

SCALABLE HA-MAILCLUSTER MIT ANSIBLE

→ **Heinlein Support**

- IT-Consulting und 24/7 Linux-Support mit ~35 Mitarbeitern
- Eigener Betrieb eines ISPs seit 1992
- Täglich tiefe Einblicke in die Herzen der IT aller Unternehmensgrößen

→ **24/7-Notfall-Hotline: 030 / 40 50 5 - 110**

- 35 Spezialisten mit LPIC-2 und LPIC-3
- Für alles rund um Linux & Server & DMZ
- Akutes: Downtimes, Performanceprobleme, Hackereinbrüche, Datenverlust
- Strategisches: Revision, Planung, Beratung, Konfigurationshilfe

**Sorry girls & guys,
No live demo today.**

:-(

Inhalt

- Motivation
- Wie skaliert man nun so einen Cluster
- Automatisierung: Ansible
- High Availability und Loadbalancing in Mailclustern

Jeder fängt mal klein an.



mta1

Postfix

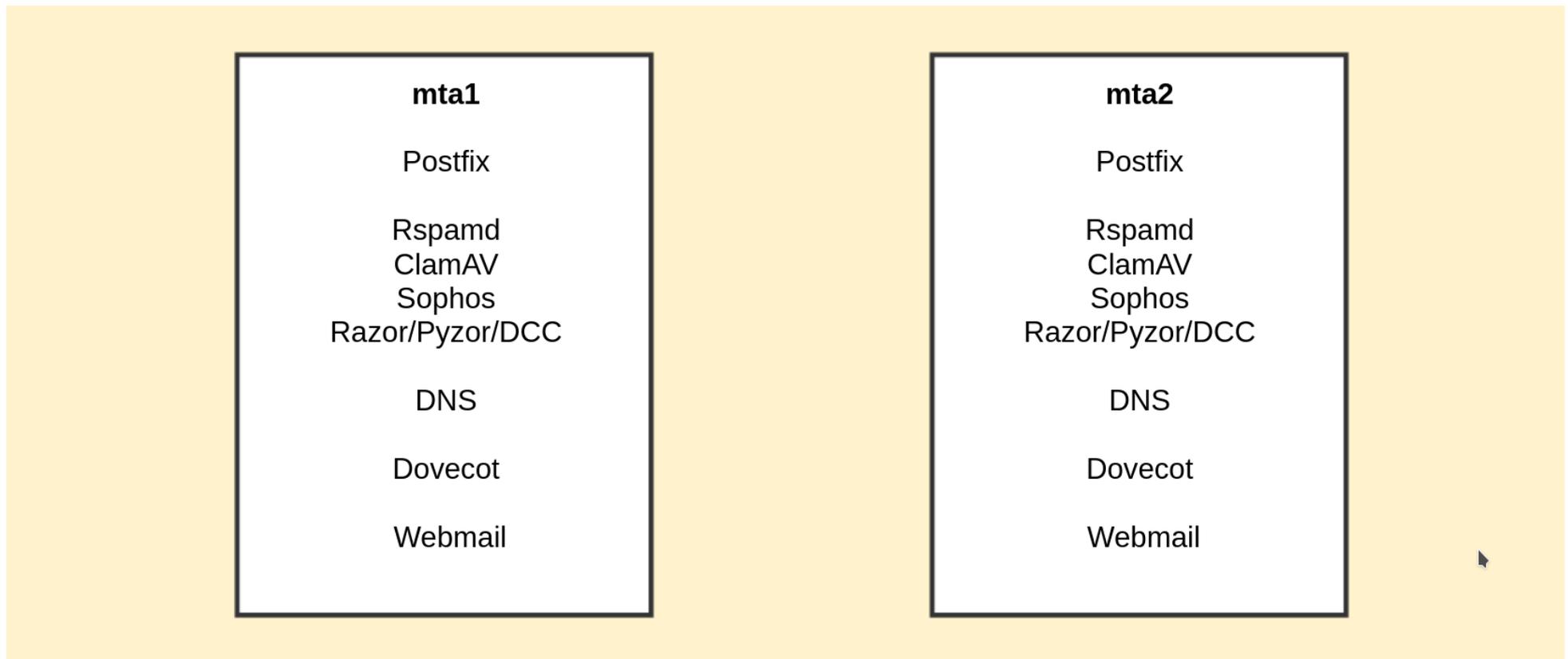
Rspamd
ClamAV
Sophos
Razor/Pyzor/DCC

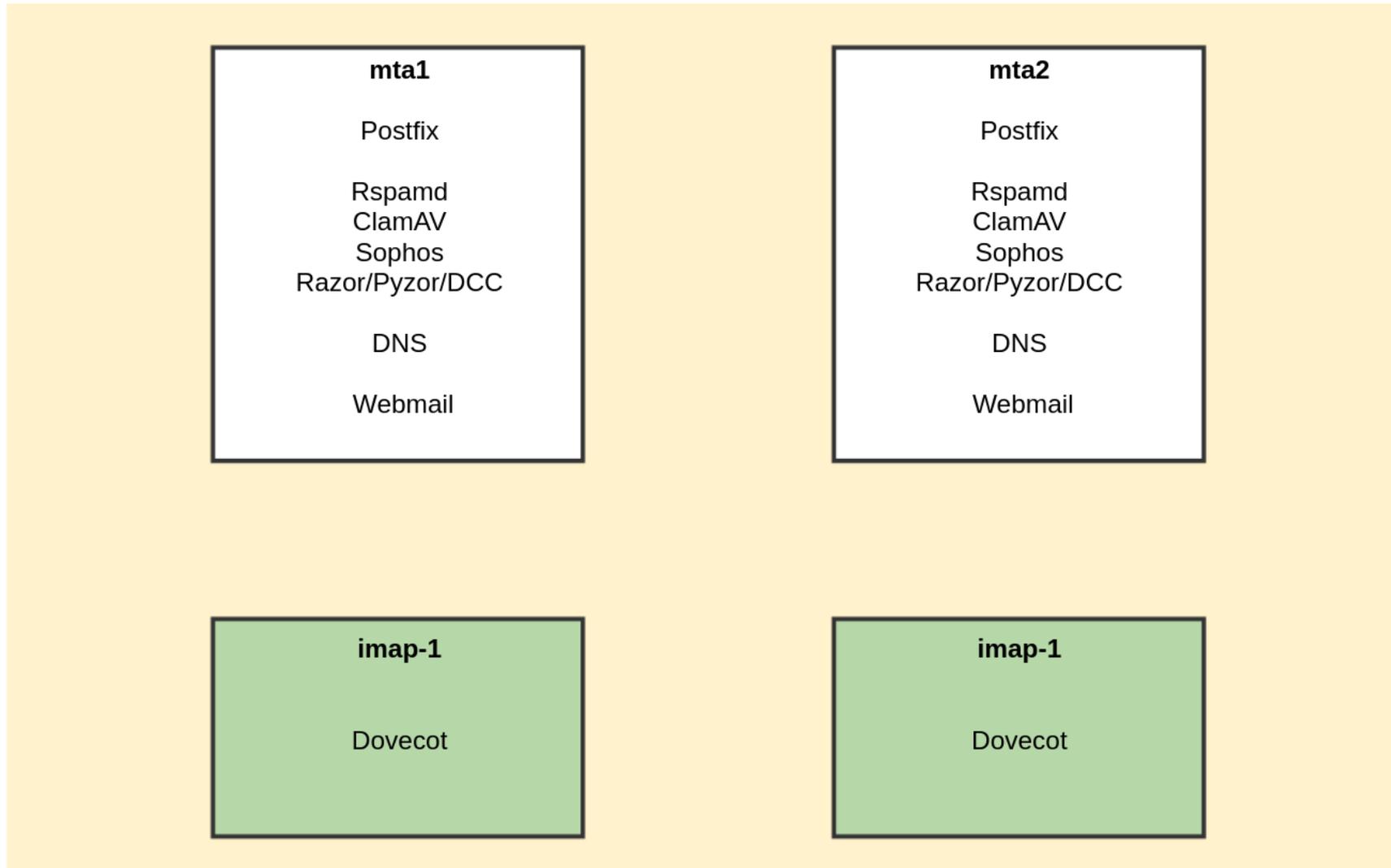
DNS

Dovecot

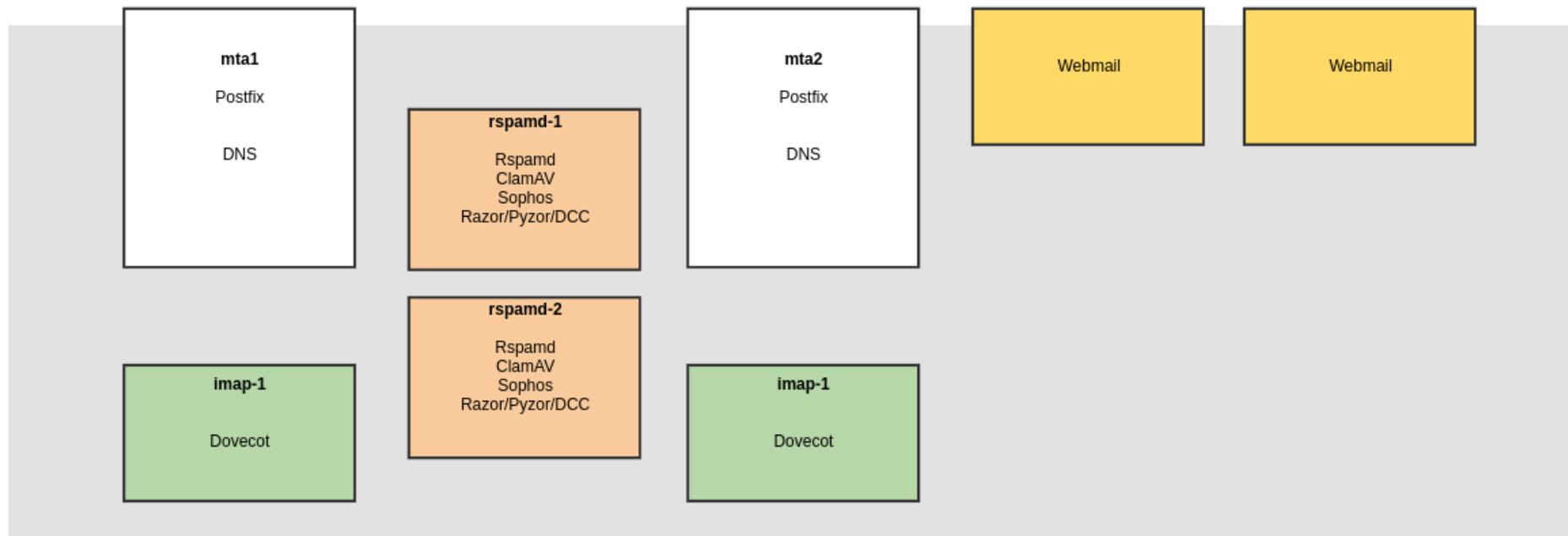
Webmail

Jeder braucht irgendwann einen Partner.

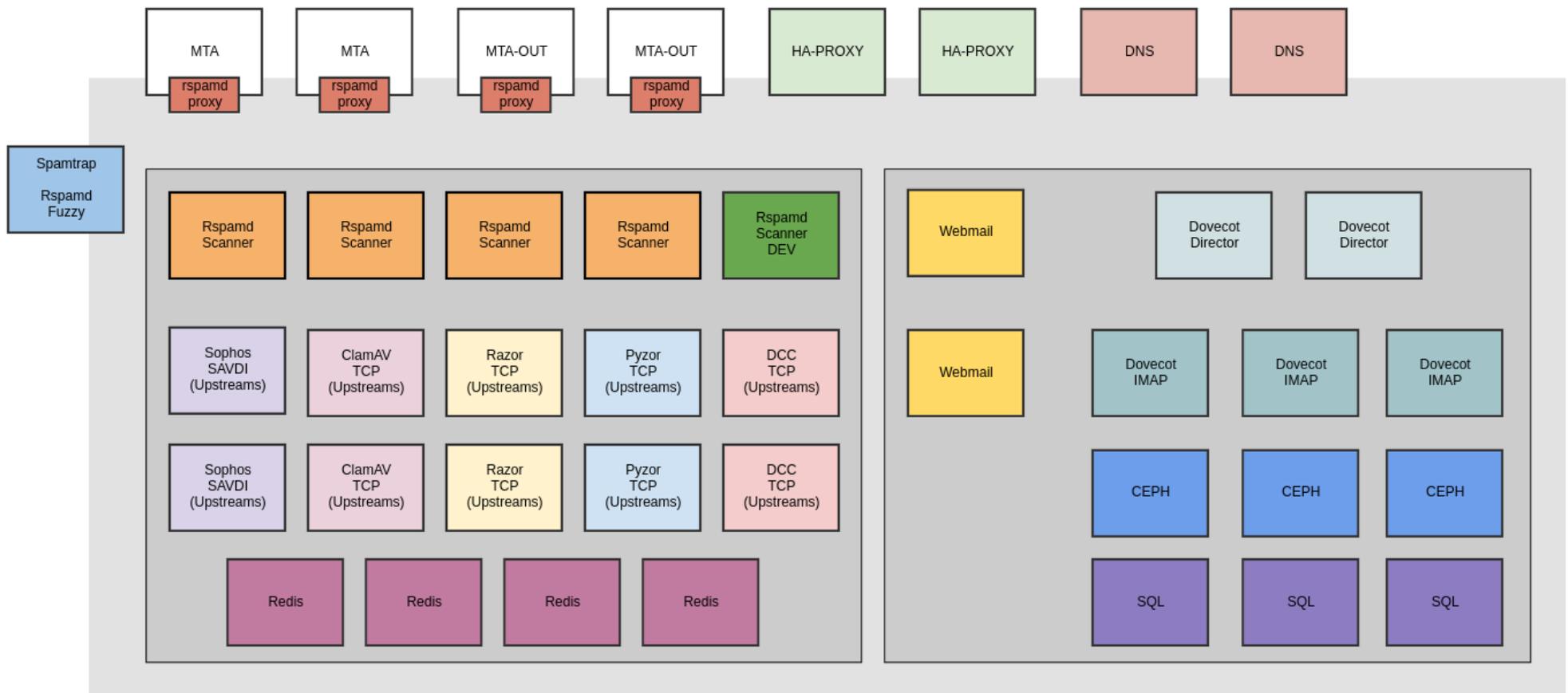




Gleich ist die Familie komplett.



Jetzt ist der Zoo komplett ;)



Automatisierung: Ansible

- Geht auch mit Puppet, Salt, Chef, etc
- Schneller
- Manuelles, stupides Wiederholen verhindern
- Definierter Zustand des Systems
- Dokumentation
- Fertige Vorlagen nutzen

Schreib nicht alles noch einmal

→ Import aus Ansible-Galaxy, Github, eigenem Repository

```
# from galaxy  
- src: yatesr.timezone
```

```
# from GitHub  
- src: https://github.com/bennojoy/nginx
```

```
# from GitHub, overriding the name and specifying a specific tag  
- src: https://github.com/bennojoy/nginx  
  version: master  
  name: nginx_role
```

```
# from a webserver, where the role is packaged in a tar.gz  
- src: https://some.webserver.example.com/files/master.tar.gz  
  name: http-role
```

```
# from Bitbucket  
- src: git+http://bitbucket.org/willthames/git-ansible-galaxy
```

Playbooks können aufeinander aufbauen

- Besser kleinere Rollen, die atomare Funktionen umsetzen
- Beispiel:
 - Dovecot-basic
 - kann Dovecot auf allen möglichen Distributionen installieren
 - Liefert distributionsspezifische Pfade
 - Dovecot-default-config
 - Best-Practice Einstellungen auf allen Dovecots
 - Dovecot-Director
 - Macht aus einem generischen Dovecot einen Director
 - Dovecot-Director-MAILBOX.org
 - Kundenspezifische Einstellungen

Settings (Variablen) mergen

- Variablen mergen zur Laufzeit nach fest definierter Reihenfolge
 - https://docs.ansible.com/ansible/2.4/playbooks_variables.html#variable-precedence-w-here-should-i-put-a-variable

- Beispiel:

```
/role/default/main.yml
```

```
rspamd_upstream_server: 127.0.0.1
```

```
/inventory/group_vars/all.yml
```

```
rspamd_upstream_server: 10.1.1.100, 10.1.1.101, 10.1.1.102
```

```
/inventory/host_vars/my_host.yml
```

```
rspamd_upstream_server: 10.1.1.100, 10.1.1.101
```

```
- name: Set Setup_Facts variables
```

```
set_fact:
```

```
rspamd_upstream_server: 10.1.1.100
```

role defaults

inventory file or script group vars

inventory group_vars/all

playbook group_vars/all

inventory group_vars/*

playbook group_vars/*

inventory file or script host vars

inventory host_vars/*

playbook host_vars/*

host facts

play vars

play vars_prompt

play vars_files

role vars (defined in role/vars/main.yml)

block vars (only for tasks in block)

task vars (only for the task)

role (and include_role) params

include params

include_vars

set_facts / registered vars

Funktionen auf Hosts oder Groups ausführen

→ z.B. Erstellen von IP-Listen aller Gruppen-Mitglieder:

```
mariadb_galera_servers:
```

```
"{{ groups['galera'] | map('extract', hostvars, ['ansible_default_ipv4',  
'address']) | join(',') }}"
```

```
=> mariadb_galera_servers: 10.1.1.100, 10.1.1.101, 10.1.1.102
```

→ Oder einfach DNS?

```
mariadb_galera_servers: sql.slac2018.de
```

Schleifen: über Listen, Gruppen ...

→ Lade alle vorhandenen Template auf das System

```
- name: WORKER - Configure Rspamd Templates /etc/rspamd/local.d/  
  template:  
    src: "{{ item }}"  
    dest: "/etc/rspamd/local.d/{{ item | basename | regex_replace('\.j2','') }}"  
  with_fileglob:  
    - "../templates/worker/etc/local.d/*.j2"
```

Schleifen: über Listen, Gruppen ...

→ Setze für alle Mitglieder einer Gruppe einen DNS A Record

```
- name: Add A Records for all hosts
  powerdns_record:
    name: "{{ item }}.{{ common_domain }}"
    zone: "{{ common_domain }}"
    type: A
    content: "{{ hostvars[item].ansible_host }}"
    ttl: 60
    pdns_host: "{{ pdns_primary_server }}"
    pdns_port: "{{ pdns_primary_server_port }}"
    pdns_api_key: "{{ pdns_api_key }}"
  with_items: "{{ groups['all'] }}"
```

Schleifen: über Listen, Gruppen ...

→ Schleifen auch in Templates Nutzen

```
# Load Balancing for Galera Cluster
```

```
listen galera
```

```
    bind 0.0.0.0:3306
```

```
    balance source
```

```
    mode tcp
```

```
    option tcpka
```

```
    option mysql-check user haproxy
```

```
    {% for host in mariadb_galera_servers_group %}
```

```
    server {{ hostvars[host].host }} {{ hostvars[host].ipv4 }}:3306 check weight 1
```

```
    {% endfor %}
```

Ansible kann auch Cloud

- Es gibt Module für Amazon, Azure, Cloudstack, Docker, Google, LXC, LXD, Openstack, Ovirt, Rackspace, Vmware ...
 - VM erstellen oder löschen
 - Infos zu laufenden Vms abholen
 - Firewall-Einstellungen in der Virtualisierung
- Oft werden dazu dann aber dynamische Inventories (Host-Listen) benötigt

Cloudstack VM erstellen

```
- name: create nodes
  local_action:
    module: cs_instance
    name: "{{ item }}"
    project: "{{ hostvars[item].acs_project }}"
    template: "{{ hostvars[item].acs_template }}"
    service_offering: "{{ hostvars[item].acs_service }}"
    ssh_key: "{{ acs_ssh_key_name }}"
    network: "{{ hostvars[item].acs_network }}"
    state: deployed
  with_items: "{{ groups['acs'] }}"
  register: cs_return
  run_once: True
  become: no
```

Cloudstack Firewall

```
- name: open Firewall Ports for IP1
  local_action:
    module: cs_firewall
    project: "{{ acs_project }}"
    ip_address: "{{ acs_ip1 }}"
    port: "{{ item.port }}"
    protocol: "{{ item.protocol }}"
    cidr: 0.0.0.0/0
  with_items:
    - { port: '53', protocol: 'tcp' }
    - { port: '53', protocol: 'udp' }
    - { port: '80', protocol: 'tcp' }
    - { port: '443', protocol: 'tcp' }
    - { port: '5322', protocol: 'tcp' }
    - { port: '25', protocol: 'tcp' }
    - { port: '7000', protocol: 'tcp' }
```

Ausführung einschränken mit `--limit` und `--tags`

- `--limit`
 - groups
 - hosts

- `--tags`
 - install
 - configure
 - clean_up

Ansible kann auch zentral

- Typischerweise wird Ansible vom lokalen Rechner gestartet
- Zentrale Ansätze: Ansible Tower, Jenkins, Kubernetes ...
 - Lokal testen (Vagrant)
 - Aber von zentral ausrollen - Dev-Ops
 - System-Updates automatisieren

Ansible ist auch Doku

- saubere und konsistente Playbooks sind Voraussetzung
- Idempotente Playbooks
- Jeder kann im Code nachlesen was gemacht wurde
- Besser als jede Beschreibung des IST-Zustandes

Ansible Erweiterbarkeit durch eigene Module

- Ansible basiert auf Python
- https://docs.ansible.com/ansible/latest/dev_guide/developing_modules.html
- Viele interne Module sind Community gepflegt
- Viele externe Module auf Github

Skalieren in Ansible

- Wenn alles perfekt vordefiniert ist - einfach einen neuen Host im Inventory hinzufügen:

```
[director]
```

```
dir1
```

```
dir2
```

```
dir3
```

- Wenn wir dir4 hinzufügen wird eine neue VM erstellt und Dovecot als Director vollständig installiert

Wo Templates Spaß machen

- Überall wo Konfigurationen inkludiert und gemerged werden

- z.B. Rspamd
 - /etc/rsnspamd/local.d/redis.conf

- z.B. Amavis
 - /etc/amavis/conf.d/99-myconfig.conf

- Dovecot
 - /etc/dovecot/conf.d/99-myconfig.conf

Und sonst so?

- Keine manuellen Änderungen
- Am Besten nicht auf bestehende Systeme deployen
 - Neue Infrastruktur erstellen
- Oder nur kleine Änderungen durchführen
 - SSH-Key Deployment
- Ansonsten: Wegwerfen und neu machen geht oft schneller als aufwendig den Fehler finden

High Availability und Loadbalancing in Mailclustern

- High Availability: Redundanz in Form von
 - Master / Slave
 - Master / Master

- Load Balancing: Lastverteilung über mehrere Knoten
 - DNS
 - Level 2 / Level 3 Loadbalancer
 - Proxy Server

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"))
```

SQL / LDAP - eine Datenbank braucht jeder

- Master / Slave und Master / Master Replikation möglich

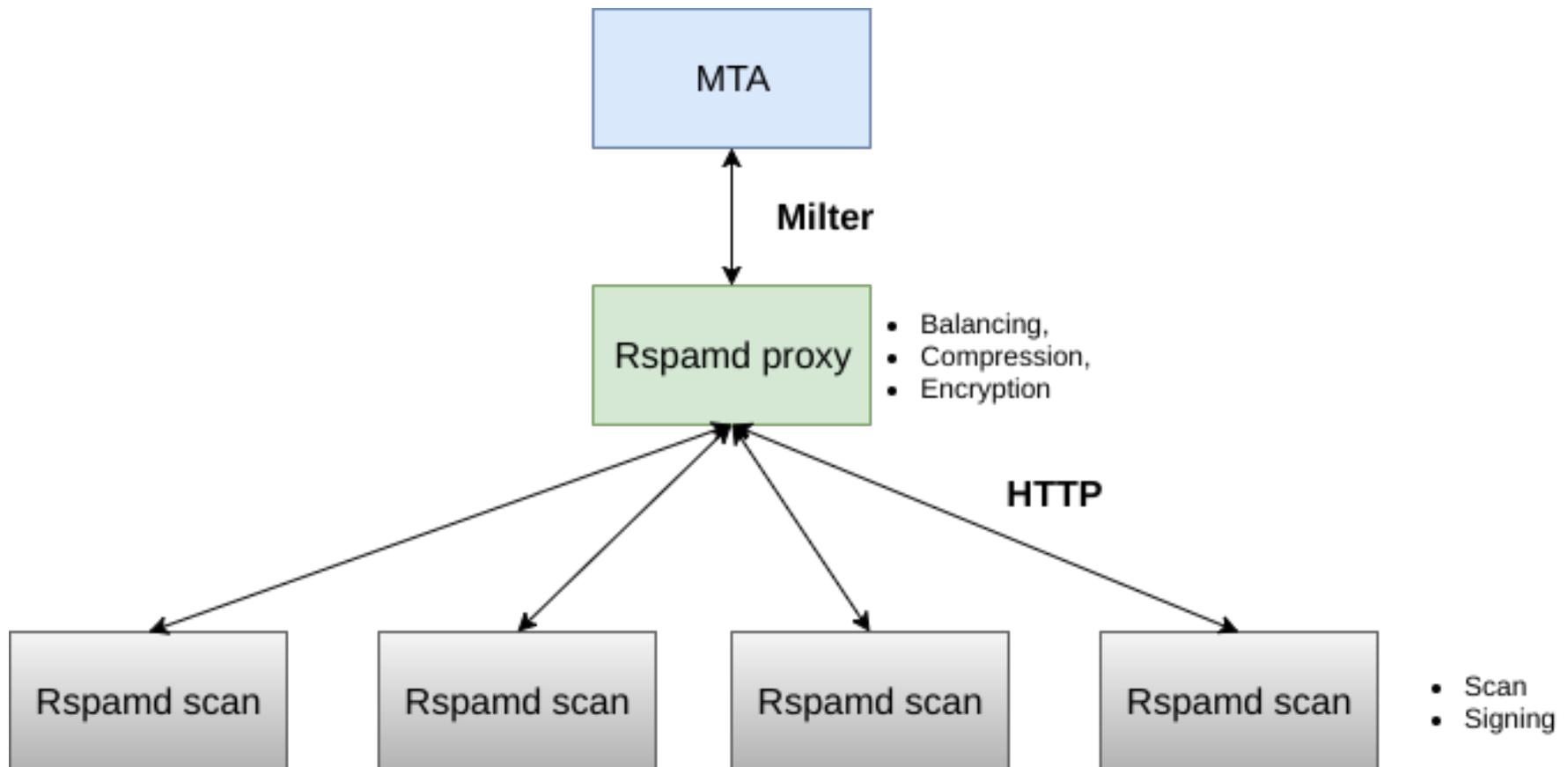
- Mysql
 - Galera
 - Mysql Group Replication
 - Mysql Binlog Master / Slave

- OpenLdap
 - Multi-Master Replikation

Rspamd - Upstreams

- Minimaler Rspamd Proxy direkt auf den MTA's
 - Fehler erzeugt 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

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

```
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
```

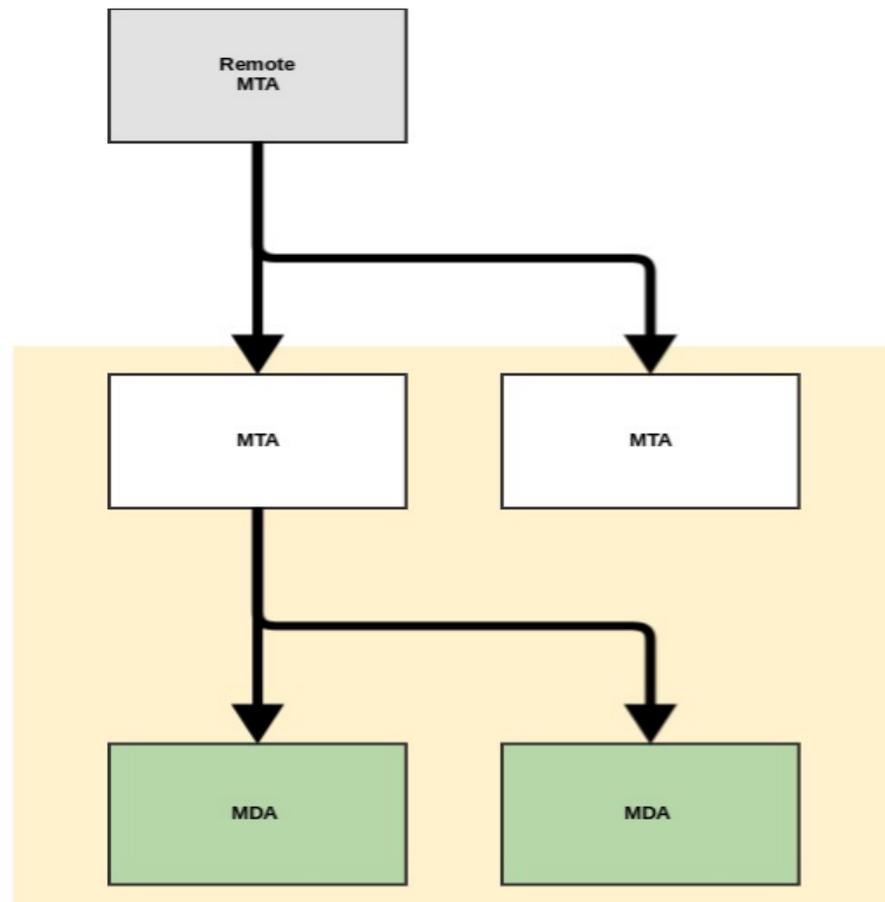
```
Mai 06 15:55:10 mx1.slac2018.de postfix/lmtp[55159]: connect to  
director.slac2018.de[10.1.1.209]:24: Connection refused
```

```
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



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
```

Dovecot - shared Storage

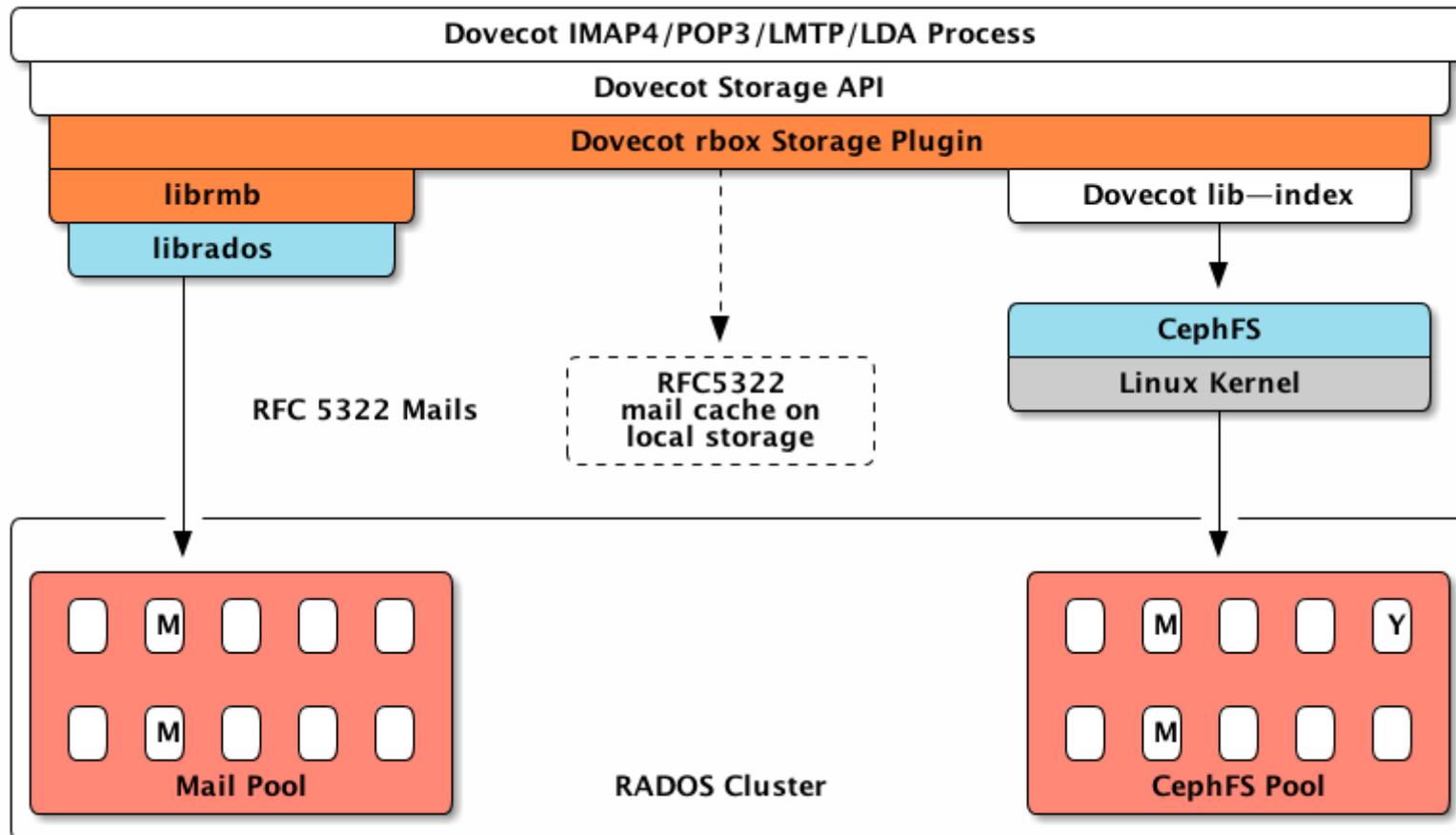
- Shared NFS Storage kann Probleme verursachen (fsync, defekte Indizes ...)
- Redundantes NFS ist kompliziert oder teuer
- Replication ist immer nur N+1
 - Migration ist aufwendig
- S3 (Objektspeicheranbindung) nur in Dovecot Pro - obox
- Neuer Ansatz - gesponsert von der Telekom:
 - Natives Ceph Plugin
 - Noch in Entwicklung und noch nicht Feature Complete

Was ist dieses Ceph?

- Objektspeicher
 - Ich übergebe dem System ein Objekt und es kümmert sich selbst um Speicherung und nachweisbare Redundanz
 - Es wird kein RAID, Fiberchannel, iSCSI ... benötigt
 - Ceph Monitore sorgen für Redundanz 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 Reorganisierung aller Daten
 - Backup?: Ceph kann Snapshots und die Änderungen auf ein anderes Ceph System übertragen

Ceph + Dovecot

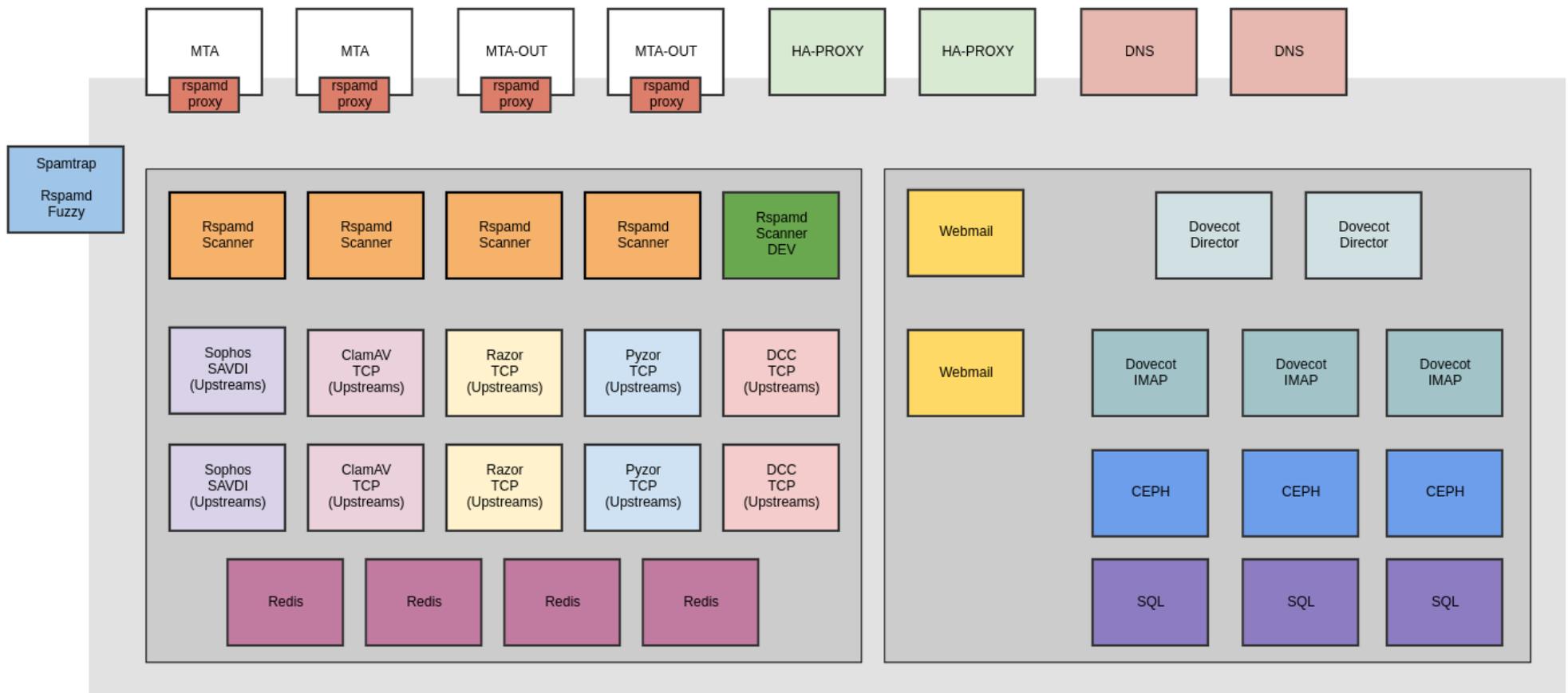
- <https://github.com/ceph-dovecot/dovecot-ceph-plugin>
- Hybrider Entwicklungsansatz
 - Erweiterungen sowohl für Ceph als auch Dovecot
 - Mails werden als Objekte direkt über Rados abgelegt
 - Dovecot Dicts im Ceph Rados omap key/value
 - Dovecot Index *noch* auf CephFS
 - CephFS kann man ungefähr mit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



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 an eigenen Implementierungen für Ceph

Jetzt ist der Zoo komplett ;)



- Auftrennung der Systeme nach Funktionen
- Redundanz schaffen (nicht immer gleich der teure Loadbalancer)
- Trennung von Nutzdaten und Services
- Automatisierung
 - Erstellung, Provisionierung und Löschen von Systemen
 - Updates und Konfigurationsmanagement
- 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 / Alerting
 - Eine Rspamd Maschine rejected gar keinen Spam mehr
 - Scheinbar neue Spamwelle (Monday 3am ???)
 - Löschen von Systemen wenn nicht genügend Last da ist

- 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.