High Availability, Loadbalancing und Scale-Out in Mailclustern

High Availability und Loadbalancing

- High Availability ist? *Redundanz in Form von*
  - Master / Slave
  - Master / Master

- Load Balancing ist? *Verteilung der Anfragen über mehrere Knoten*
  - DNS (z.B. multiple DNS Records)
  - Level 2 / Level 3 Loadbalancer
  - Proxy Server

Scale-Out oder doch Scale-Up?

- Scale-... ist?
  - *Kapazitätserweiterung durch zusätzliche Ressourcen*

- Scale-Up
  - Mehr CPU-Kerne, mehr RAM, mehr Netzwerkbandbreite
  - Skalieren auch die Prozesse mit?
  - Erhöhung der Redundanz sowie der HA ist so schwierig

- Scale-Out
  - Mehr Instanzen
  - Gute Lastverteilung muss sicher gestellt werden

- *Heutezutage wollen wir ganz sicher Scale-Out*

Mailcluster brauchen gar keine Loadbalancer!

→ OK, an einer Stelle doch :(

Dienste eines Mailclusters

- Mail Transfer Agent (Postfix, Exim...)
- Mail Delivery Agent (Dovecot ...)
- Datenbanken (SQL, LDAP, Redis ...)
- DNS (Bind, PowerDNS, Unbound)
- Content-Filter (Amavis/Spamassassin, Rspamd)
  - + Zusatztools: RBL, Razor, Pyzor, DCC ...
  - + kommerzielle Appliances
- Anti-Virus (ClamAV, kommerzielle AV)
- Webmailer / Webserver (Apache, Nginx ...)
- Proxy / Loadbalancer (HA-Proxy, Apache, Nginx, keepalived ...)
- Storage (local, NFS, SCSI, S3

DNS

- High Availability und Loadbalancing by Design
  - Mehrere NS Server

- Kennt Ihr den schon? PowerDNS
  - Datenbank-Backends, kann Bind Zonen nutzen, HTTP-API, Supermasters, einfaches DNSSec
  - Supermasters:
    - Slave Akzeptiert jede neue Domain, welche vom Supermaster kommt
    - slac2020.de auf Master anlegen
    - SLAVE als NS aufnehmen
    - Slave transferiert die Zone sowie der Master ein notify schickt

PowerDNS

- DNSSec mit Autosign aktivieren:

```
pdnsutil secure-zone slac2018.de

pdnsutil add-zone-key slac2018.de ksk 2048 active rsasha256
pdnsutil add-zone-key slac2018.de zsk 1024 active rsasha256
```

Dnsdist - Loadbalancer für DNS

- Kann Loadbalancing, Routing, Firewalling, Query Limiting

```
setLocal('10.1.1.12:53')
setACL({'0.0.0.0/0', '::/0'}) -- Allow all IPs access

newServer({address='10.1.1.36:5301', pool='recursor', order=2})
newServer({address='10.1.1.151:5301', pool='recursor', order=1})
setServerPolicy(firstAvailable) -- first server within its QPS limit

newServer({address='10.1.1.36:53', pool='auth', order=1})
newServer({address='10.1.1.151:53', pool='auth', order=2})
setServerPolicy(firstAvailable) -- first server within its QPS limit

recursive_ips = newNMG()
recursive_ips:addMask('10.1.1.0/24') -- These network masks are the ones from
allow-recursion in the Authoritative Server

addAction({'slac2018.de.'}, PoolAction("auth"))
addAction({'1.1.10.in-addr.arpa.'}, PoolAction("auth"))
addAction(NetmaskGroupRule(recursive_ips), PoolAction('recursor'))
```

SQL / LDAP & Co - Datenbank & Value-Store

- Master / Slave und Master / Master Replikation möglich

- Mysql
  - Galera
  - Mysql Group Replication
  - Mysql Binlog Master / Slave

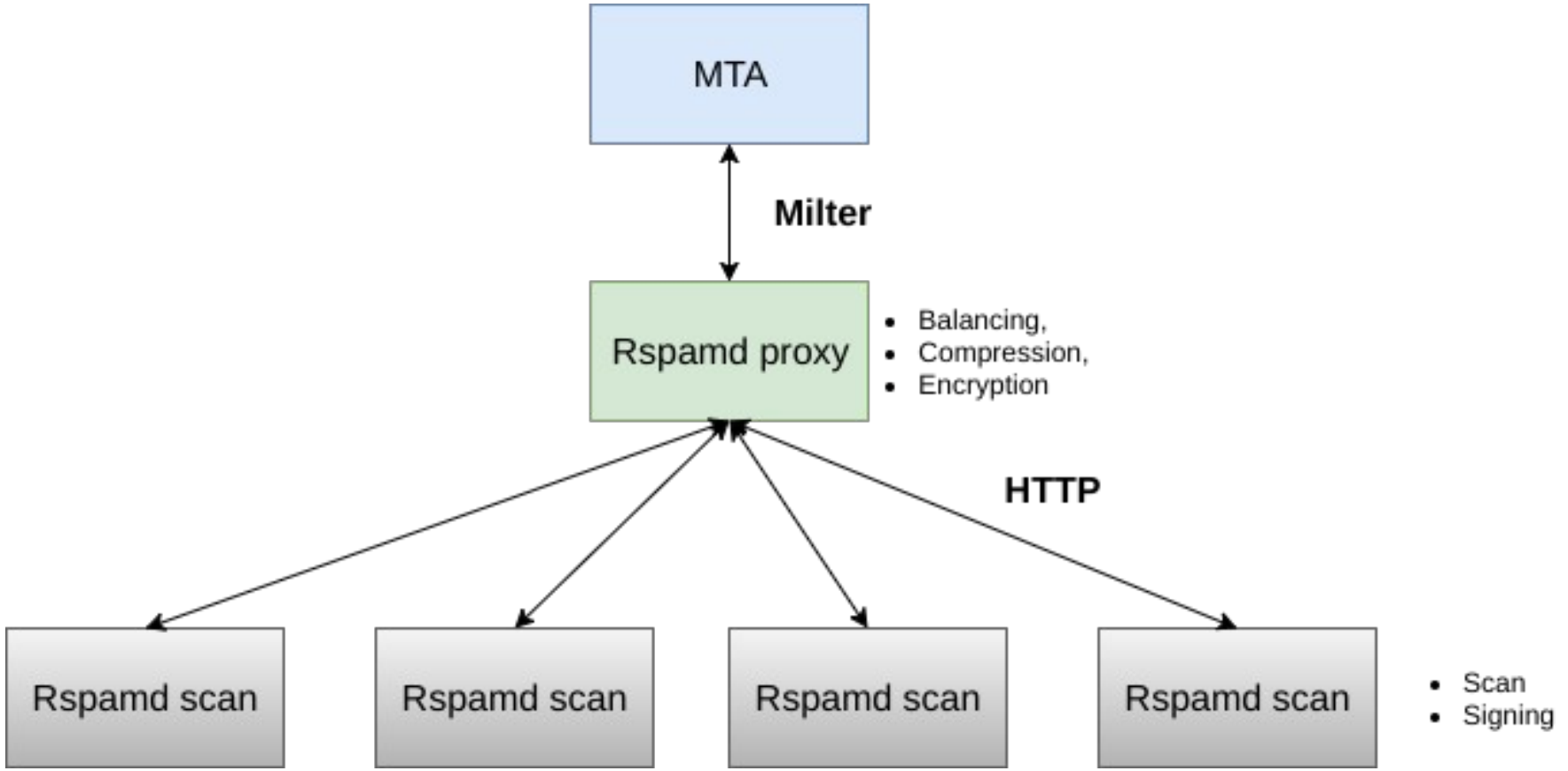
- OpenLdap
  - Multi-Master Replikation

- Replikation auch verfügbar für Redis, Memcache ...

Rspamd - Upstreams

- Minimaler Rspamd Proxy direkt auf den MTA's
  - Absturz erzeugt nur einen 4xx Fehler im Postfix
- Rspamd Upstreams Implementierung
  - Rspamd monitored seine Upstreams selbstständig
  - Host-Liste mit Gewichtung
  - Loadbalancing Schema: round-robin, master-slave ...
  - Wo kommt es zum Einsatz:
    - Rspamd Proxy → Rspamd Worker
    - Rspamd Worker → Antivirus Server
    - Rspamd Worker → Redis Server
    - ...

Rspamd - Multiserver



Webmailer

- Typischerweise mit vorgeschaltetem HTTP-Proxy
  - HA-Proxy, Apache, nginx ...
- Session – Replikation ist ein Problem
  - Kann z.B. in Redis erledigt werden
  - OpenXchange integriert Hazelcast dafür

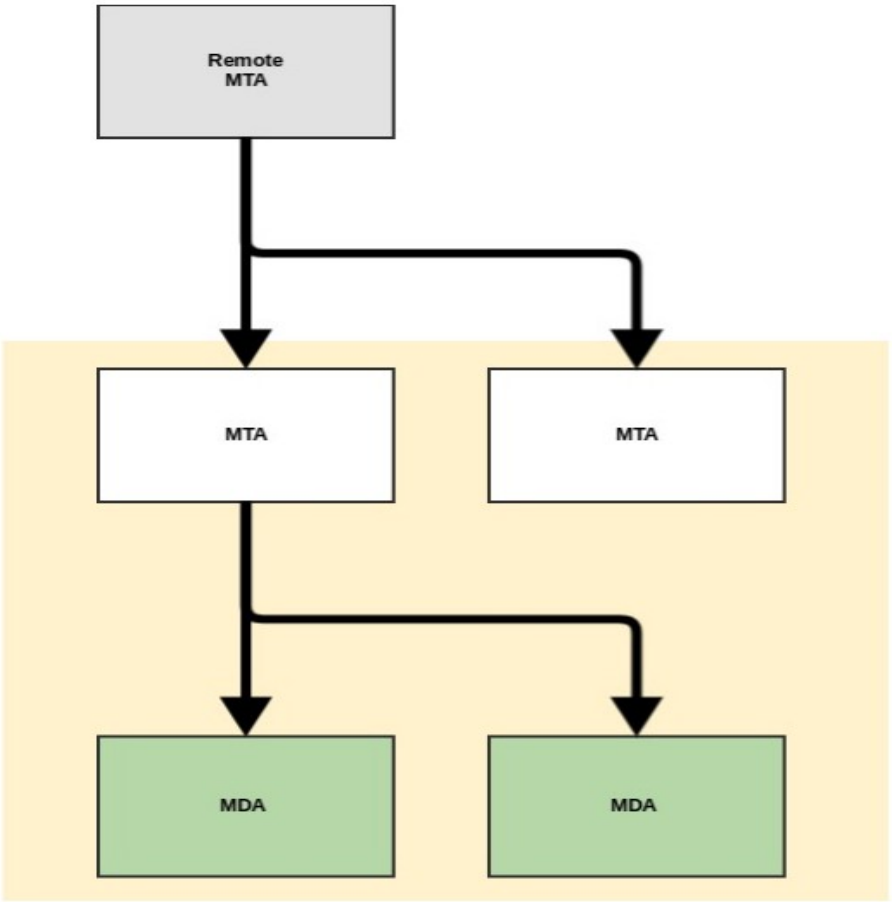
MTA Postfix – HA & LB by Design

- Bei SMTP ist die Redundanz und die Lastverteilung im Protokoll bzw in den RFC verankert
- Alle 4xx Fehler sind temporär und der einliefernde MTA muss es neu versuchen
- Typischerweise wird sofort ein weiterer MX probiert
- Das kann man auch intern verwenden
  - SMTP-AUTH → SMTP-OUT Server
  - MX → SMTP-Intern
- LMTP definiert das leider nicht – keine MX Records ;(
  - ABER …

Postfix - HA & LB by Design - auch bei LMTP!

```
cat /etc/postfix/relay_domains
   slac2018.de   lmtp:director.slac2018.de

dig director.slac2018.de +short
   10.1.1.133
   10.1.1.209

Mai 06 15:55:10 mx1.slac2018.de postfix/lmtp[55159]: connect to
director.slac2018.de[10.1.1.133]:24: Connection refused
Mai 06 15:55:10 mx1.slac2018.de postfix/lmtp[55159]: connect to
director.slac2018.de[10.1.1.209]:24: Connection refused

lmtp_fallback_relay = dir-1.slac2018.de:24,dir-2.slac2018.de:24

Mai 06 15:55:20 mx1.slac2018.de postfix/lmtp[55159]: connect to dir-
1.slac2018.de[10.1.1.209]:24: Connection refused
Mai 06 15:55:20 mx1.slac2018.de postfix/lmtp[55159]: connect to dir-
2.slac2018.de[10.1.1.133]:24: Connection refused
```

Postfix - HA & LB by Design



MDA - Dovecot

- IMAP, POP3, LMTP haben keine Redundanz per Design
  - DNS nicht praktikabel (zu langsam)
  - Loadbalancer
- Mailboxen:
  - Gemeinsamer Speicher
    - Teilweise problematisch (Indizes)
  - Replikation
- Frontend: eigener Loadbalancer Dovecot Director

Dovecot Director

- Director ist ein Level-7 LB bzw. Proxy für IMAP, POP3, LMTP, Sieve und neu auch SMTP
- Kann User authentifizieren
- Aber besser noch routen an Hand von Eigenschaften (Sharding)
- Mehrere Direktoren mit mehreren Backends möglich
- Direktoren bilden einen Ring und tauschen Informationen aus
- Alle Verbindungen eines Users (Desktop, Mobile) landen immer auf dem gleichen Backend
- Backends können auch nicht Dovecot Server sein: Cyrus, Courier, Exchange, Notes ...
- Backends werden überwacht

Dovecot Datenbank Anbindung

→ Dovecot kann mehrere Datenbanken nacheinander abfragen

```
passdb {
  driver = sql
  args = /etc/dovecot/dovecot-sql.conf.ext
}
passdb {
  driver = sql
  args = /etc/dovecot/dovecot-sql2.conf.ext
}

  # v2.2.10+:
  skip = never
  result_failure = continue
  result_internalfail = continue
  result_success = return-ok
```

https://wiki.dovecot.org/Authentication/MultipleDatabases
https://wiki.dovecot.org/UserDatabase

Scale Out - MTA, MDA, Content-Filter

- Postfix
  - Wenn reines MX Routing: MX und SPF anpassen, fertig
  - Direkte Datenabfragen und Umschreibungen in Postfix eleminieren
  - Postfix verify & Aliase in Dovecot auflösen

- Dovecot
  - Nutzung der Direktoren + eventuell HA-Proxy
  - HA-User-DB bzw mehrere Datenbanken anbinden
  - Backends bei Replikation 1-fach redundant
  - Backends bei shared Storage n-fach redundant

- Rspamd & Tools
  - Backends wegen Upstreams n-fach redundant
  - Rspamd-proxy (auf localhost) SpoF aber Postfix ist ja redundant
  - Alle von rspamd genutzten Tools lassen sich n-fach auslegen

Dovecot – shared Storage

- Shared NFS Storage kann Probleme verursachen (fsync, defekte Indizes ...)
- Redundantes NFS ist kompliziert oder teuer und kann die o.g. Probleme gar nicht lösen
- Replication immer genau zwischen 2 Servern ohne shared Storage
  - Migration ist aufwendig
- S3 (Objektspeicheranbindung) nur in Dovecot Pro - obox
- Neuer Ansatz – gesponsert von der Telekom:
  - Natives Ceph Plugin
  - Noch in Entwicklung und noch nicht Feature Complete

Was ist dieses Ceph?

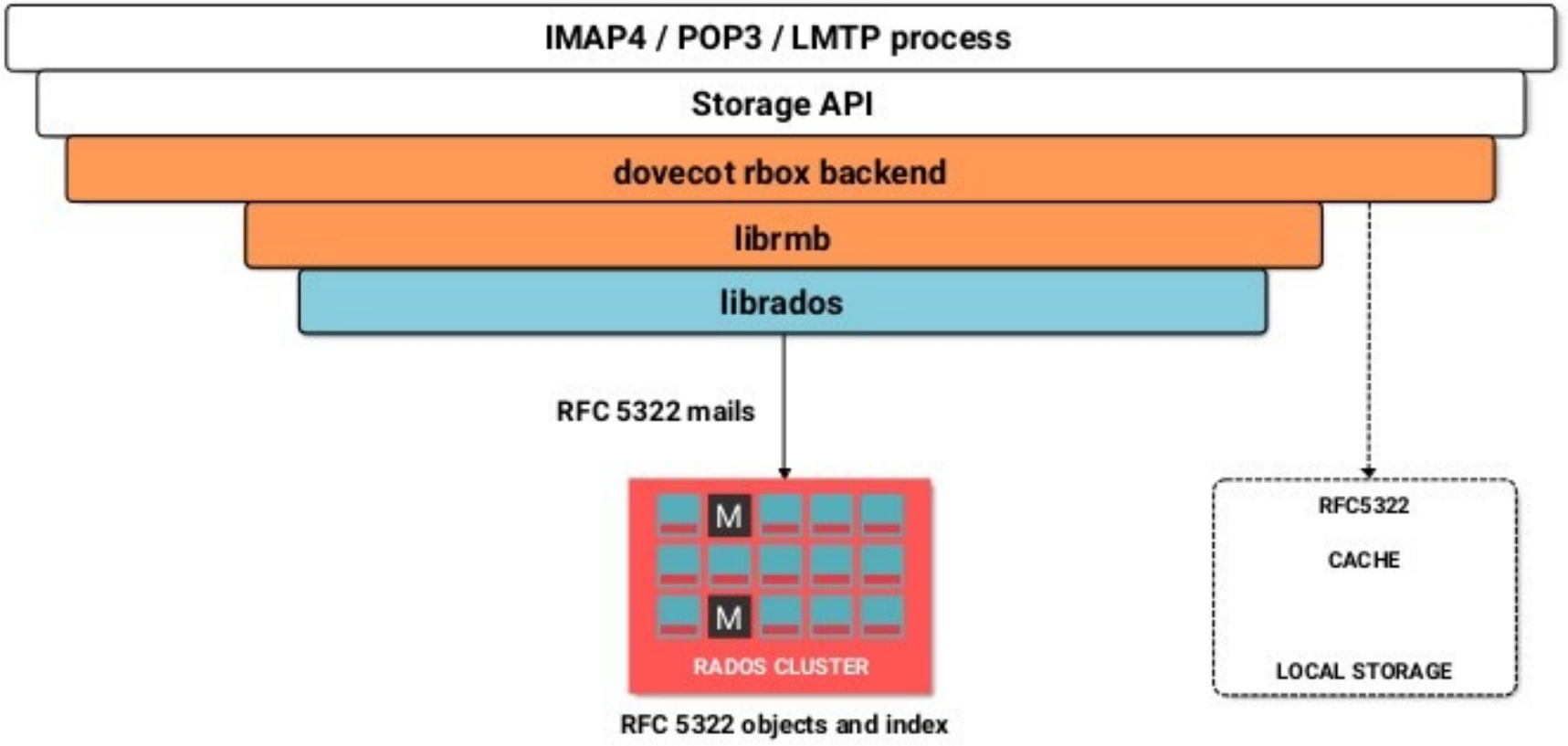
- Objektspeicher
  - Ich übergebe dem System ein Objekt (Datei) und es kümmert sich selbst um Speicherung und nachweisbare Redundanz
  - Es wird kein RAID, Fiberchannel, iSCSI ... benötigt
  - Ceph Monitore sorgen für Lastverteilung der Anbindung
  - Ceph verwaltet einzelne Festplatten
  - Konfiguration und Algorithmus entscheiden wie oft und auf welcher Festplatte ein Objekt gespeichert wird
  - Kümmert sich automatisch beim Hinzufügen und beim Ausfall eines Systems um die Reorganisation aller Daten
  - Backup? Ceph kann Snapshots und die Änderungen auf ein anderes Ceph System übertragen

Ceph + Dovecot

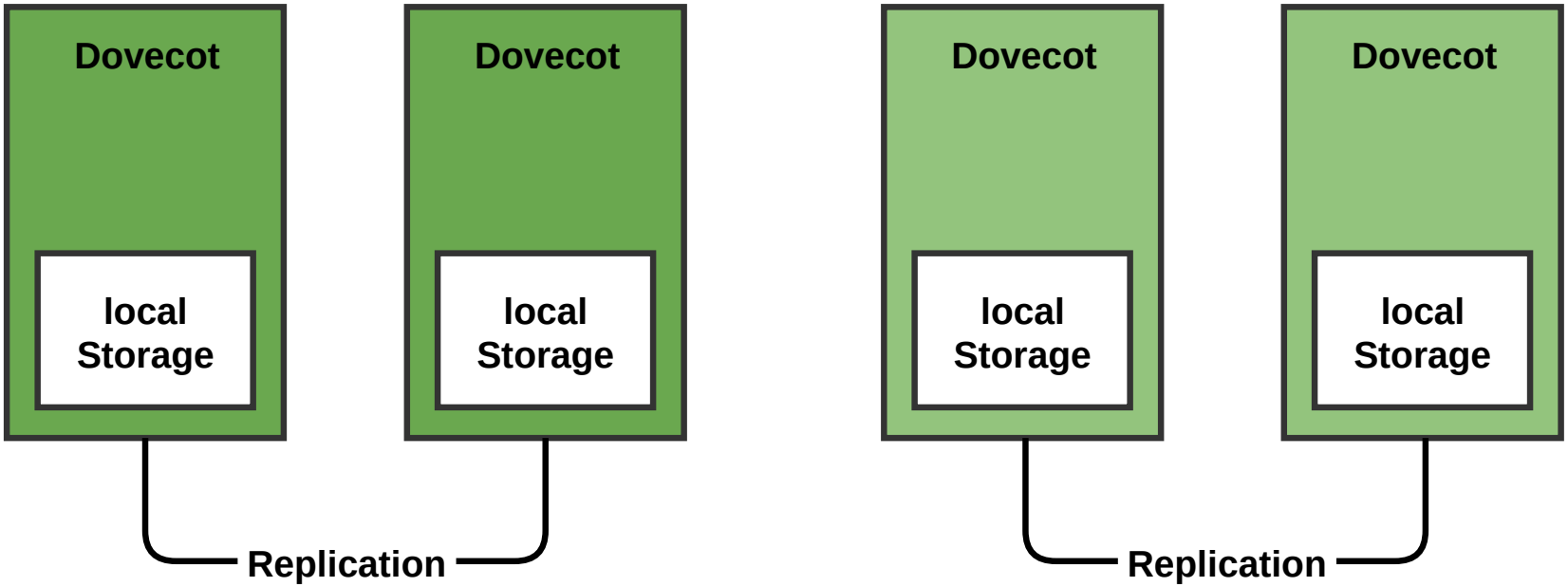
- https://github.com/ceph-dovecot/dovecot-ceph-plugin
- Hybrider Entwicklungsansatz
  - Erweiterungen sowohl für Ceph als auch Dovecot
  - Mails werden als Objekte direkt über Rados abgelegt
  - Dovecot Dicts im Ceph Rados omap key/value
  - Dovecot Index *noch* auf CephFS
    - CephFS kann man ungefähr mit redundantem NFS beschreiben
- Erste Version soll im Herbst fertig sein
- Projekt gesponsert von der Telekom - Wie viele Postfächer hat die Telekom? Was könnte sie einsparen?

Further Development

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

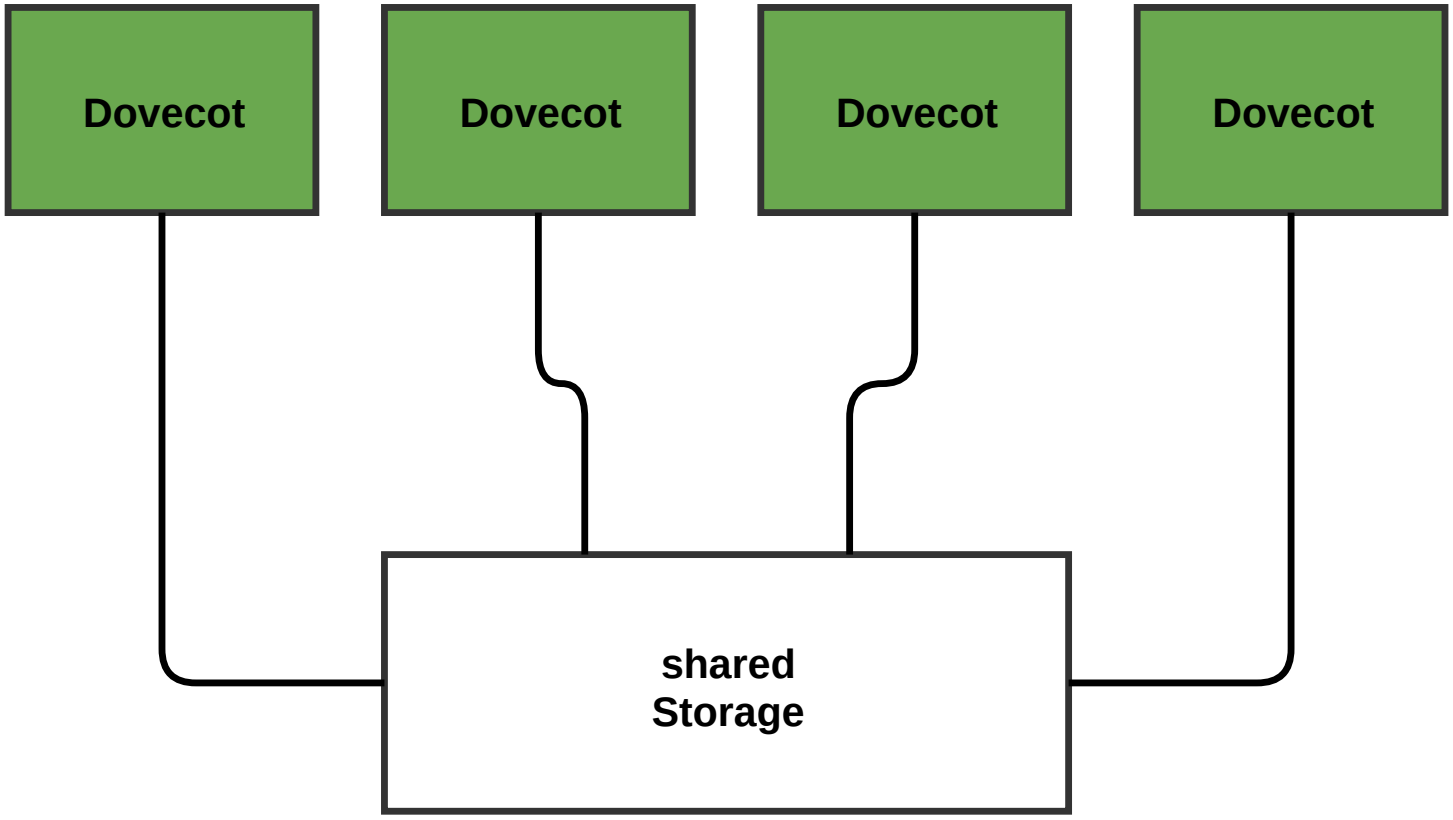


Dovecot - local Storage & Replication



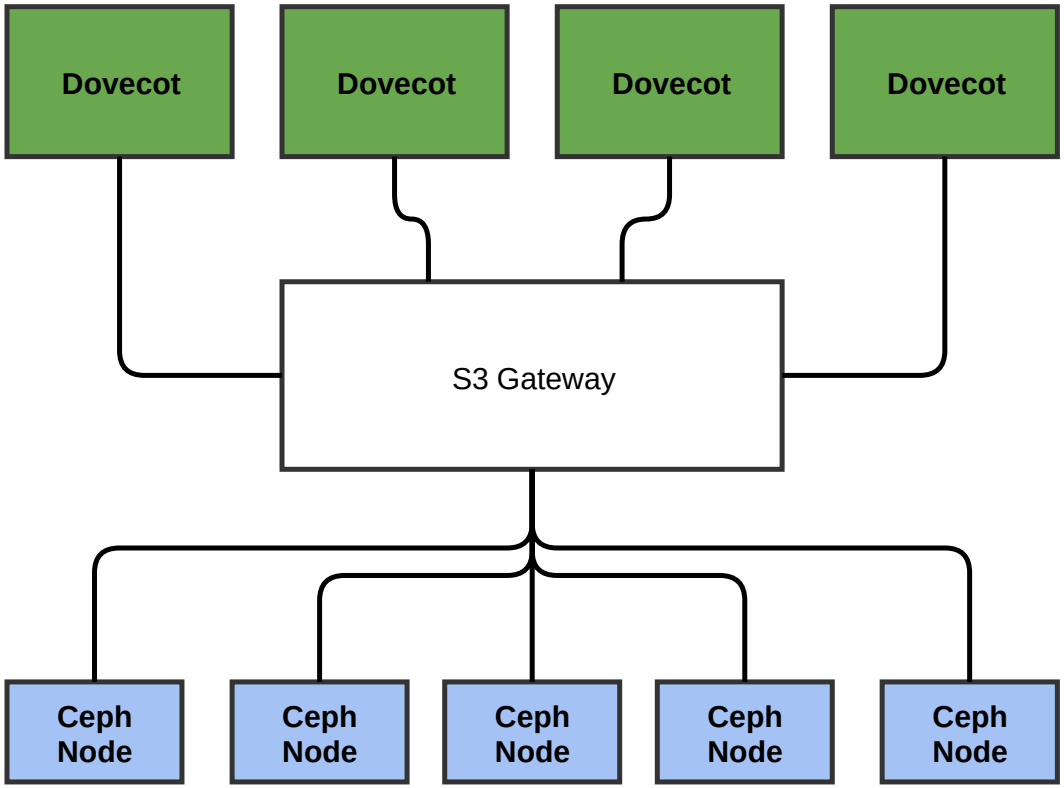
- Mailreplikation + Datenredundanz per Design
- Replikation geht immer nur zwischen 2 Servern
- Kompliziertes Scale-Out und manuelle Migration

Dovecot - shared Storage - NFS



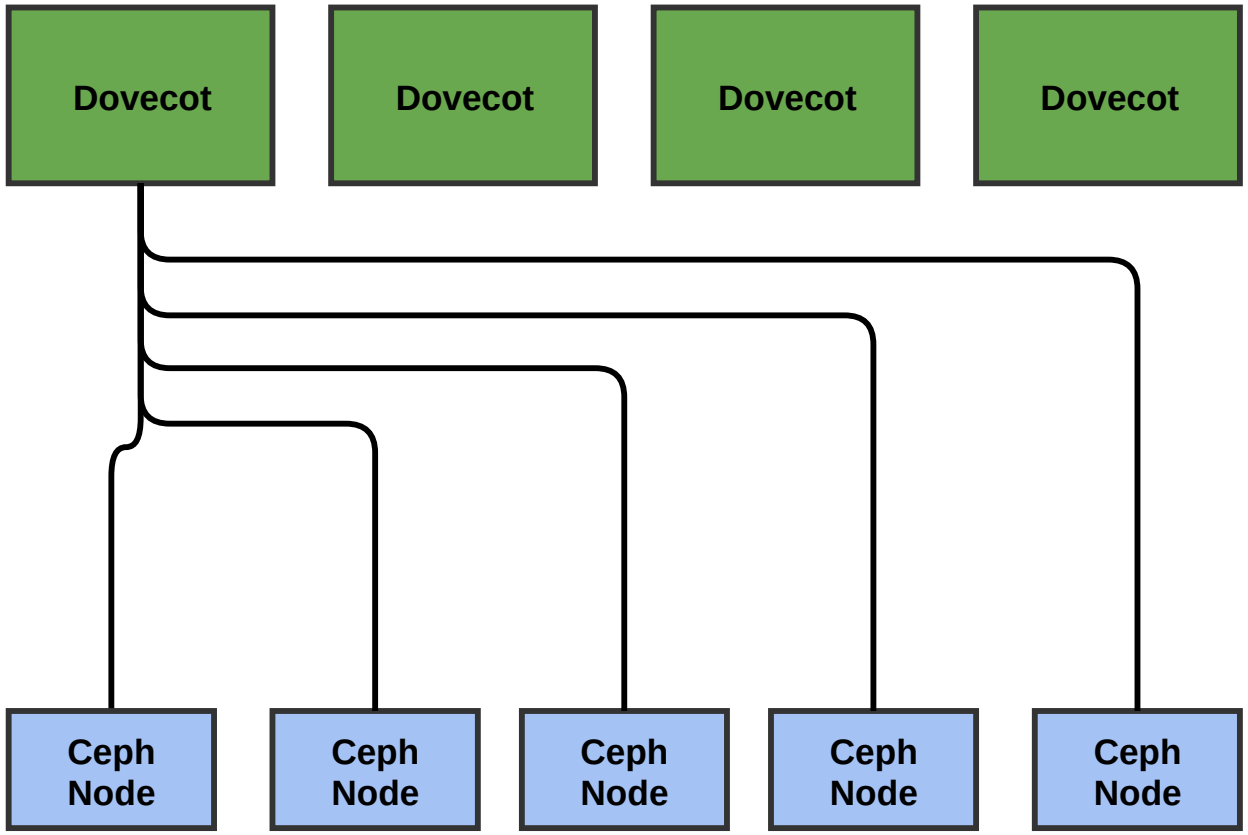
- → Shared Storage skaliert nicht gut mit
- → Probleme mit Dovecot Indexes

Dovecot PRO – Obox (S3) + Ceph



- Ceph und Dovecot schön skalierbar
- S3 Gateway(s) sind der Flaschenhals
- Index + MetaCache werden lokal zwischengespeichert

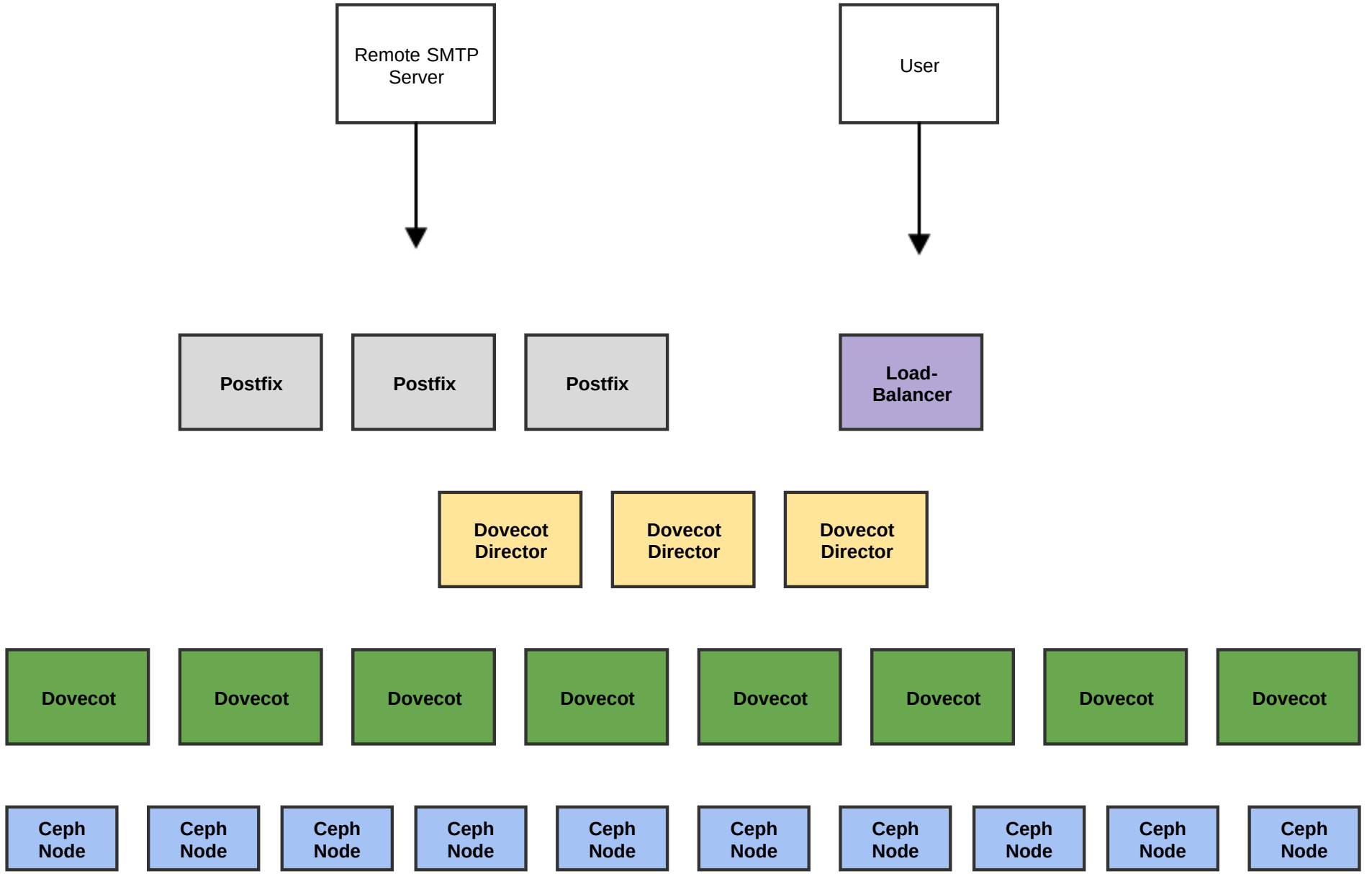
Dovecot Ceph Plugin – Ceph direkt



- Dovecot holt die Daten von den Ceph Nodes direkt ab
- Sehr gute Skalierung bei Ceph und Dovecot
- Index und Metadata direkt im Ceph, lokaler MailCache

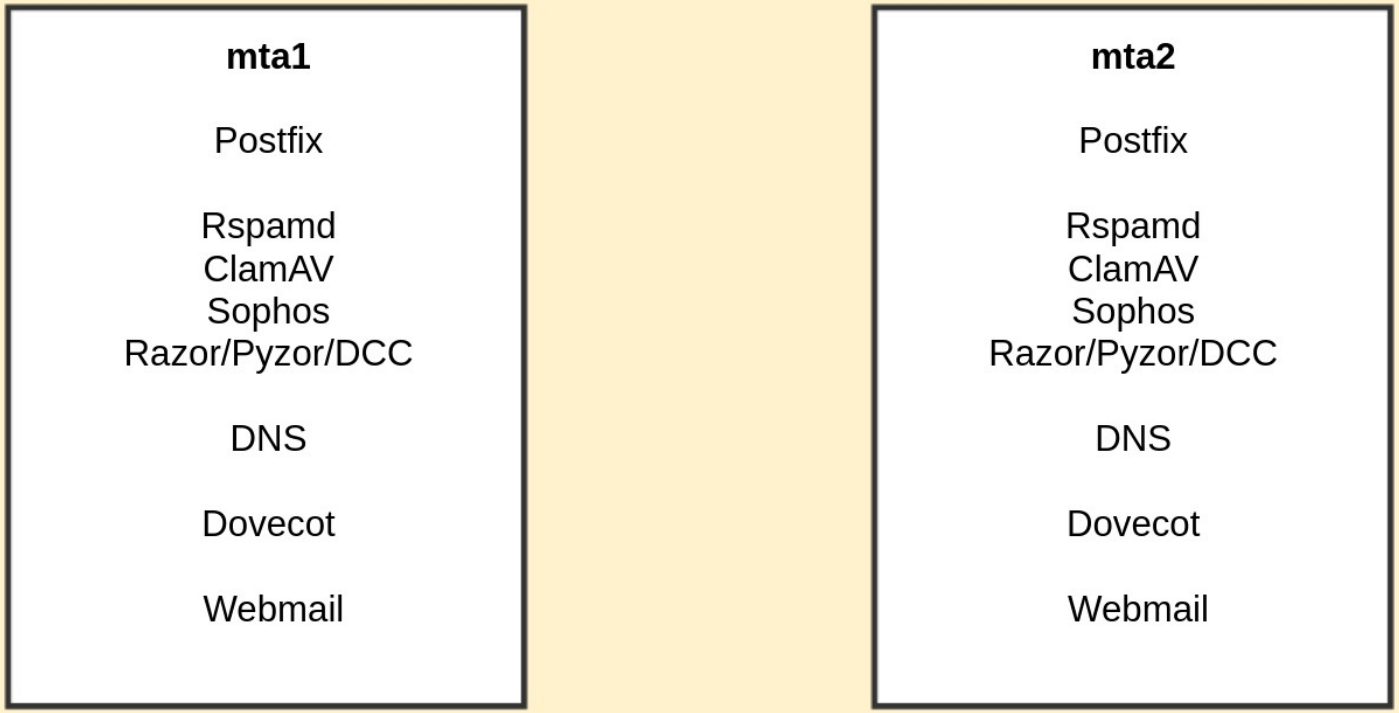
Ceph + Dovecot

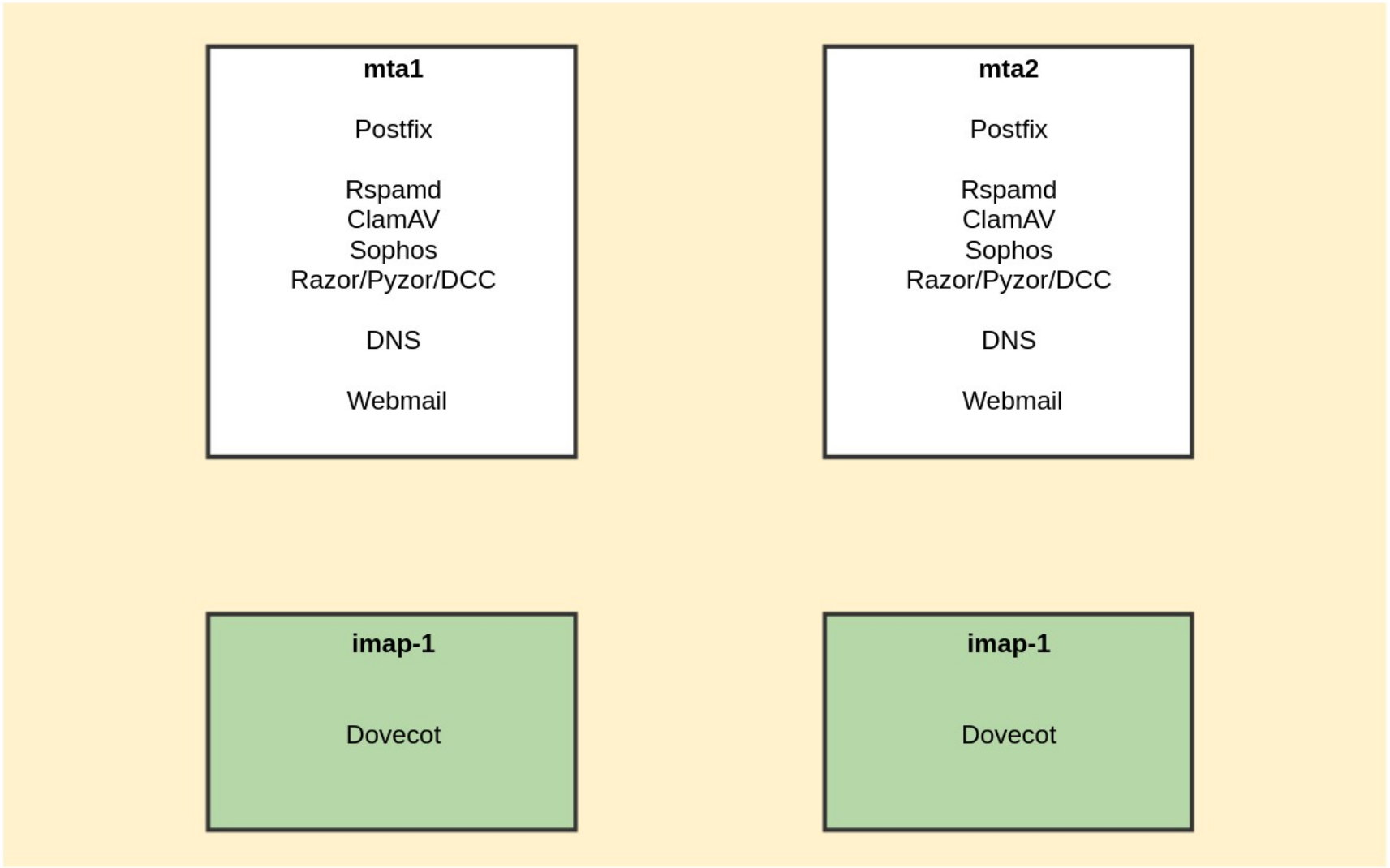
- Dovecot-Ceph hebt die Notwendigkeit von Replikation und Sharding wieder auf
- Läßt die Dovecot Backends wunderbar und ohne Migration in die Breite skalieren
- Dovecot (Pro) und OpenXchange arbeiten wohl an eigenen Implementierungen für Ceph
- Dovecots obox (S3) soll grundlegend freigegeben werden. Optimierungen (für Enterprise) bleiben aber im Pro


Remote SMTP Server
User
Postfix
Postfix
Postfix
Load-Balancer
Dovecot Director
Dovecot Director
Dovecot Director
Dovecot
Dovecot
Dovecot
Dovecot
Dovecot
Dovecot
Dovecot
Dovecot
Ceph Node
Ceph Node
Ceph Node
Ceph Node
Ceph Node
Ceph Node
Ceph Node
Ceph Node
Ceph Node
Ceph Node

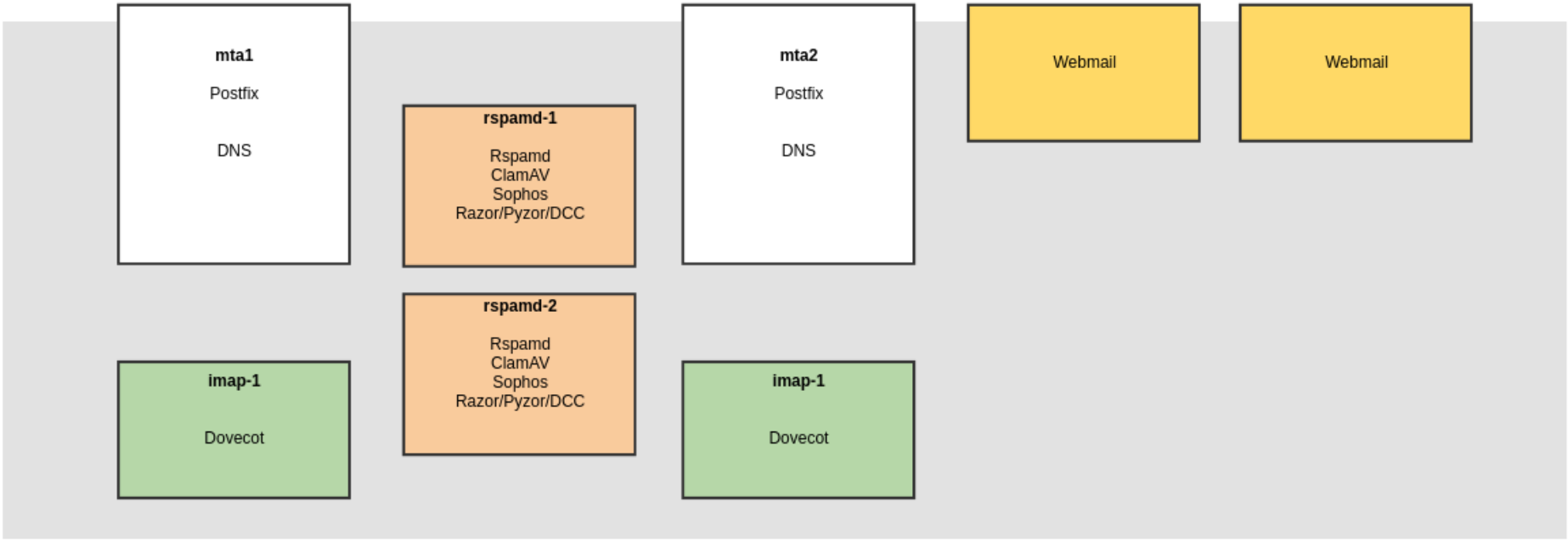
Jeder fängt mal klein an.

Jeder braucht irgendwann einen Partner.

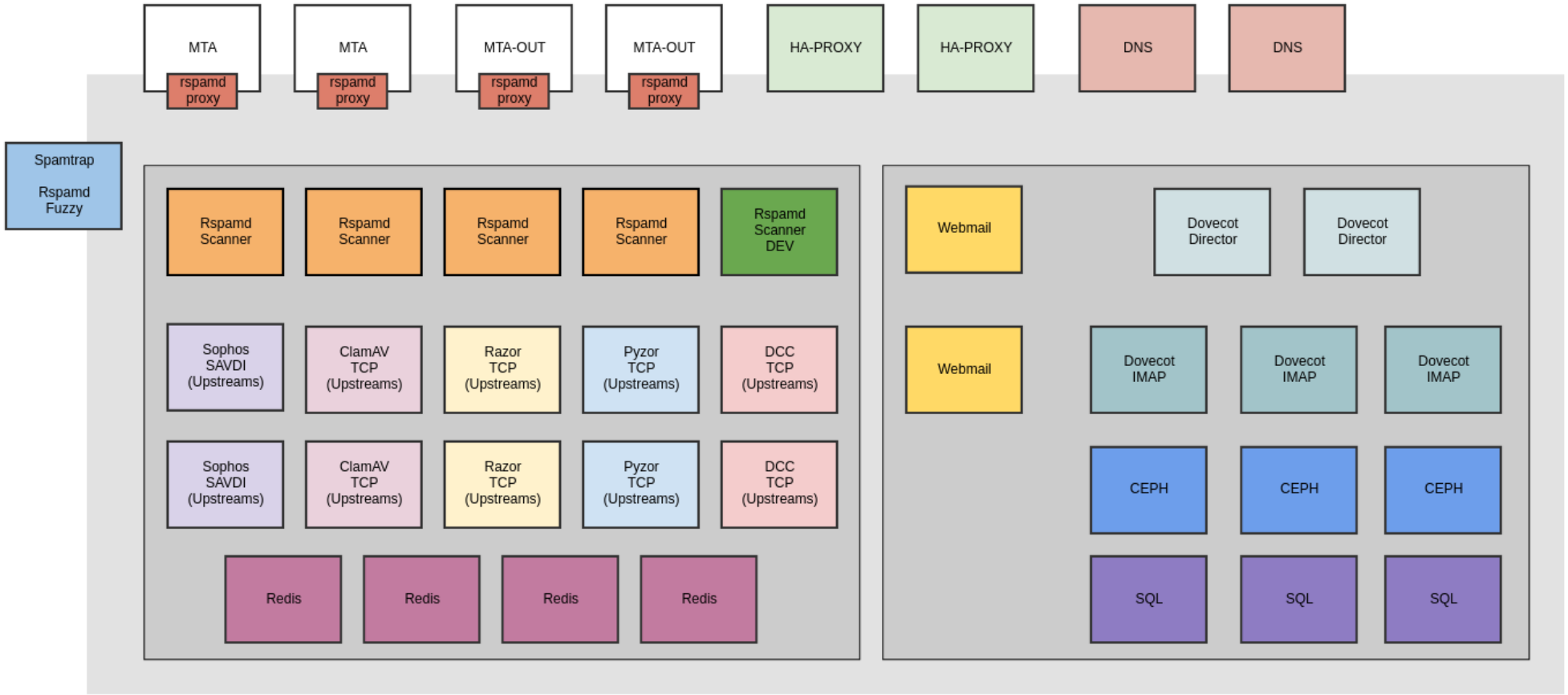



mta1
Postfix
Rspamd
ClamAV
Sophos
Razor/Pyzor/DCC
DNS
Webmail
mta2
Postfix
Rspamd
ClamAV
Sophos
Razor/Pyzor/DCC
DNS
Webmail
imap-1
Dovecot
imap-1
Dovecot

Die Familie wächst.



Jetzt ist der Zoo komplett ;)



Summary

- Auftrennung der Systeme nach Funktionen
  - Leichtgewichtige VMs oder Container
  - Steal Time??? (!)
  - https://www.heinlein-support.de/sites/default/files/SLAC%202018_virtuell-ne-physikalisch.pdf
- Trennung von Nutzdaten und Services
  - LB und HA von Storage und Diensten lassen sich einzeln skalieren
- Redundanz schaffen (nicht immer gleich der teure Loadbalancer)
  - SMTP: MX-Routing
  - LMTP: Multiple A/AAAA Records oder lmtp_fallback ...
  - Datenbanken: mehrere Replica anbinden oder LB
  - Content Filter: z.B. Rspamd Upstreams
  - Storage: HA-NFS, Ceph, (S3)

Summary #2

- Automatisierung
  - Erstellung, Provisionierung und Löschen von Systemen
  - Updates und Konfigurationsmanagement
  - Einfaches Scale-Out auch bei vielen Maschinen
  - Beziehungen zwischen Diensten (IPs, Ports ...) leichter und konsistent managebar
  - Langwierige manuelle Arbeiten vermeiden (Hostnamen, IP, Config, DNS, Firewall ...)

- Bei Ärger wegschmeißen und neu machen
- Bei Performanceproblemen → Breitenskalierung durch zusätzliche Nodes

Next?

- Kubernetes / Container oder Hybrid??
  - Automatische Skalierung und Lastverteilung auf Bare Metal

- Metrikauswertung / APM / Alerting
  - Dynamische Anpassung der benötigten Ressourcen
  - Eine Rspamd Maschine rejected gar keinen Spam mehr
  - Scheinbar neue Spamwelle (Monday 3am ???)
  - Löschen von Systemen wenn nicht genügend Last da ist
  - IP auf einer Blacklist?

- Natürlich und gerne stehe ich Ihnen jederzeit mit Rat und Tat zur Verfügung und freue mich auf neue Kontakte.
  - Carsten Rosenberg
  - Mail: c.rosenberg@heinlein-support.de
  - Telefon: 030/40 50 51 – 46
- Wenn's brennt:
  - Heinlein Support 24/7 Notfall-Hotline: 030/40 505 - 110

Soweit, so gut.

Gleich sind Sie am Zug:
Fragen und Diskussionen!

Wir suchen:
Admins, Consultants, Trainer!

Wir bieten:
Spannende Projekte, Kundenlob, eigenständige Arbeit, keine Überstunden, Teamarbeit

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

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

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.