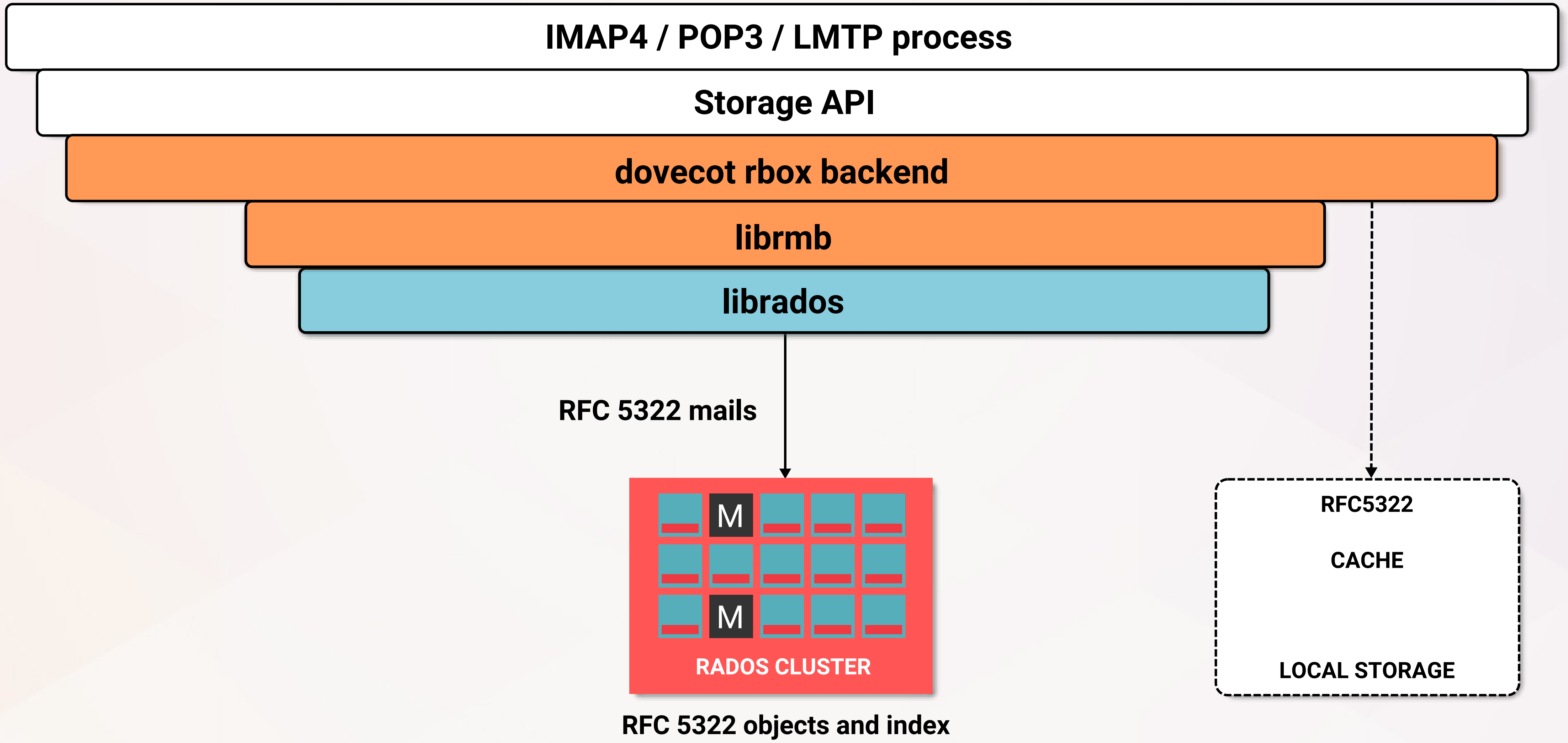
Move to Production

Production

- verify if all requirements are fulfilled
- integrate in production
- migrate users step-by-step
- extend to final size
  - 128 HDD nodes, 1200 OSDs, 4,7 PiB
  - 15 SSD nodes, 120 OSDs, 175 TiB

Further Development

Goal: Pure RADOS backend, store metadata/index in Ceph omap



Conclusion

Summary and conclusions

Ceph can replace NFS

- mails in RADOS
- metadata/indexes in CephFS
- BlueStore, EC

librmb and dovecot rbox

- Open Source, LGPLv2.1
- librmb can be used in non-dovecot systems
- still under development

PoC in progress

Be invited to: Participate!

Try it, test it, feedback and report bugs!
Contribute!

github.com/ceph-dovecot/

Thank you.

Questions?

ceph