Peer Heinlein <p.heinlein@heinlein-support.de>

Dovecot: Einfach clustern.

Cluster-Varianten im Überblick

- Active/Passive-Cluster (DRBD, Shared SAN)
  - Ausfallsicherheit
- Active/Active Shared Storage (NFS, Cluster-Filesystem)
  - Ausfallsicherheit, Breitenskalierung
- Active/Active Replikation mit individuellem Storage
  - Ausfallsicherheit, Breitenskalierung
- Partitionierter Cluster (Aufteilung der Nutzer auf mehrere Server)
  - Breitenskalierung

Peer Heinlein <p.heinlein@heinlein-support.de>

Breitenskalierung im partitionierten Cluster

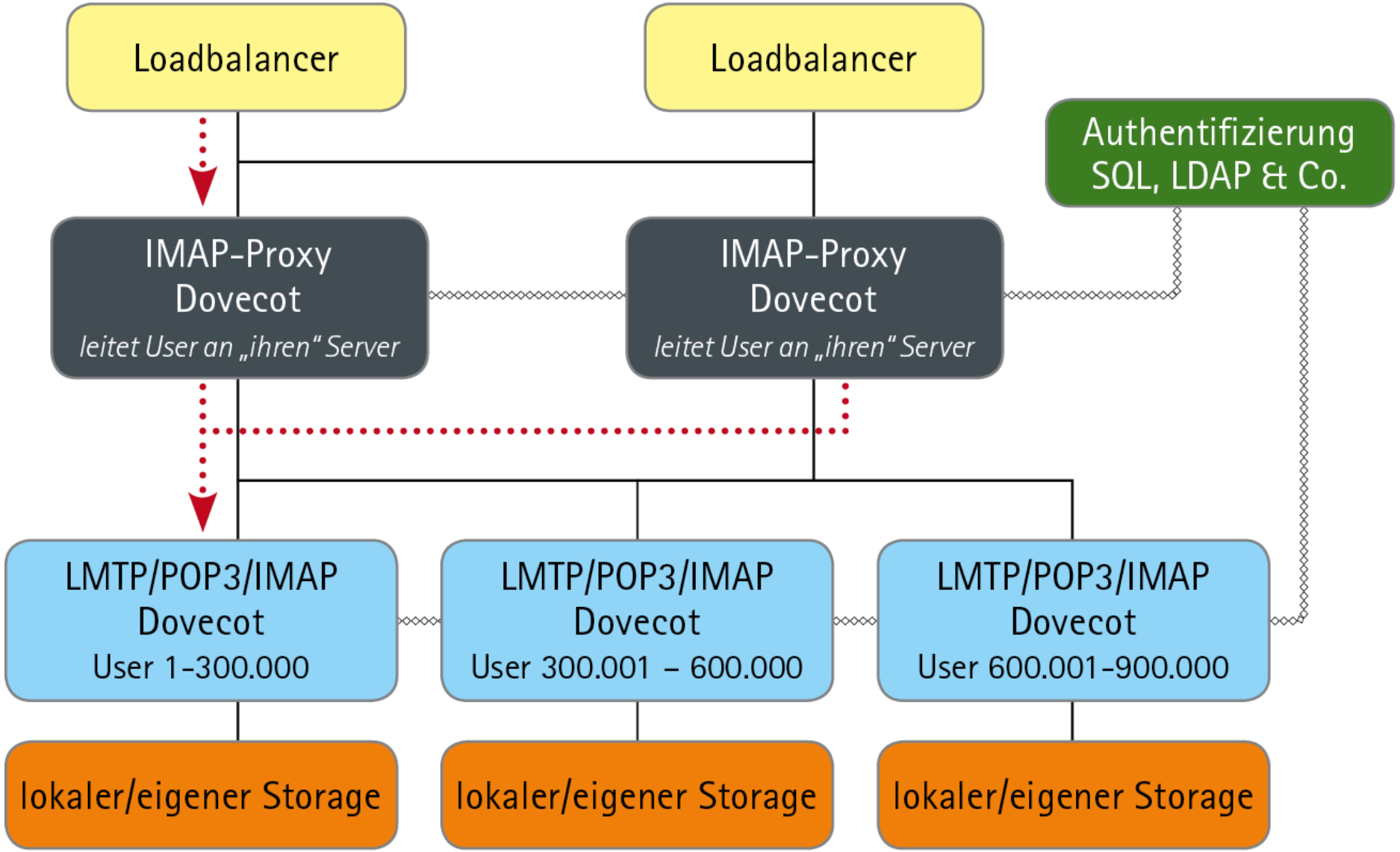
Breitenskalierung im partitionierten Cluster

- Nutzer haben spezifischen Home-Server
  - Hinterlegt als LDAP-Attribut o.ä.
- Keine Ausfallsicherheit des einzelnen Nodes
- Im Prinzip mehrere einzelne IMAP-Server
- Lastverteilung über alle Nodes
- Hat nichts mit Verfügbarkeit zu tun

- Klassischer „Cyrus Murder Cluster“

heinlein

Peer Heinlein <p.heinlein@heinlein-support.de>



Die Userverteilung im partitionierten Cluster

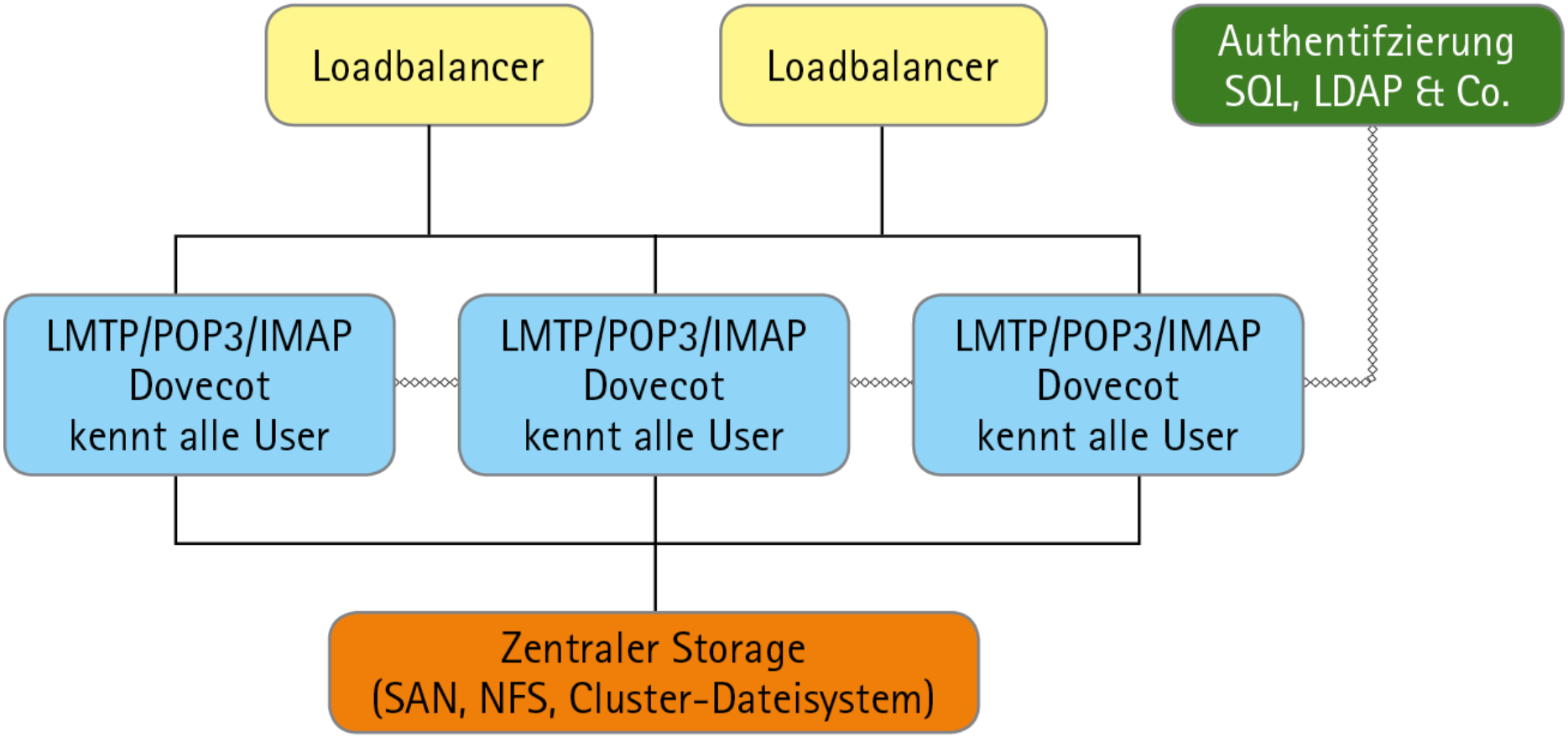
- Layer-7-Loadbalancing nötig („IMAP-Proxy“)
  - Früher: Perdition
  - Heute: Dovecot kann selbst als IMAP-Proxy agieren
- Dedizierte Proxy-Server vor dem eigentlichen Backend
  - „proxy=yes“ und „host=<host>“ als Ergebnis der UserDB-Abfrage
  - Proxy-Server sind dumm und trivial – keine Mail-Daten, nur Authentifizierung
- Implizite Proxy-Server im Backend
  - Connected ein Nutzer auf dem „falschen“ Server wird er transparent zum richtigen Zielsystem per TCP/IP weitergeleitet
  - „proxy=maybe“ und „host=<host>“ als Ergebnis der UserDB-Abfrage
  - Gilt für POP3, IMAP und auch LMTP (!)

Peer Heinlein <p.heinlein@heinlein-support.de>

Active/Active Shared Storage (NFS, Cluster-Filesystem)

Active/Active mit Shared Storage

- Mehrere Frontend-Server teilen sich einen gemeinsamen Storage
- NFS: Ein zentrales Dateisystem, überall verfügbar
  - NFS-Server ist Single Point of Failure?!
- Cluster-Filesystem unterschiedlichster Art
  - Ausfallsicher redundant oder ebenfalls Single Point of Failure?


Loadbalancer
Loadbalancer
Authentifzierung
SQL, LDAP & Co.
LMTP/POP3/IMAP
Dovecot
kennt alle User
LMTP/POP3/IMAP
Dovecot
kennt alle User
LMTP/POP3/IMAP
Dovecot
kennt alle User
Zentraler Storage
(SAN, NFS, Cluster-Dateisystem)

„Shared Storage" hat Designprobleme

- Schützt vor Hardwareausfall, schützt nicht vor Filesystem-Problemen
- Nicht Hardware, sondern der logische Datenbestand ist Ausfallrisiko Nr. 1!
  - Administrationsfehler (rm -rf *, chown, chmod, mv)
  - Defekte Dateisysteme
  - Defekte Index-Datenbanken bei mdbox o.ä.
- Blockreplikation repliziert alle Probleme des Dateisystems in Echtzeit

Shared Storage hat Performanceprobleme

- Dovecots I/O-Optimierungen müssen bei NFS/Cluster-FS abgeschaltet werden
  - mmap_disable = yes
  - dotlock_use_excl = no # only needed with NFSv2, NFSv3+ supports O_EXCL and it's faster
  - mail_fsync = always
  - mail_nfs_storage = yes
  - mail_nfs_index = yes
- Dovecot fährt „mit angezogener Handbremse“
  - I/O suboptimal
  - Cache suboptimal
  - Ggf. Latenzprobleme

Breitenskalierung bei Performancengpässen

- Aber: Die meisten Performanceprobleme sind durch fehlende I/O-Leistung begründet
  - CPU + Netzwerk nicht das Problem!
- Breitenskalierung von Frontend-Nodes mit Shared Storage (NFS) erhöht nicht I/O-Leistung, bremst aber im Zugriff.
  - Absolut kontraprodukiv!
- Besser: Ein starker Server mit lokalem Storage und viel Cache mit normalem Dateisysteme

Sonderfall Cluster-Filesysteme

- Es gibt sehr performante Cluster-Dateisysteme, aber zusätzliche Komplexitätsschicht und Fehlerquelle
  - Cluster-Filesysteme können I/O-Leistung erhöhen
  - Achtung: Auch hier ggf. Performance-Optimierungen abschalten/prüfen, sonst defekter Index möglich!
- Gut, wenn man sie täglich zu administrieren weiß, doof, wenn man sich nicht perfekt dabei auskennt.
- Split-Brain-Problematiken im Cluster-FS möglich (SPoF!)

Peer Heinlein <p.heinlein@heinlein-support.de>

Active/Passive Cluster (DRBD, Shared SAN)

Blockreplikation mit DRBD

- Schützt vor Hardware-Defekten der Festplatten
- Kaum Performanceinbußen, weil immer noch lokales Dateisystem
- Aber: Datensicherheit bei Administrationsfehlern nach wie vor nicht gegeben
- Split-Brain-Problematiken möglich!

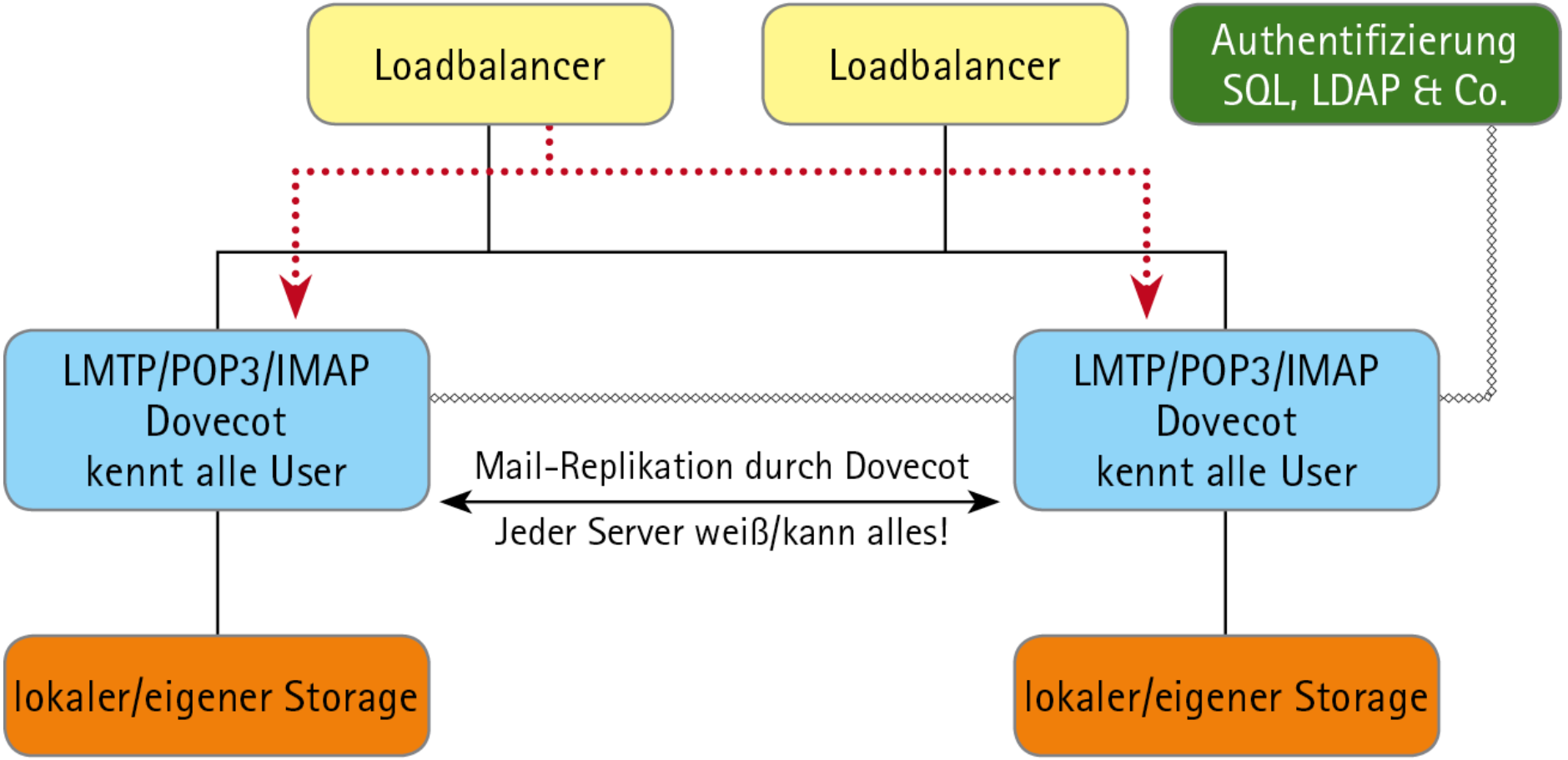
Peer Heinlein <p.heinlein@heinlein-support.de>

Active/Active Replikation mit individuellem Storage

Replikation auf Ebene einzelner E-Mails

- Zwei Dovecot-Nodes mit identischer Konfiguration/Userliste gleichen Mail-Events ab
  - Neue E-Mails
  - Verschobene E-Mails
  - Gelöschte E-Mails
  - Änderung an Metadaten

- Zur Sicherheit alle Postfächer regelmäßig über einen Cron-Job synchronisieren!


Loadbalancer
Loadbalancer
Authentifizierung
SQL, LDAP & Co.
LMTP/POP3/IMAP
Dovecot
kennt alle User
Mail-Replikation durch Dovecot
Jeder Server weiß/kann alles!
LMTP/POP3/IMAP
Dovecot
kennt alle User
lokaler/eigener Storage
lokaler/eigener Storage

Replikationen mit „dsync“

- doveadm kennt mit "dsync" bereits ein Verfahren zum bidirektionalen (!) synchronisieren von Postfächern
  - Früher: Kommando „dsync mirror“
  - Heute: „doveadm sync“
- doveadm auf dem Server über TCP-Port ansprechbar
- Symmetrischer Schlüssel schützt Kommunikation
- Zwei Nodes können dsync über TCP/IP fahren

Replikation ist bidirektional!

- Beide Nodes können parallel angesprochen werden!
- Active/Active-Setup möglich
- dsync kommt sehr gut mit Split-Brain-Situationen klar
  - Sehr geringes Risiko!

- Darum: Warum noch Active/Passive mit DRBD?
- Stattdessen einfacher und sicherer Active/Active mit Replikation!

Replikation und Performance?

- Beide Nodes schreiben parallel -- keine Einsparung
- Aber: Lesezugriffe verteilen sich auf beide Hosts
- Schafft immerhin eine gewisse Entlastung

Replikation und Loadbalancing

- Active/Passive:
  Eine Service-IP über Heartbeat, Pacemaker oder VRRP
- Active/Active:
  Layer-4-Loadbalancer verteilt die Verbindungen auf die Nodes
  Kleine Setups: Jeweils nur einen Node ansprechen => Cache!
- Active/Active stets einfacher und besser. Prima!

Maintenance mit „doveadm replicatior“

- doveadm replicator status
  - Liefert Statistik
- doveadm replicator status '*'

```
username            priority fast sync full sync failed
susi@example.com       none     00:01:20  00:07:52  y
klaus@example.com      none     00:01:20  00:07:52  y
```

- doveadm replicator replicate '*'
- doveadm replicator remove test@example.com
  - Sperrt einzelnen User

Beachtenswertes bei Replikation

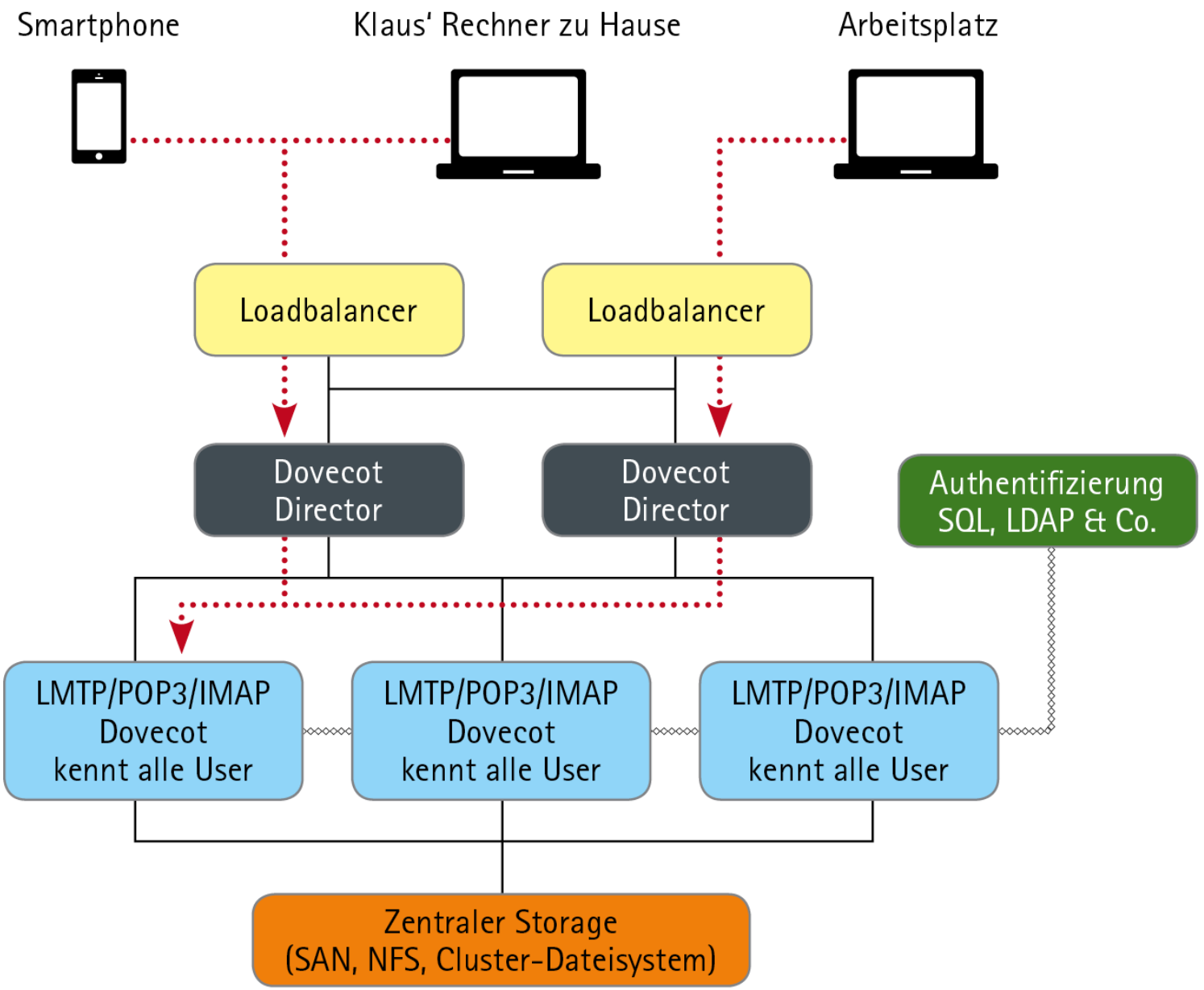
- Dovecot repliziert E-Mails + Sieve-Scripte, sonst nichts.
- Mailserver dürfen keine gemeinsame Quota-Datenbank nutzen
  - Sonst wird alles doppelt gezählt
- Jeder Index ist auf sich selbst angewiesen
  - `doveadm purge`-Kommando auf jedem Server ausführen
  - `doveadm force-resync` oder `doveadm quota recalc` wirken auch nur individuell pro Server
- Im Cluster muß jeder Host einen eigenen Hostnamen haben

- Sonderfälle bei Public Namespace & Co.

Dovecot Director

Get most out of your cluster experience

- Nutzer haben viele parallele IMAP-Verbindungen (Desktop, Office, Handy)
- Für bestes Caching-Verhalten gleichen Nutzer immer auf gleichen Ziel-Server terminieren
- Aber: Verschiedene Source-IPs!
  - „Sticky connections“ / „persistente Verbindungen“ setzen gleiche Source-IP voraus
- Layer-4-Loadbalancer kann Nutzer nicht bündeln!


Smartphone
Klaus' Rechner zu Hause
Arbeitsplatz
Loadbalancer
Loadbalancer
Dovecot Director
Dovecot Director
Authentifizierung SQL, LDAP & Co.
LMTP/POP3/IMAP Dovecot kennt alle User
LMTP/POP3/IMAP Dovecot kennt alle User
LMTP/POP3/IMAP Dovecot kennt alle User
Zentraler Storage (SAN, NFS, Cluster-Dateisystem)

Dovecot Director als Layer-7-Balancer

- Der Director kennt die Dovecot-Backends und deren Zustand
- Dynamisches Management möglich
  - Nodes abschalten/hinzufügen!
- Verteilt Nutzer auf die Backends
  - gleiche Logins immer zum gleichen Server
  - Gewichtete Verteilung auf die Backends möglich!
- Dovecot Director kann man nutzen, muß man i.d.R. aber nicht, solange keine FS-Probleme zu erwarten sind
- Kann Active/Active/Active parallel betrieben werden – die Directoren besitzen über ein Protokoll gemeinsames Wissen
  - Ausfallsicherheit: Layer-4-Loadbalancer verteilt auf mehrere Directoren

```
director_servers = 192.168.3.10:9090 192.168.3.11:9090
director_mail_servers = 192.168.50.161 192.168.50.162 192.168.50.163

service director {
  unix_listener login/director {
    mode = 0666
  }

  inet_listener {
    port = 9090
  }
}

service imap-login {
  executable = imap-login director
}
service pop3-login {
  executable = pop3-login director
}

# Enable director for LMTP proxying:
protocol lmtp {
  auth_socket_path = director-userdb
}
```

Dovecot Director Maintenance

- doveadm director ring status

```
director ip    port type last failed
192.168.50.161 9090 self never
192.168.50.162 9090 l+r  never
```

- doveadm director status

```
mail server ip vhosts       users
192.168.50.161    100        5427
192.168.50.162    100        5877
```

Dovecot Director Maintenance

- doveadm director add 192.168.50.161 150
  - Fügt Host mit Gewichtung „150“ hinzu
- doveadm director remove 192.168.50.161
  - Entfernt Host aus Balancing
- doveadm director dump
  - Sichert aktuelle dynamische Konfiguration für späteren Neustart

Dovecot Director Maintenance

- doveadm director status klaus@example.com

  Current: not assigned
  Hashed: 192.168.50.161
  Initial config: 192.168.50.161

- doveadm director map <host>
  - Zeigt aktuelle User-Zuordnungen

- doveadm director flush 192.168.50.161 (auch: all)
  - Löscht Zuordnungen dieses Hosts

Idealsetup

- Ein Active/Passive Layer-4-Loadbalancer-Pärchen
  - (Oder Multicast?)
- verteilt auf > zwei Directoren
- verteilen auf repliziert <n> partitionierte Server
  - (ggf. nochmal durch Layer-4-Balancing ergänzt)
- die jeweils im Pärchen repliziert betrieben werden

Peer Heinlein <p.heinlein@heinlein-support.de>

Howto Dovecot-Replikation

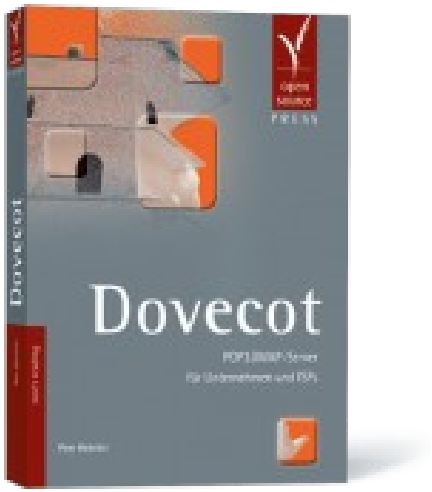
Das Howto zum Replikations-Cluster

- „doveadm user '*'“ muß funktionieren
  - "Iterate Query" in der userdb-Config
  - Generiert Userliste für vollen Replikationsdurchlauf
- Replikations-Config aktivieren (nächste Folie)
- „doveadm replicator“-Kommando anwerfen
  - Zweiten Node beobachten
  - Logfile lesen

```
mail_plugins = $mail_plugins notify replication

service aggregator {
  fifo_listener replication-notify-fifo {
    user = vmail
  }
  unix_listener replication-notify {
    user = vmail
  }
}

service replicator {
  process_min_avail = 1
  unix_listener replicator-doveadm {
    mode = 0600
  }
}

service doveadm {
  inet_listener {
    port = 12345
  }
}

doveadm_port = 12345
doveadm_password = secret

plugin {
  mail_replica = tcp:192.168.50.161
}
replication_dsync_parameters = -d -n INBOX -l 30 -U
```

Wenn es um echtes Papier geht:

- „Dovecot – POP3/IMAP-Server für Unternehmen und ISPs“
  - Das erste Dovecot-Buch auf 400 Seiten
  - Shared Folder, Quota, Cluster: Alles drin.



- Das Postfix-Buch
  Sichere Mailserver mit Postfix
  - Der Klassiker mit rund 1000 Seiten
  - Beinhaltet auch Spamschutz und Rechtsgrundlagen

- Natürlich und gerne stehe ich Ihnen jederzeit mit Rat und Tat zur Verfügung und freue mich auf neue Kontakte.

Peer Heinlein

Mail: p.heinlein@heinlein-support.de

Telefon: 030/40 50 51 - 42

- Wenn's brennt:
  - Heinlein Support 24/7 Notfall-Hotline: 030/40 505 - 110

heinlein

Peer Heinlein <p.heinlein@heinlein-support.de>

Ja, diese Folien stehen auch als PDF im Netz...
http://www.heinlein-support.de/vortrag

Soweit, so gut.

Gleich sind Sie am Zug:
Fragen und Diskussionen!

Wir suchen neue Kollegen für:

Helpdesk, Administration, Consultanting!

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

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

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

Und nun...

- Vielen Dank für's Zuhören...
- Schönen Tag noch...
- Und viel Erfolg an der Tastatur...

Bis bald.

Heinlein Support hilft bei allen Fragen rund um Linux-Server

HEINLEIN AKADEMIE

Von Profis für Profis: Wir vermitteln in Training und Schulung 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 für Archivierung, IMAP und Anti-Spam und speziell für den Serverbetrieb konzipierte Software rund ums Thema E-Mail.