

# Skalierendes Mailstorage mit Dovecot und ceph



2016032101 / Stephan Seitz <s.seitz@heinlein-support.de>

## Wer ist die Heinlein Support GmbH?

- Wir bieten seit 20 Jahren Wissen und Erfahrung rund um Linux-Server und E-Mails
- IT-Consulting und 24/7 Linux-Support mit 20 Mitarbeitern
- Eigener Betrieb eines ISPs seit 1992
- Täglich tiefe Einblicke in die Herzen der IT aller Unternehmensgrößen

# Inhalt

- Problembeschreibung
  - Grenzen von Mailservern
- Lösungen
  - aktuelle „klassische“ Möglichkeiten
  - Lösung mit horizontal skalierbaren Komponenten
- Werkzeuge
  - ceph / librados
    - S3 Gateway
  - Dovecot
    - obox / fscache Backend
    - obox / fscache mit S3
- Hands On
  - Aufbau der Komponenten
- Analyse und Messwerte



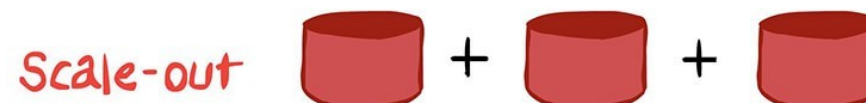
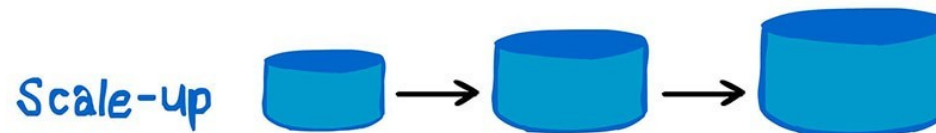
## Problembeschreibung

- IMAP statt POP
  - weitere Dienste bei Mailanbietern
    - Dokumentenspeicher (z.B. dropbox-Alternativen)
- Mails liegen beim Provider
  - möglicherweise über Jahre
  - und dauerhaft verfügbar
- Anforderungen steigen
  - Speicherkapazität
  - Zugriffsgeschwindigkeit
  - Nachhaltigkeit



## Ziele

- frei skalierbares Storage
- frei skalierbare Mail-Infrastruktur
- Für alle Komponenten
  - einfacher Aufbau
  - Software-defined (Idealerweise: Open Source)
  - Horizontale Skalierbarkeit / Scale Out



## Lösungen / aktuelle „klassische“ Möglichkeiten

- Partitionieren bzw. koordinierte und geplante Ressourcenaufteilung
  - z.B. Postfächer a-d auf srv01, e-h auf srv02, etc.
  - Konsequenz
    - administrative Koordinationsarbeit
    - Komplexere HA-Konzepte
    - geplante Replikationsstrategien
    - Backup beeinflusst die Performance
- Storage-Solutions
  - NetApp, EMC z.B. proprietäre Cluster / Metro-Cluster Lösungen
  - Konsequenz
    - Vendor Lock In
    - Umfangreiche Lizenzen



# Lösung mit horizontal skalierbaren Komponenten

- frei skalierbares Storage
  - Verwendung eines Objekt Storages
- frei skalierbare Mail-Infrastruktur
  - Mailserver Backend für objektbasierte Speicher
- einfacher Aufbau
- Software-defined





## Ein kurzer Blick auf die Werkzeuge

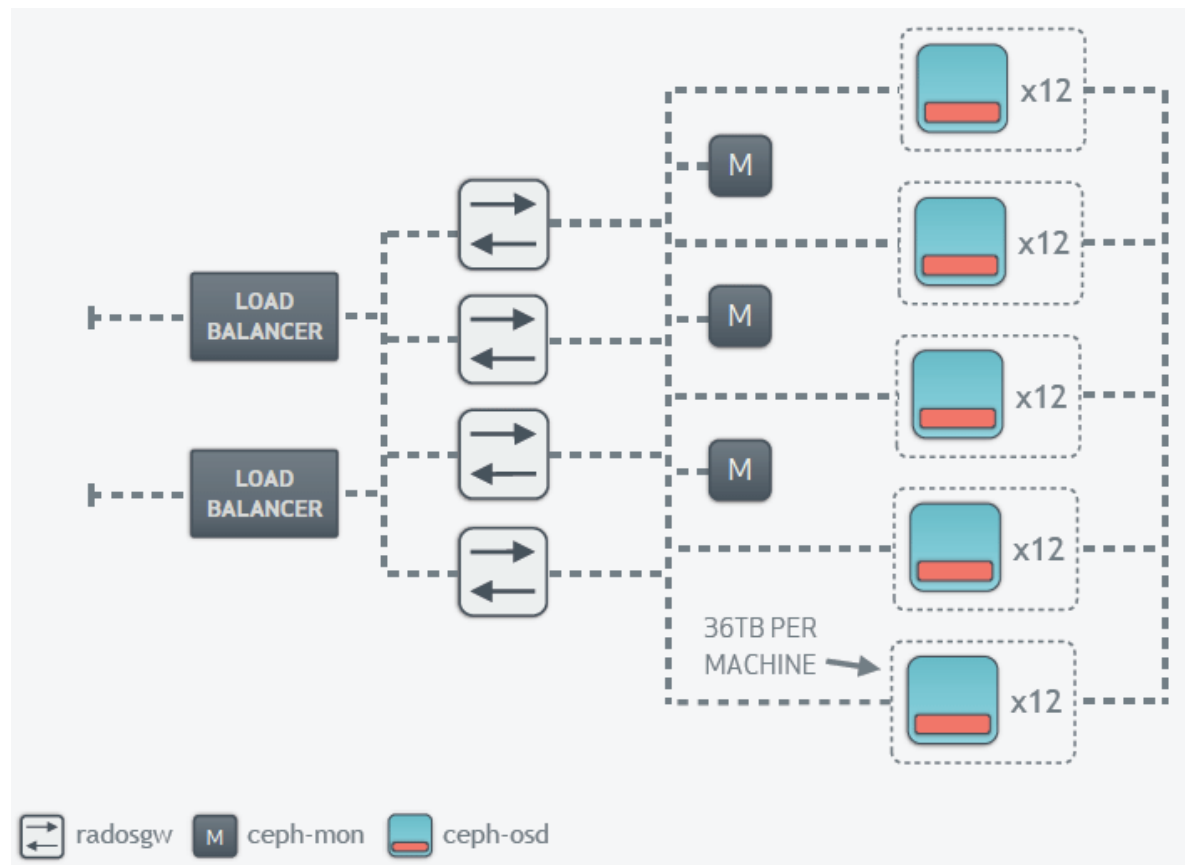


## ceph / librados



## ceph / Konzept und schematischer Aufbau

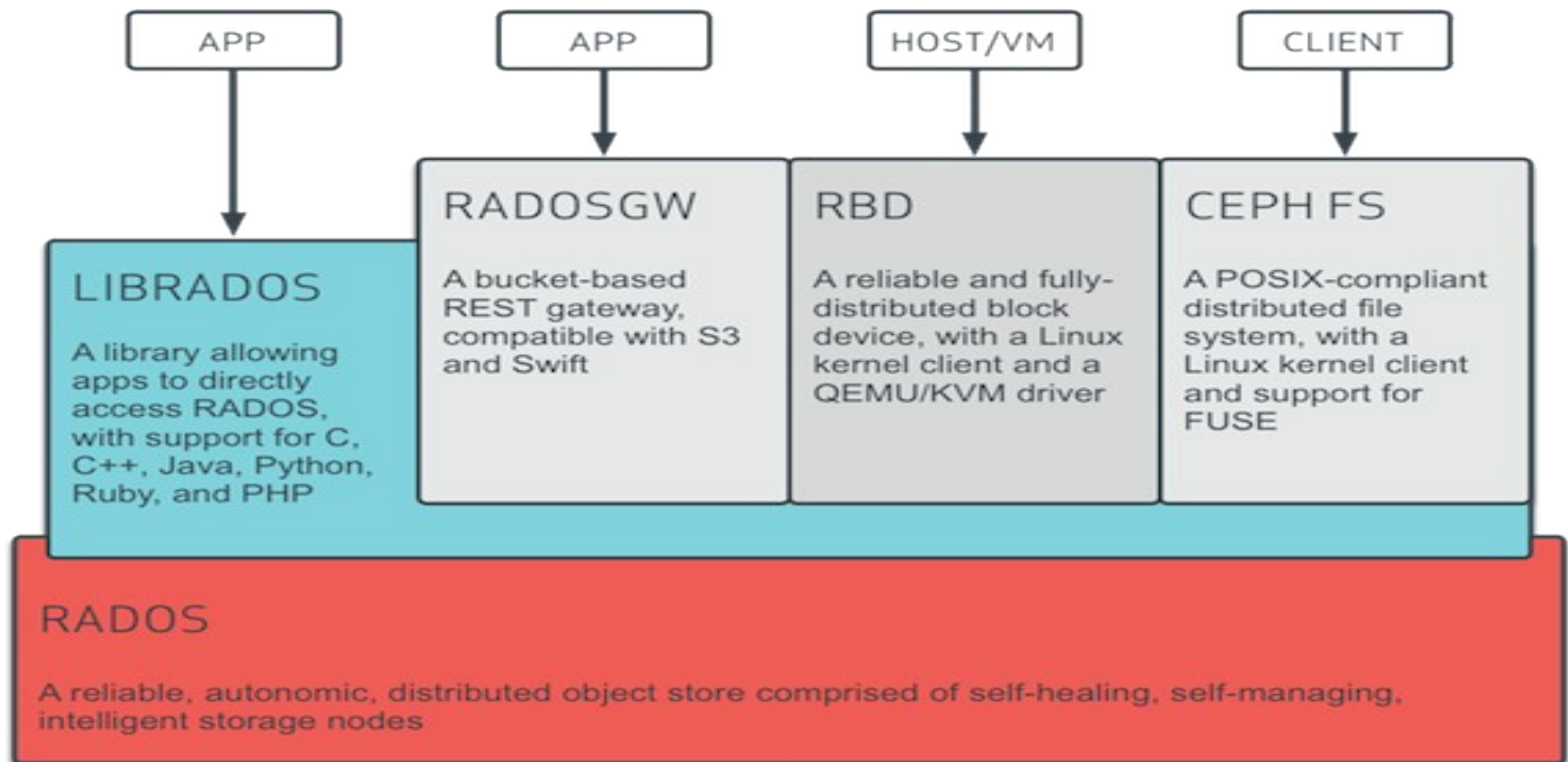
→ ceph kann parallel auf Objekte in OSD zugreifen



→ Hier mit multiplen RadosGW Instanzen und S3/HTTP Loadbalancer

## ceph / Softwarekomponenten

→ Flexibler, skalierbarer Object-Store



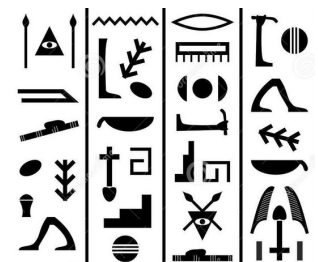


**klassischer Aufbau mit shared storage**

## Dovecot: - mit zentralen Storage-Komponenten -

- Begrenzte Skalierbarkeit durch beispielsweise
  - endliche SAN Kapazität
  - endliche I/O Performance der Controller
  - endliche I/O Performance der Infrastruktur (FC / iSCSI)
- SAN als Single-Point-of-Failure?
  - oder „nur“ als Flaschenhals?

legacy?



# Dovecot: Skalierbarer und Flexibler IMAP-Server

- Entwickelt von Timo Sirainen seit 2002
- Laut <http://openemailsurvey.org/> (Stand 2016)
  - ca. 3 Mio Mailserver in Freier Wildbahn
  - Anteil ca. 68% aller IMAP-Server
- performant / grundsätzlich geringe Block-I/O-Anforderungen
- einfach und flexibel konfigurierbar
- modularer Code / PlugIn-Struktur
- einfache Migrationsmöglichkeiten von anderen Systemen
- IMAP-Proxy ermöglicht Migrationen faktisch ohne Downtime
- administrationsfreundlich
  - sehr gute Dokumentation unter <http://wiki2.dovecot.org>
  - doveadm als „finnisches Taschenmesser“

## Dovecot: Speicherformate im Vergleich

- **Maildir:** Jede E-Mail liegt in einer einzelnen Datei
  - ✓ Einfach, manipulierbar
  - ✓ Locking- und datenbankfrei
  - ✗ Viele Dateien = Horror im Backup
- **mbox:** Eine Datei beinhaltet alle E-Mails hintereinander
  - ✓ Wenige Dateien, schnelles Backup
  - ✗ Schreib-Locks notwendig
  - ✗ I/O-lastig beim Löschen von Mails
- **mdbox:** Viele Dateien a 5-10 MB haben jeweils einige E-Mails (+Datenbank)
  - ✓ Wenige Dateien = schnelles Backup
  - ✓ Alt-Storage möglich
  - ✗ Administration nur noch über doveadm-Kommand
  - ✗ Vorsicht beim Backup wg. Datenbank

## Dovecot: Maildir: Der Aufbau

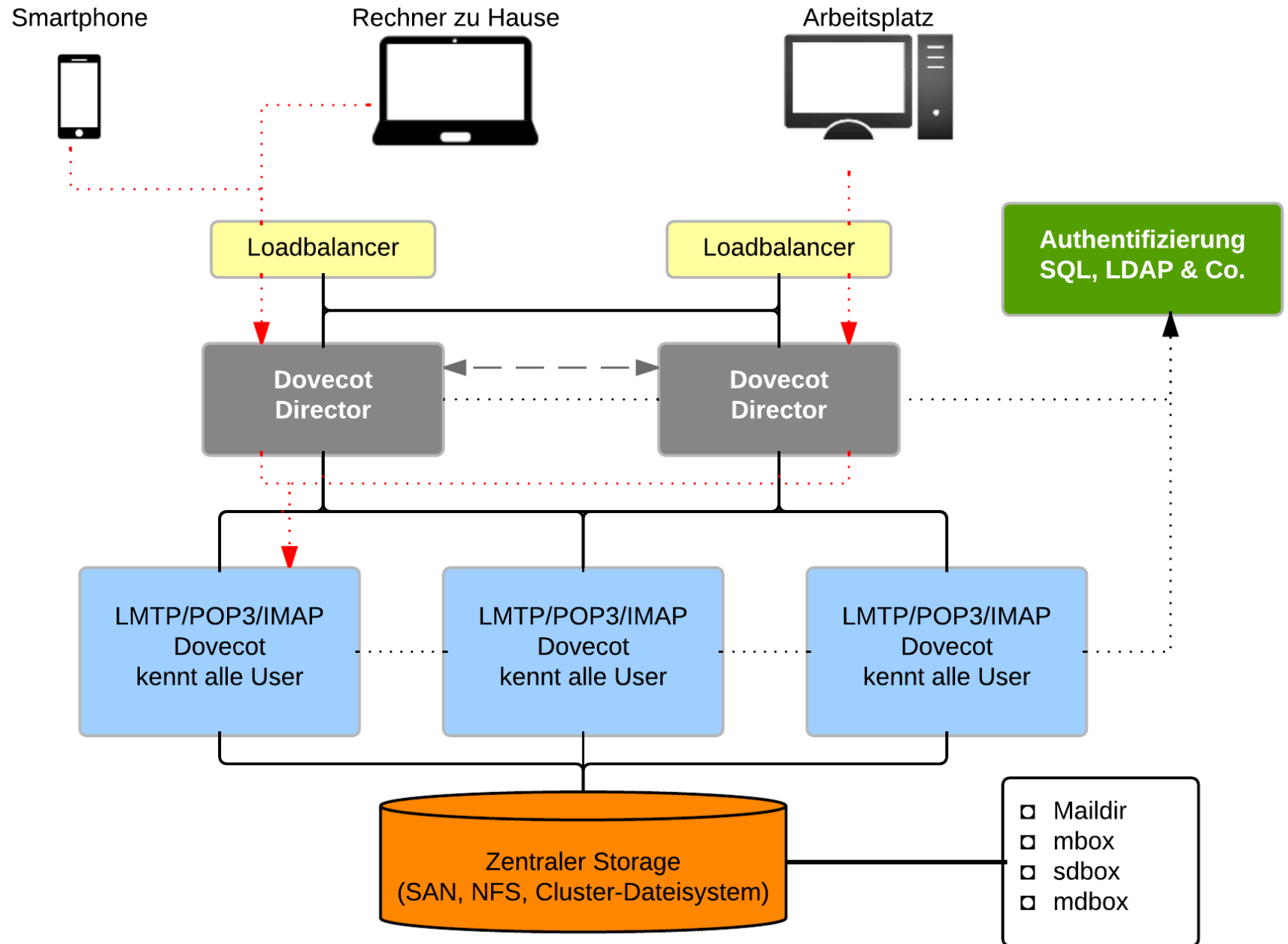
- cur | new | tmp = INBOX
  - .Urlaub = Folder (enthält cur|new|tmp)
  - .Urlaub.Spanien = Folder
- dovecot.index.\* = Caching-Index
- maildirsize = Quota-Protokoll
  - Können gelöscht werden!
- dovecot-acl-list = ACLs für Freigaben
- dovecot-uidlist = Zuordnung Mail <=> UUID
- dovecot-keywords = Keywords der Mails
- subscriptions = abonnierte Folder

```
drwx----- 2 cur
drwx----- 2 new
drwx----- 2 tmp
drwx----- 5 .Drafts
drwx----- 5 .Dummy
drwx----- 5 .Junk
drwx----- 5 .Sent
drwx----- 5 .Trash
drwx----- 5 .Urlaub
drwx----- 5 .Urlaub.Spanien
rw----- 1 dovecot.index.cache
rw----- 1 dovecot.index.log
rw----- 1 maildirsize
rw----- 1 dovecot-acl-list
rw----- 1 dovecot-uidlist
rw----- 1 dovecot-uidvalidity
rw----- 1 dovecot-uidvalidity.4ca9b4c8
rw----- 1 dovecot-keywords
rw----- 1 subscriptions
```



## Dovecot: Architektur Director/Proxy Cluster

- Linear skalierbar mit zusätzlichen Komponenten
  - ~ 50.000 gleichzeitige Verbindungen auf den Direktoren
  - ~ 15.000 gleichzeitige Verbindungen auf den IMAP-Backends
- Backends können komplett „stateless“ sein
- Direktoren arbeiten im Ringverbund
  - Informationen über User werden ausgetauscht
- Direktoren wissen über den Zustand der Backends
- Jedes System kann ausfallen
- Indizierung geschieht „on the Fly“ beim Ausliefern der Nachricht
- Index-Files sind selbstheilend





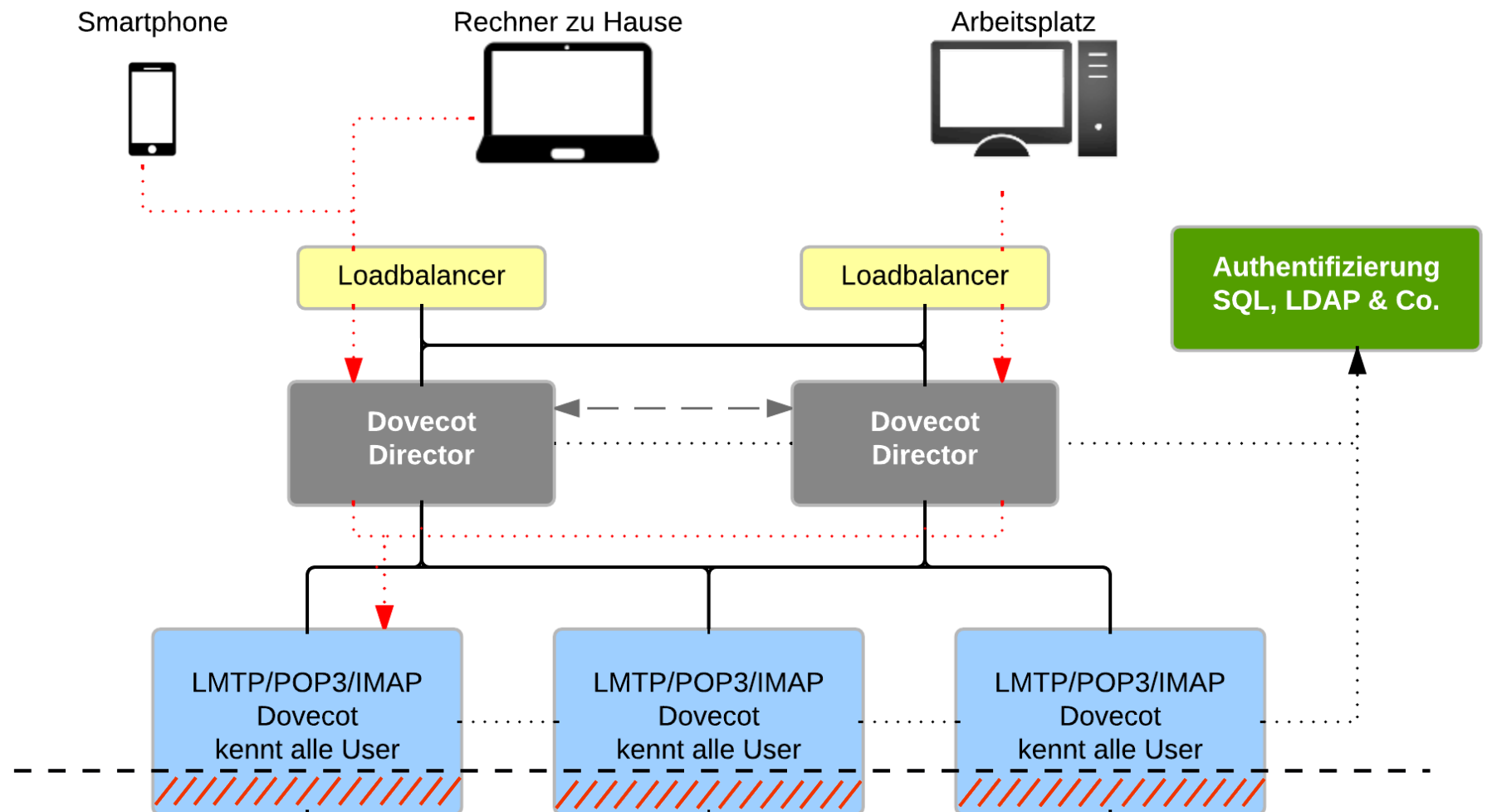
## **Aufbau mit object storage**

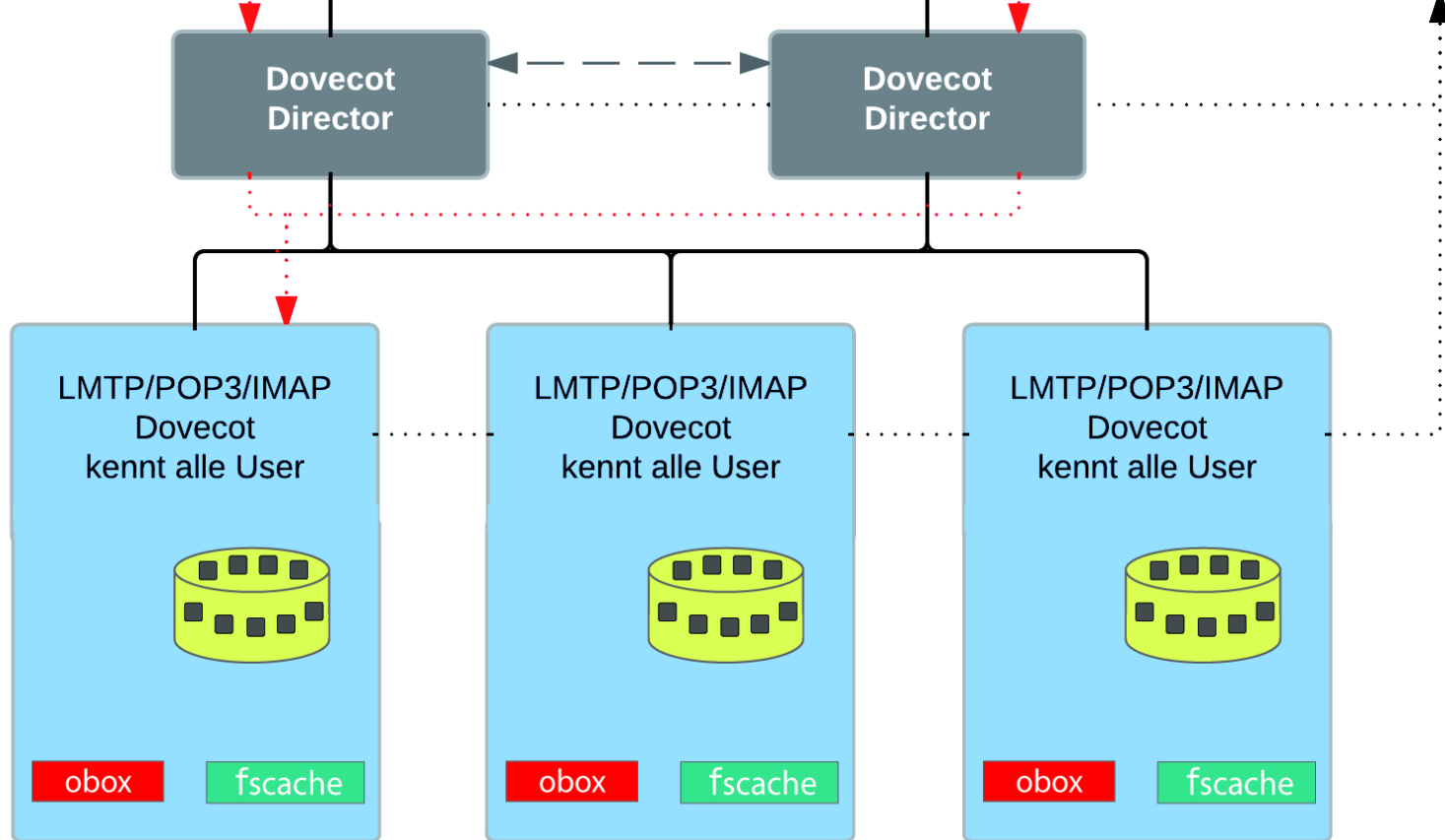
## Dovecot: Object-Storage pro/kontra

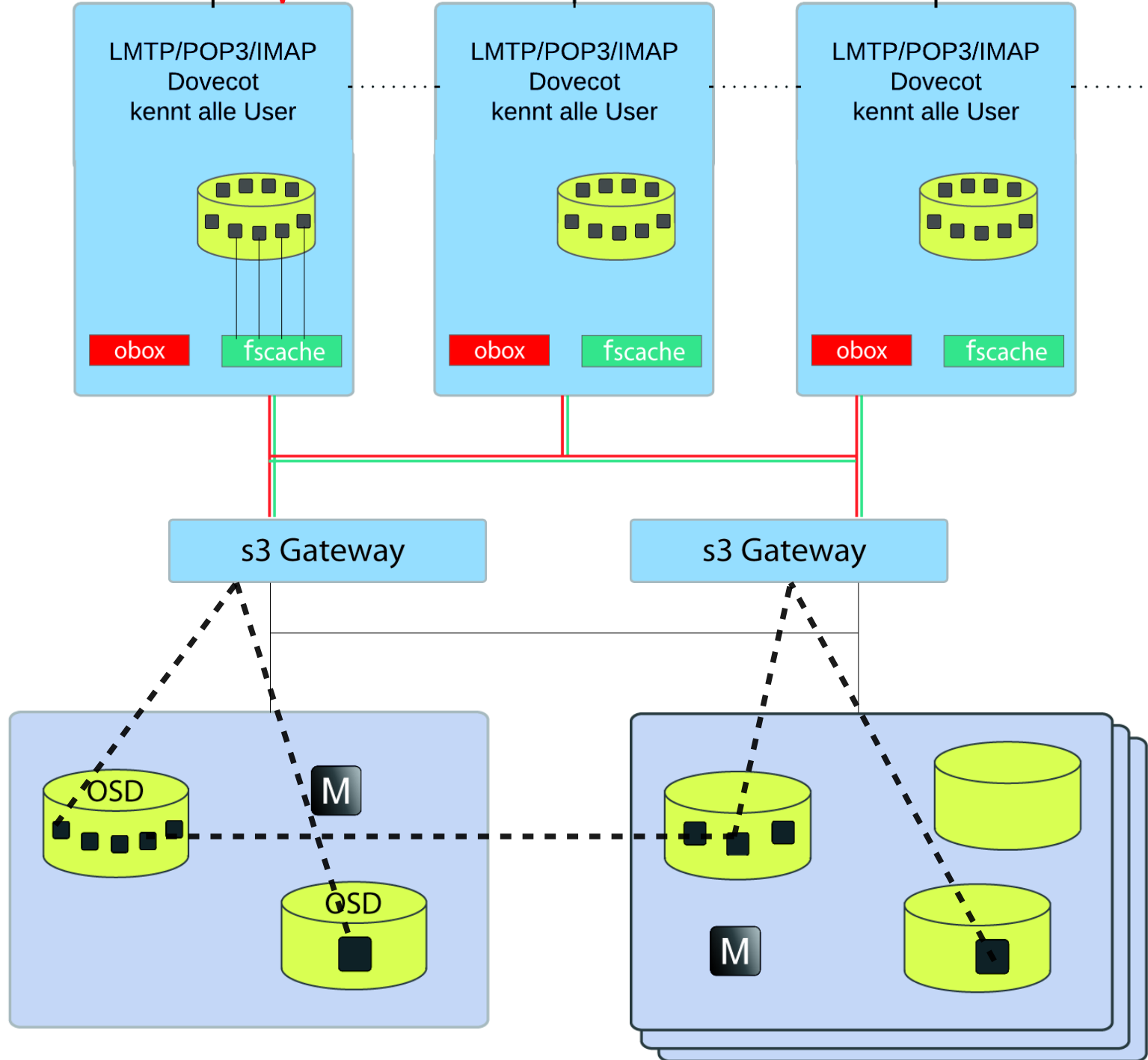
- ✓ Mails bei Amazon AWS „offen“ speichern?
- ✓ ceph skaliert massiv horizontal
  - ✓ radosgw bietet S3-API
- ✓ Partitionierung nach Geo-Location möglich
- ✓ Hohe Verfügbarkeit, d.h.
  - ✓ alle Komponenten sind cluster-/HA-fähig
- ✗ Latenzen sind deutlich höher als bei lokalem Storage
- ✗ Abrufen vieler Mails braucht dementsprechen mehr Zeit

## Dovecot: Object-Storage (obox)

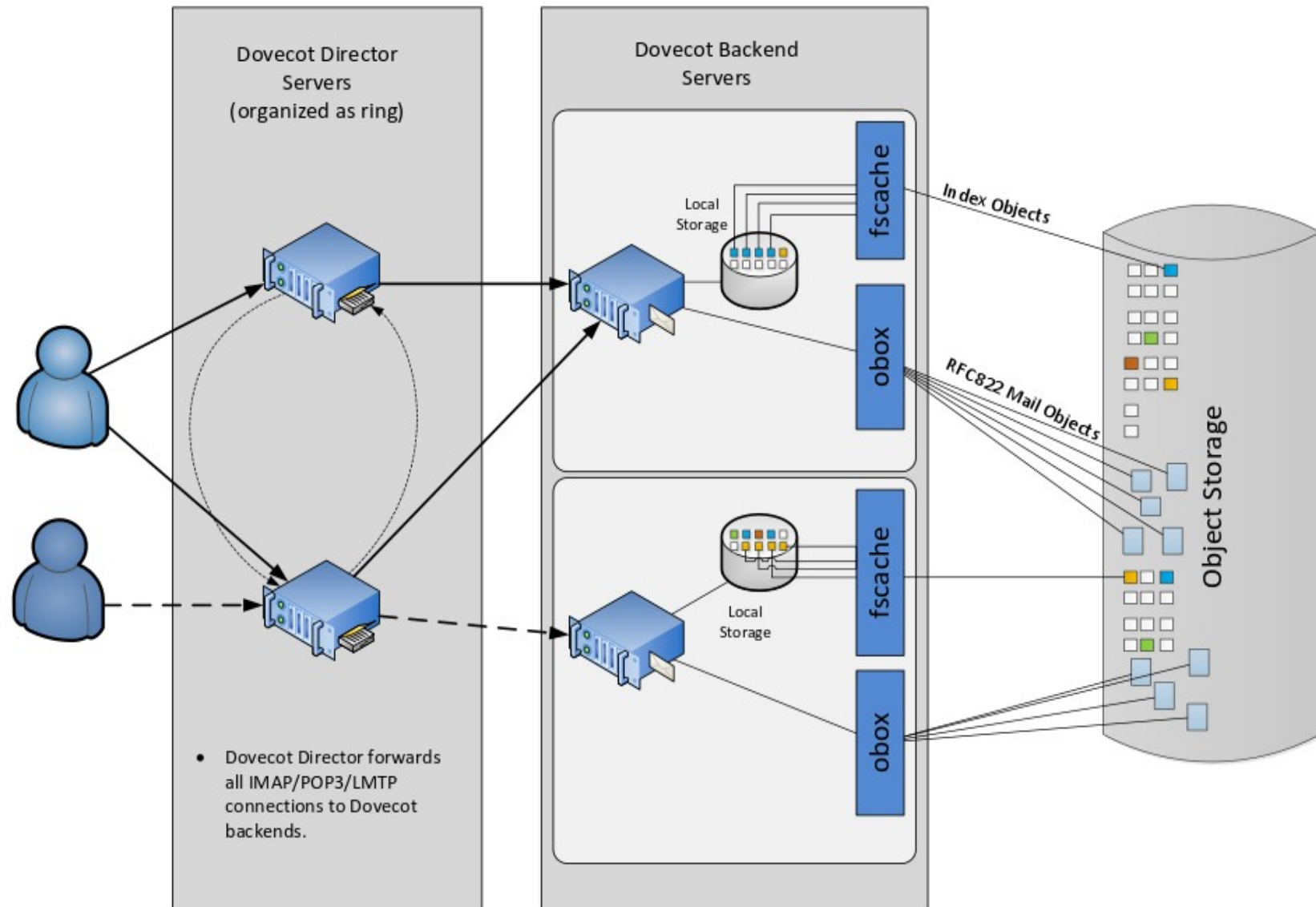
- obox - Mailbox-Format
  - ein Objekt ↔ eine Mail
    - Mails liegen als Objekte vor
  - Massiv paralleles Lesen und Schreiben
- Index-Files werden via *fscache* bereitgestellt
  - **/var/lib/dovecot/cache** ... auch via RBD
- Cache-Indexfiles im Dateisystem oder RAM
- Änderungen werden alle paar Minuten in den Object-Store geschrieben
- Dank Direktor landet der User immer auf dem selben Server
  - und benutzt demzufolge den gleichen Cache

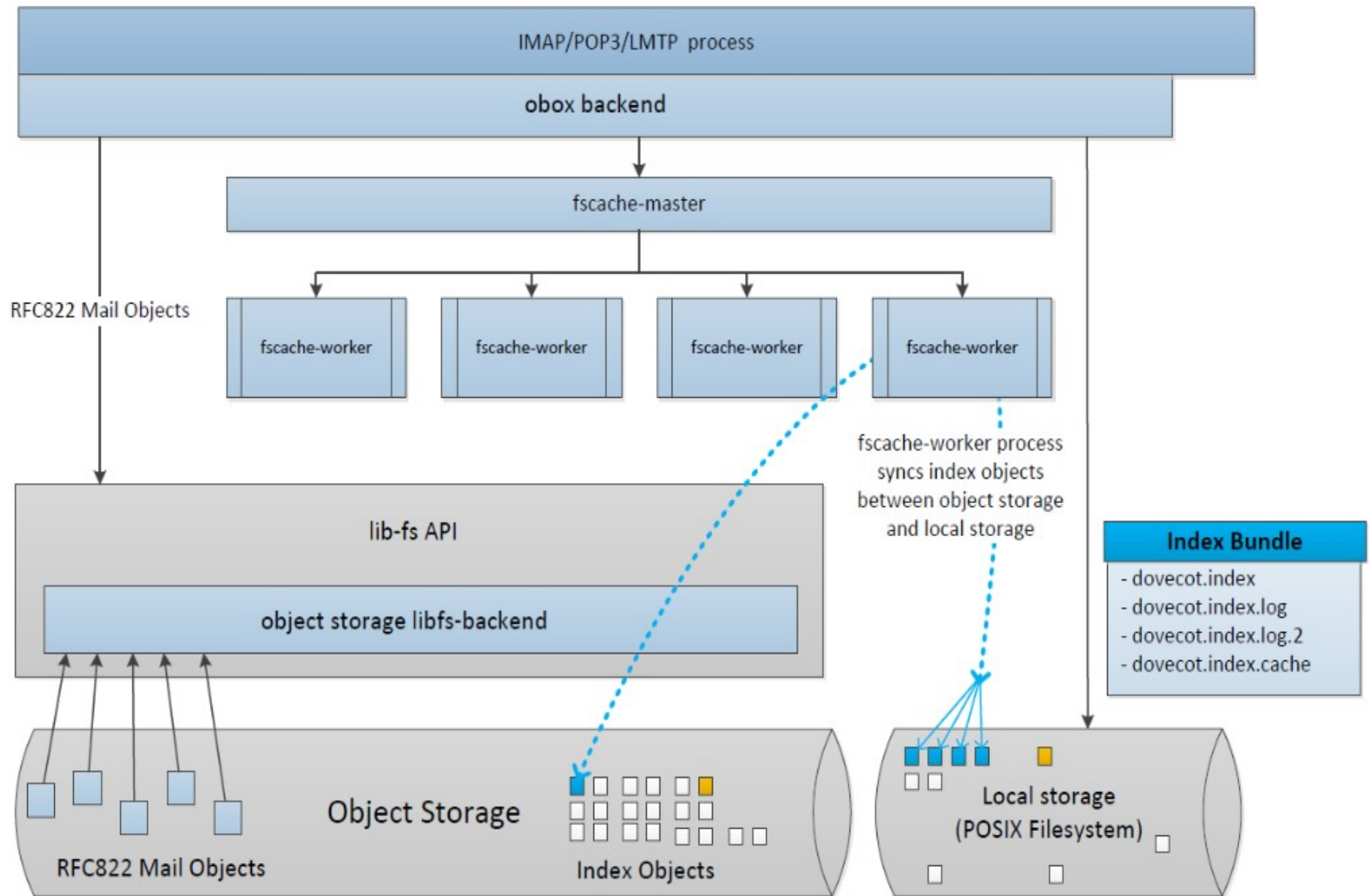












# **Hands On!**

## **Implementierung und Test**



## ceph: notwendige Pools und radosgw-Konfiguration

→ `ceph osd pool create {poolname} {pgs} {pgps} replicated`

```
root@ceph-s3-gateway:~# ceph osd lspools | sed 's/,/\n/g'
47 .rgw.root
48 .rgw.control
49 .rgw
50 .rgw.gc
51 .users.uid
52 .users.email
53 .users
54 .rgw.buckets.index
55 .rgw.buckets
56 rbd
```

→ <http://docs.ceph.com/docs/master/radosgw/config/>

→ `ceph.conf` / gateway Konfiguration

```
root@ceph-s3-gateway:~# cat /etc/ceph/ceph.conf | tail -n 8
[client.radosgw.ceph-s3-gateway]
host = ceph-s3-gateway
keyring = /etc/ceph/ceph.client.radosgw.keyring
rgw socket path = ""
rgw frontends = fastcgi socket_port=9000 socket_host=127.0.0.1
log file = /var/log/radosgw/client.radosgw.gateway.log
rgw print continue = false
rgw dns name = ceph-s3-gateway.clubmate.rocks
```

## ceph: notwendige Pools und radosgw-Konfiguration

→ pools via rados betrachtet

```
root@ceph-s3-gateway:~# rados df
```

| pool name          | category | KB        | objects | clones | degraded | unfound | rd     | rd KB   | wr     | wr KB    |
|--------------------|----------|-----------|---------|--------|----------|---------|--------|---------|--------|----------|
| .rgw               | -        | 2         | 7       | 0      | 0        | 0       | 67     | 52      | 22     | 8        |
| .rgw.buckets       | -        | 663759    | 86579   | 0      | 0        | 0       | 16028  | 191567  | 537152 | 21754160 |
| .rgw.buckets.index | -        | 0         | 4       | 0      | 0        | 0       | 440122 | 673740  | 422843 | 0        |
| .rgw.control       | -        | 0         | 8       | 0      | 0        | 0       | 0      | 0       | 0      | 0        |
| .rgw.gc            | -        | 0         | 32      | 0      | 0        | 0       | 36962  | 39352   | 49572  | 0        |
| .rgw.root          | -        | 1         | 3       | 0      | 0        | 0       | 194    | 128     | 3      | 3        |
| .users             | -        | 1         | 3       | 0      | 0        | 0       | 14     | 8       | 5      | 4        |
| .users.email       | -        | 1         | 2       | 0      | 0        | 0       | 0      | 0       | 4      | 3        |
| .users.uid         | -        | 1         | 6       | 0      | 0        | 0       | 212    | 199     | 152    | 5        |
| rbd                | -        | 4880353   | 1203    | 0      | 0        | 0       | 16957  | 1593286 | 224602 | 48152302 |
| total used         |          | 19813468  | 87847   |        |          |         |        |         |        |          |
| total avail        |          | 168828656 |         |        |          |         |        |         |        |          |
| total space        |          | 188642124 |         |        |          |         |        |         |        |          |

## ceph: notwendige Pools und radosgw-Konfiguration

→ Apache 2.4 VHost-Konfiguration für radosgw

```
root@ceph-s3-gateway:~# cat /etc/apache2/sites-enabled/rados-simple.conf
<VirtualHost *:80>
    # notwendige DNS Eintraege (inkl. Wildcard)
    ServerName ceph-s3-gateway.clubmate.rocks
    ServerAlias *.ceph-s3-gateway.clubmate.rocks

    ServerAdmin s.seitz@heinlein-support.de
    DocumentRoot /var/www
    RewriteEngine On

    RewriteRule .* - [E=HTTP_AUTHORIZATION:%{HTTP:Authorization},L]

    SetEnv proxy-nokeepalive 1

    ProxyPass / fcgi://127.0.0.1:9000/

    #AllowEncodedSlashes On
    ErrorLog /var/log/apache2/error.log
    CustomLog /var/log/apache2/access.log combined
    ServerSignature Off
</VirtualHost>
```

## ceph: optionaler fscache via rbd

→ ceph und Kernel-rbd ist auf den dovecot-Servern notwendig ...

```
root@ceph-dovecot1:~# rbd ls
fscache-dovecot1
root@ceph-dovecot1:~# rbd info fscache-dovecot1
rbd image 'fscache-dovecot1':
    size 5120 MB in 1280 objects
    order 22 (4096 kB objects)
    block_name_prefix: rb.0.1cedf.238elf29
    format: 1
root@ceph-dovecot1:~# df -h | grep fscache
/dev/rbd0          4,8G   11M  4,6G   1% /var/lib/dovecot/fscache-dovecot1
root@ceph-dovecot1:~# mount | grep fscache
/dev/rbd0 on /var/lib/dovecot/fscache-dovecot1 type ext4 (rw,relatime,stripe=1024,data=ordered)
root@ceph-dovecot1:~# ls -la /var/lib/dovecot/fscache-dovecot1/
total 48
drwxr-xr-x 8 vmail vmail  4096 Mär 18 12:48 .
drwxr-xr-x 5 root  root   4096 Mär 18 12:06 ..
drwx----- 2 vmail vmail  4096 Mär 18 10:25 45
drwx----- 2 vmail vmail  4096 Mär 18 10:25 6b
drwx----- 2 vmail vmail  4096 Mär 18 10:25 6d
drwx----- 2 vmail vmail  4096 Mär 18 12:48 7f
drwx----- 2 vmail vmail  4096 Mär 18 10:16 81
-rw----- 1 vmail vmail   435 Mär 18 12:48 fscache.log
drwx----- 2 root  root  16384 Mär 17 18:22 lost+found
```

## Dovecot: Object-Storage (obox)

```
root@ceph-dovecot1:~# cat /etc/dovecot/conf.d/11-obox.conf
plugin {
  # localstore
  #obox_fs = fscache 10G:/var/lib/dovecot/cache:s3:
    http://USERNAME:PASSWORD@dovecot1.ceph-s3-gateway.clubmate.rocks/

  # rbd create --size <MB> fscache-dovecot1; rbd map fscache-dovecot1
  # mkfs.ext4 /dev/rbd/rbd/fscache-dovecot1
  # mount /dev/rbd/rbd/fscache-dovecot1 /var/lib/dovecot/fscache-dovecot1
  # Bsp-rbd 5GB, Usage 4GB # UID/GID beachten!
  obox_fs = fscache 4G:/var/lib/dovecot/fscache-dovecot1:s3:
    http://USERNAME:PASSWORD@dovecot1.ceph-s3-gateway.clubmate.rocks/
}

service metacache {
  unix_listener metacache {
    user = vmail
  }
}

service metacache-worker {
  user = vmail
}

mail_uid = vmail
```

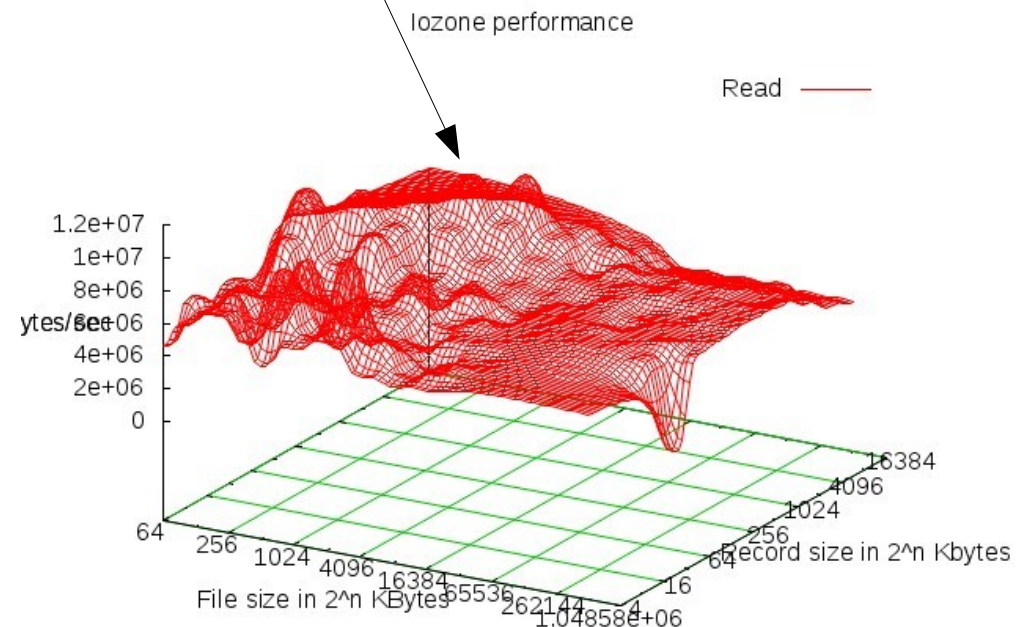
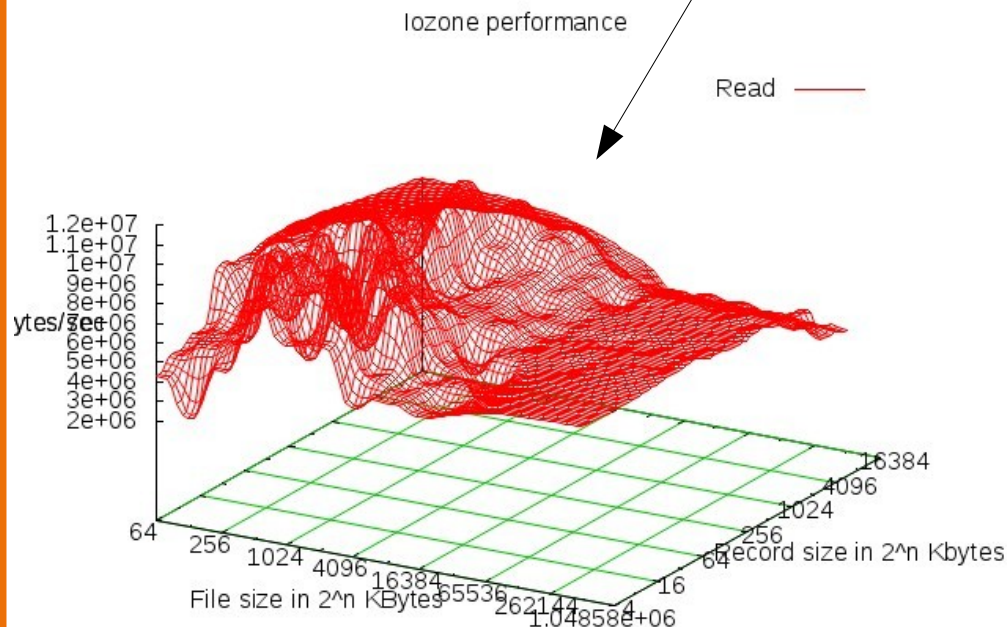


## ceph: optionaler fscache via rbd

- Verwendung von RBD
  - ceph und Kernel-rbd ist auf den dovecot-Servern notwendig
  - Netzwerk-Isolation zwischen ceph/S3 und S3/dovecot ist aufgehoben
  - fscache auf lokalen Disks (evtl. SSD) ist schneller
- Verwendung von lokalen Disks
  - Bei Verlust der Hardware sind Daten der fscache Partitionen verloren
    - Konkret: Mails die noch nicht im Object Storage gespeichert sind

## Performancewerte lokaler vs. RBD fscache

- Messung im Labor-Setup
  - Die Werte sollten nicht absolut betrachtet werden
  - $\Delta$  ist allerdings aussagekräftig

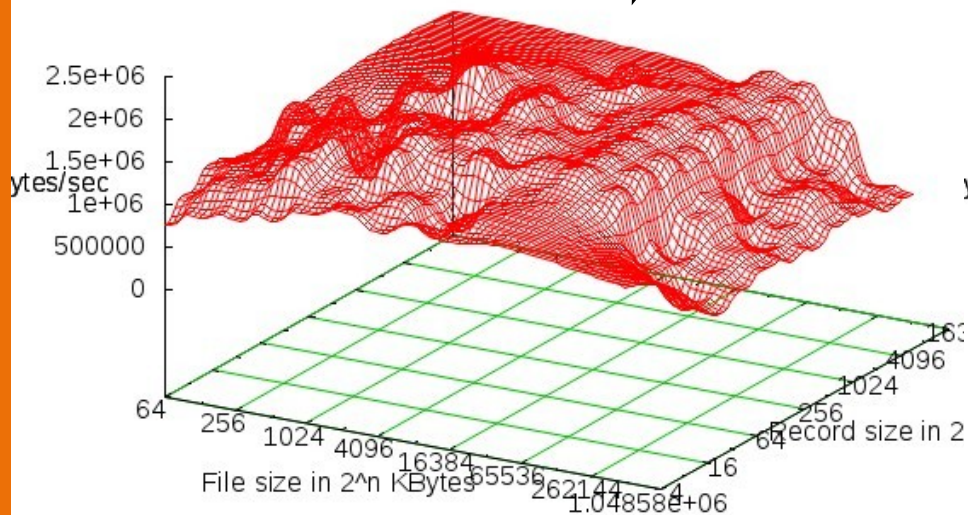


## Performancewerte lokaler vs. RBD fscache

→ Messung im Labor-Setup / write

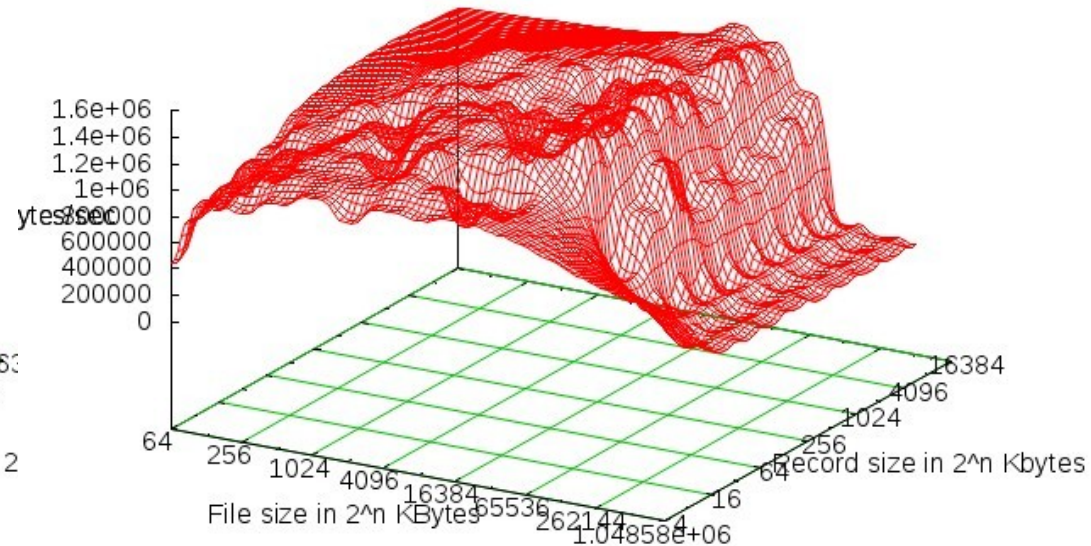
iozone performance

Write —



iozone performance

Write —

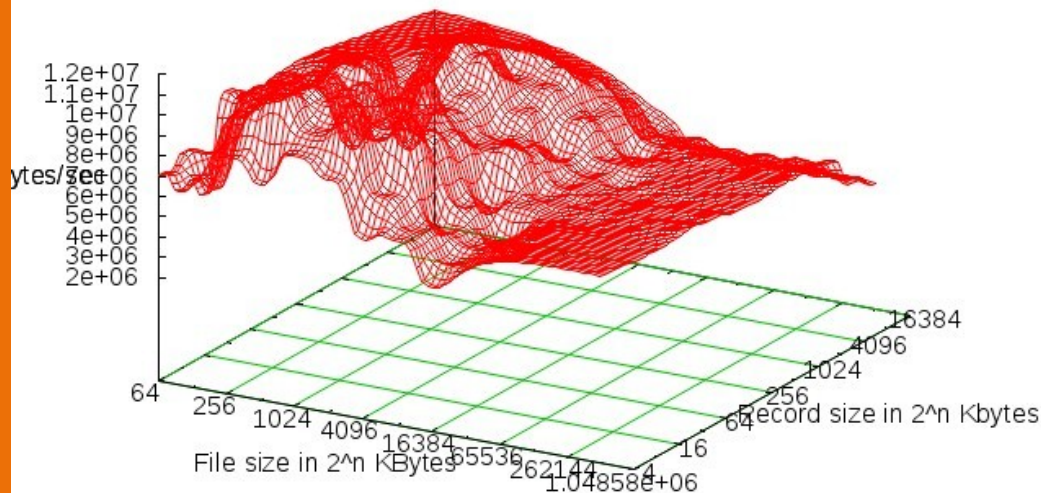


## Performancewerte lokaler vs. RBD fscache

→ Messung im Labor-Setup / reread

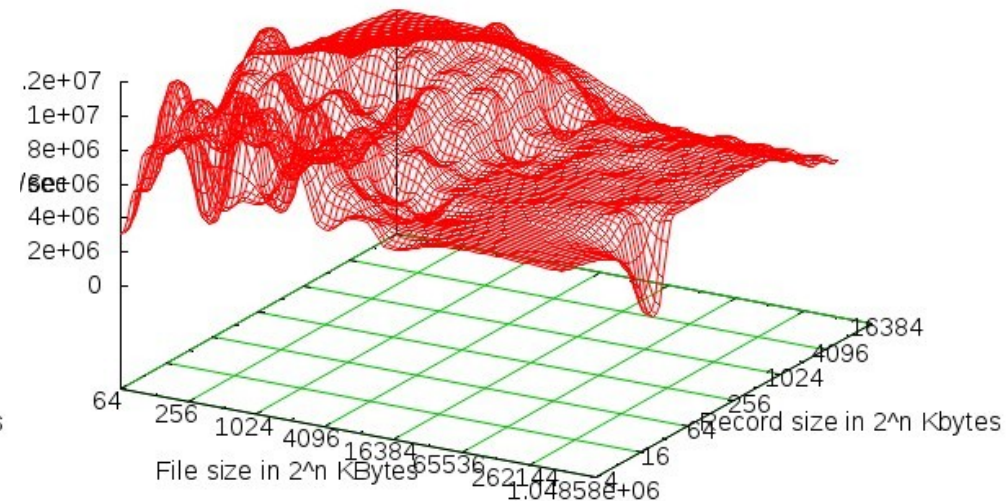
iozone performance

Reread —



iozone performance

Reread —



# Per·for·mance

*Substantiv [die]* KUNST

/pə(r)'fɔ:məns/

von einem Künstler dargebotene künstlerische Aktion.



Übersetzungen, Wortherkunft und weitere Definitionen



## Performancetest

- Virtualisiertes Setup. Die absoluten Werte sind nicht mit Hardware vergleichbar.
- Bulk-Einlieferung über lmtmp

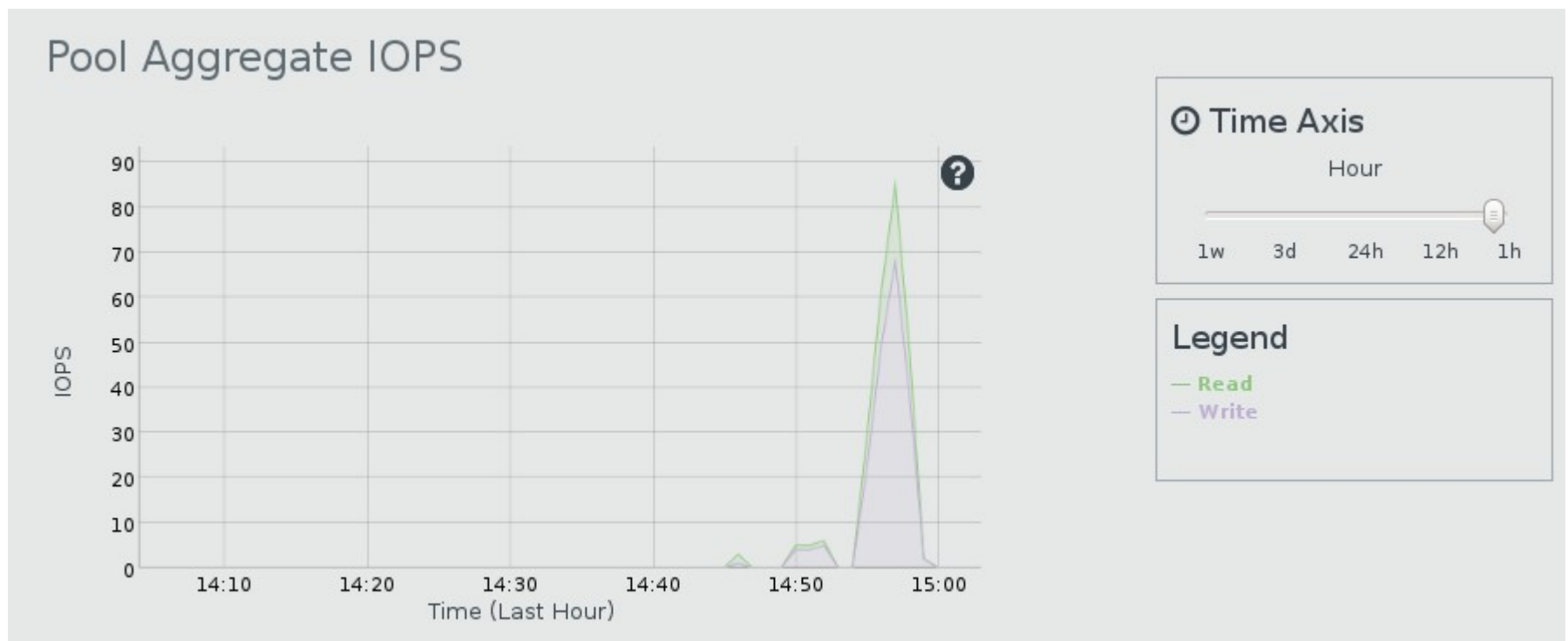
```
root@ceph-dovecot1:~# benchmark-dovecot 500
500

real 2m53.773s
user 0m0.024s
sys  0m0.032s

root@ceph-dovecot1:~# tail -n 5 /usr/local/sbin/benchmark-dovecot
time smtp-source -L -s 20 -l 5120 \
    -m $1 -c \
    -f sender@example.com \
    -t $recipient@clubmate.rocks \
    unix:/var/spool/postfix/private/dovecot-lmtmp
```

## Performancetest

- Virtualisiertes Setup. Die absoluten Werte sind nicht mit Hardware vergleichbar.
- Bulk-Einlieferung über Imtp

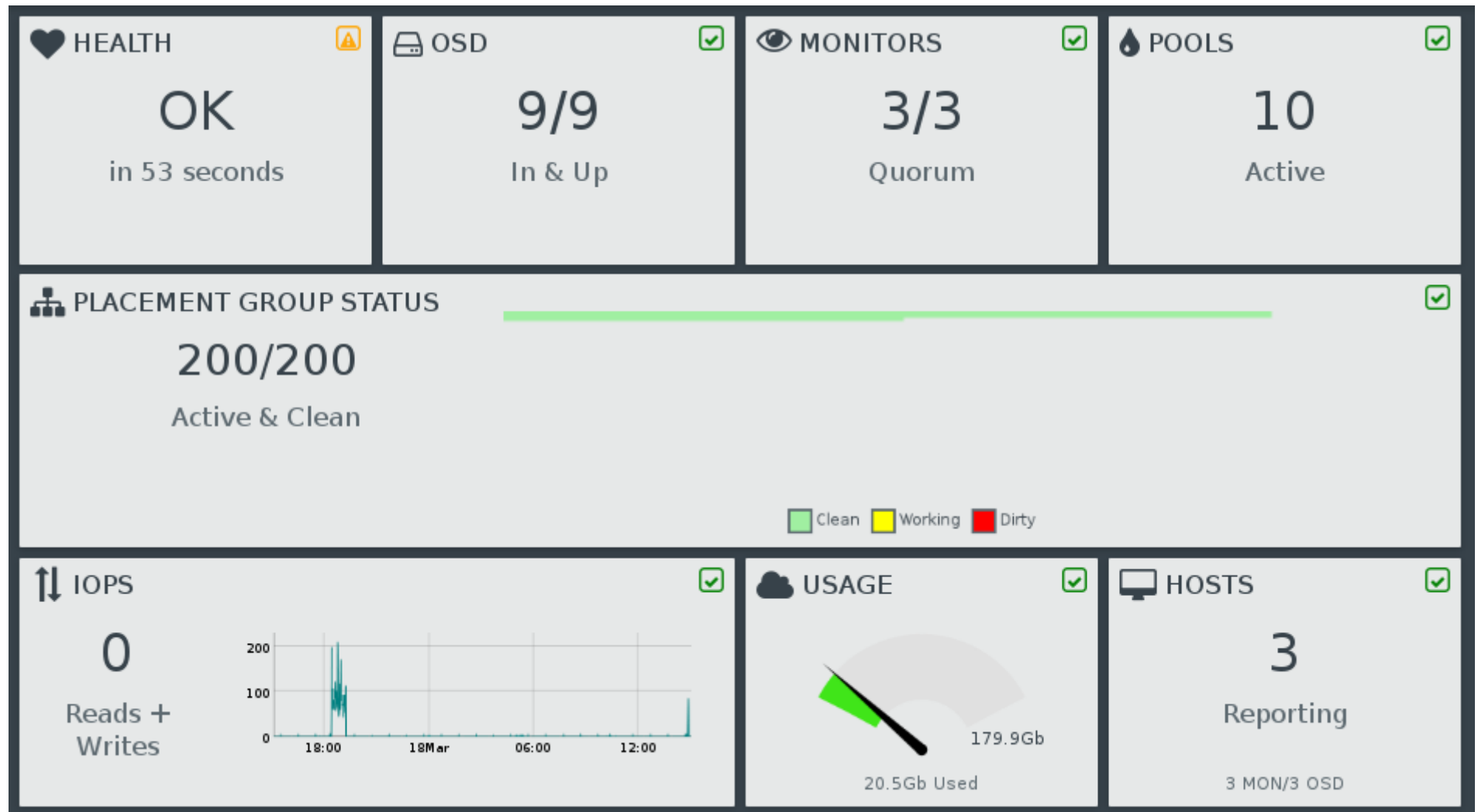


# **Hands On!**

## **Live View**







Wir suchen neue Kollegen für:

Helpdesk, Administration, Consulting!

Wir bieten:

Spannende Projekte, Kundenlob, eigenständige Arbeit,  
keine Überstunden, Teamarbeit

...und natürlich: Linux, Linux, Linux...

<http://www.heinlein-support.de/jobs>

## Quellen

- Dovecot
  - Dovecot EE - Object Storage and Enterprise Repository Installation manual
  - <http://wiki2.dovecot.org/>
  - <http://www.dovecot.org/talks/>
    - Timo Sirainen - Dovecot Recent and Future Development / SLAC 2014
- ceph
  - <http://docs.ceph.com/docs/master/radosgw/config/>
  - <http://docs.ceph.com/docs/master/radosgw/s3/>

# Heinlein Support hilft bei allen Fragen rund um Linux-Server

## HEINLEIN AKADEMIE

Von Profis für Profis: Wir vermitteln die oberen 10% Wissen: geballtes Wissen und umfangreiche Praxiserfahrung.

## HEINLEIN HOSTING

Individuelles Business-Hosting mit perfekter Maintenance durch unsere Profis. Sicherheit und Verfügbarkeit stehen an erster Stelle.

## HEINLEIN CONSULTING

Das Backup für Ihre Linux-Administration: LPIC-2-Profis lösen im CompetenceCall Notfälle, auch in SLAs mit 24/7-Verfügbarkeit.

## HEINLEIN ELEMENTS

Hard- und Software-Appliances und speziell für den Serverbetrieb konzipierte Software rund ums Thema eMail.