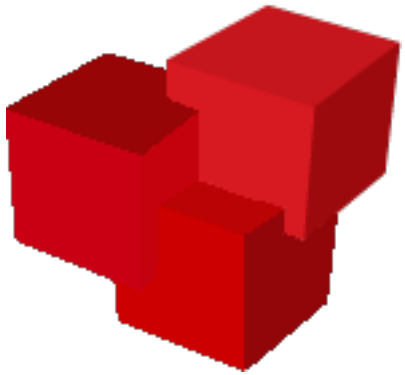
LIS AG
Linux Information Systems

Standortübergreifende Cluster mit RHEL/CentOS (stretched Cluster)

7.6.2013

Inhalt

1. generelle Übersicht des Redhat-Cluster
2. Quorum
3. Fencing
4. Konfiguration mit Conga
5. Konfiguration für SLAC-Admins



Bestandteil eines Clusters

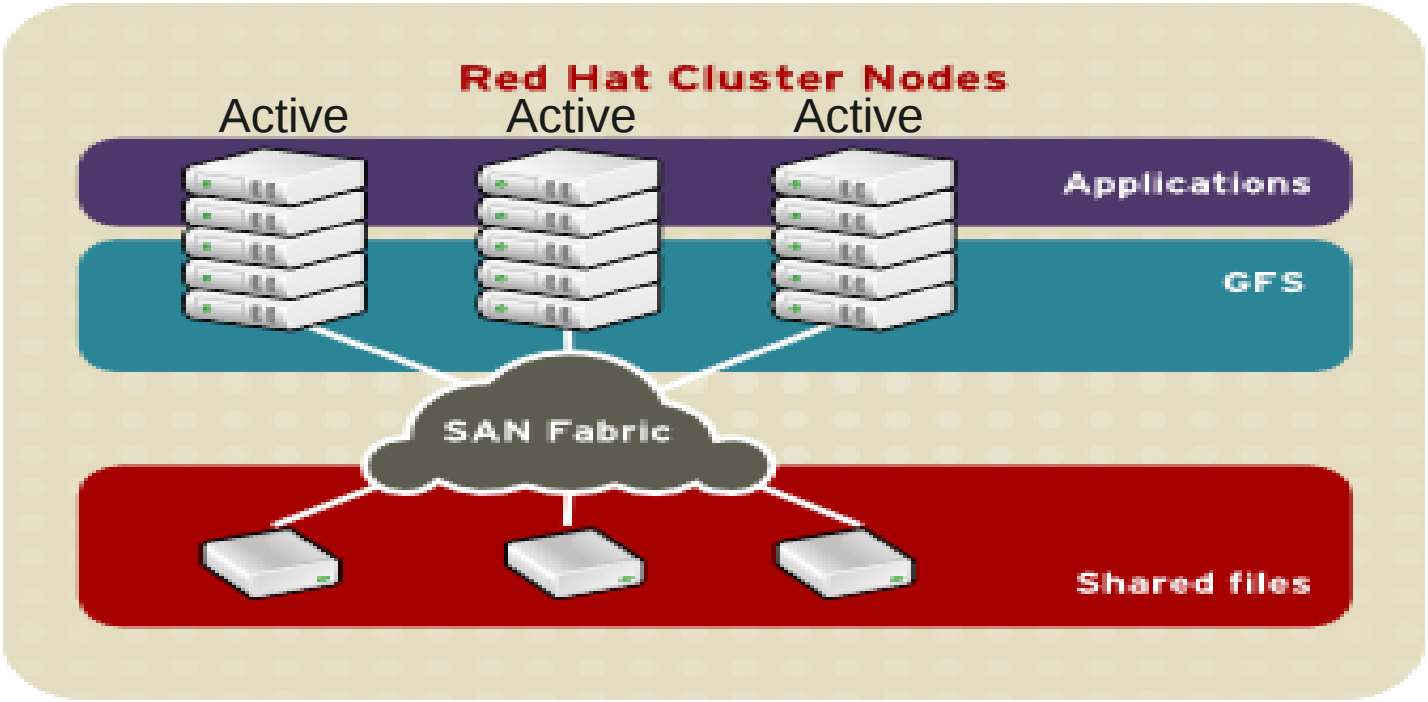
LIS AG
Linux Information Systems

- Cluster-Dienste
  - cman – Cluster Manager (benutzt Corosync)
  - rgmanager – Ressource Manager
  - qdiskd – Quorum Disk Daemon
- CLVM – Cluster LVM
- GFS2 – Shared Filesystem
- DLM – Distributed Lock Manager

Cluster-Typen

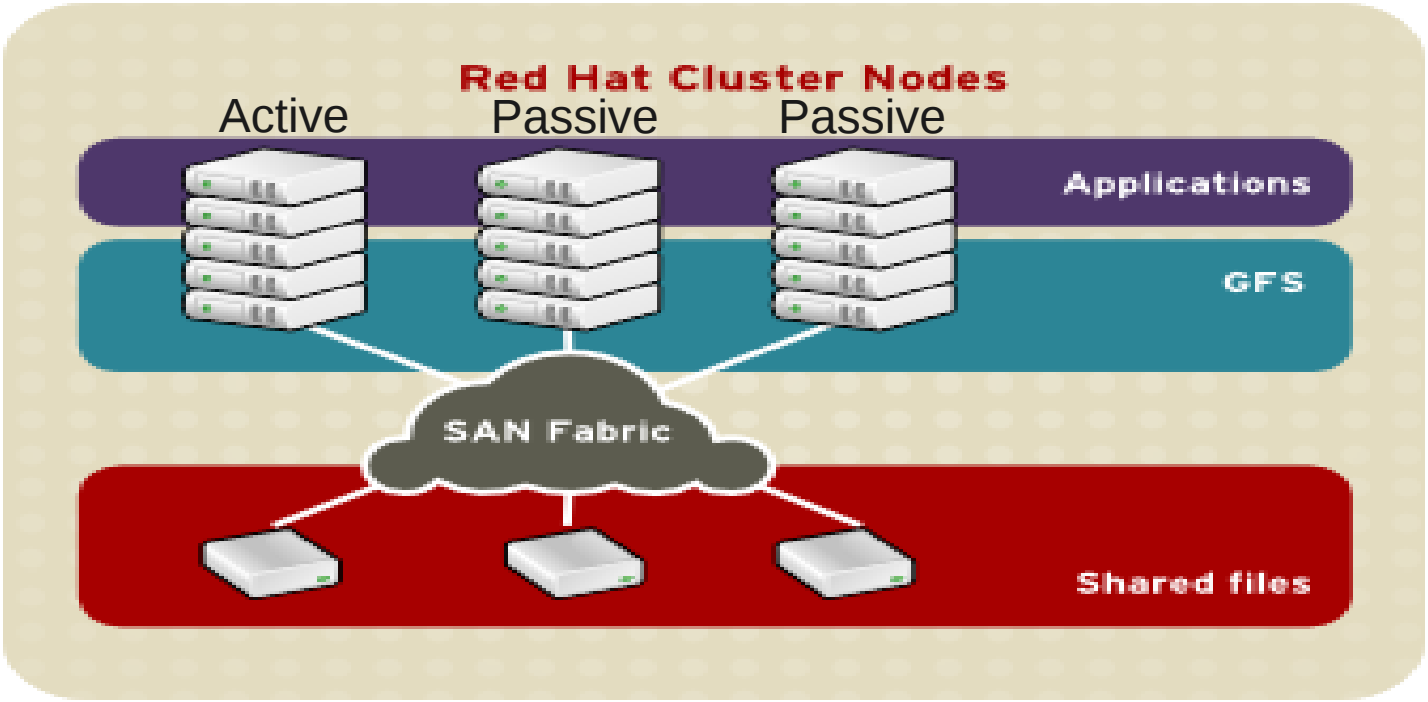
- Shared-Cluster
- HA-Cluster
- Stretched-Cluster

Shared-Cluster



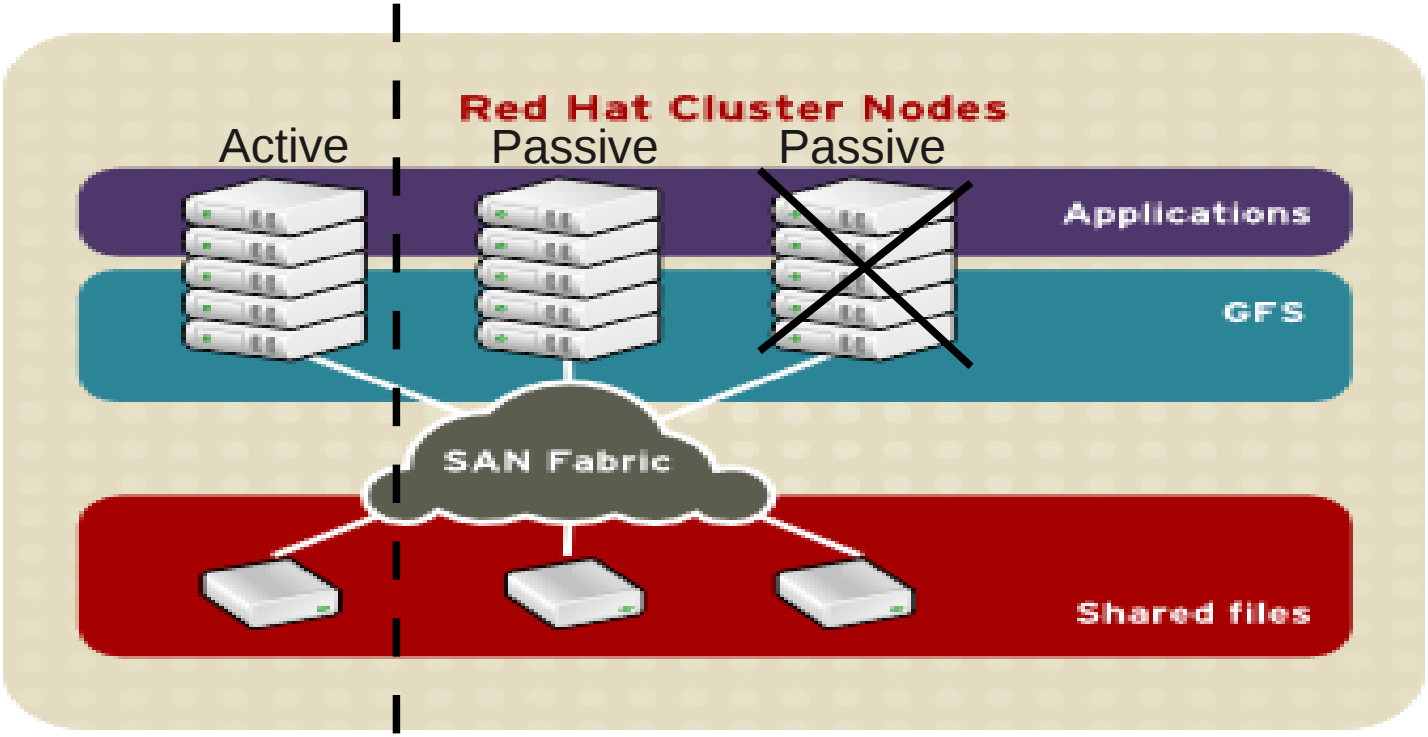
- GFS2 erlaubt Lesen und konkurrierendes Schreiben
- CLVM erlaubt LVM im Cluster
- DLM sorgt für das Locking im LVM und GFS2
- lvm.conf: `locking_type = 3` und `fallback_* = 0`

HA-Cluster



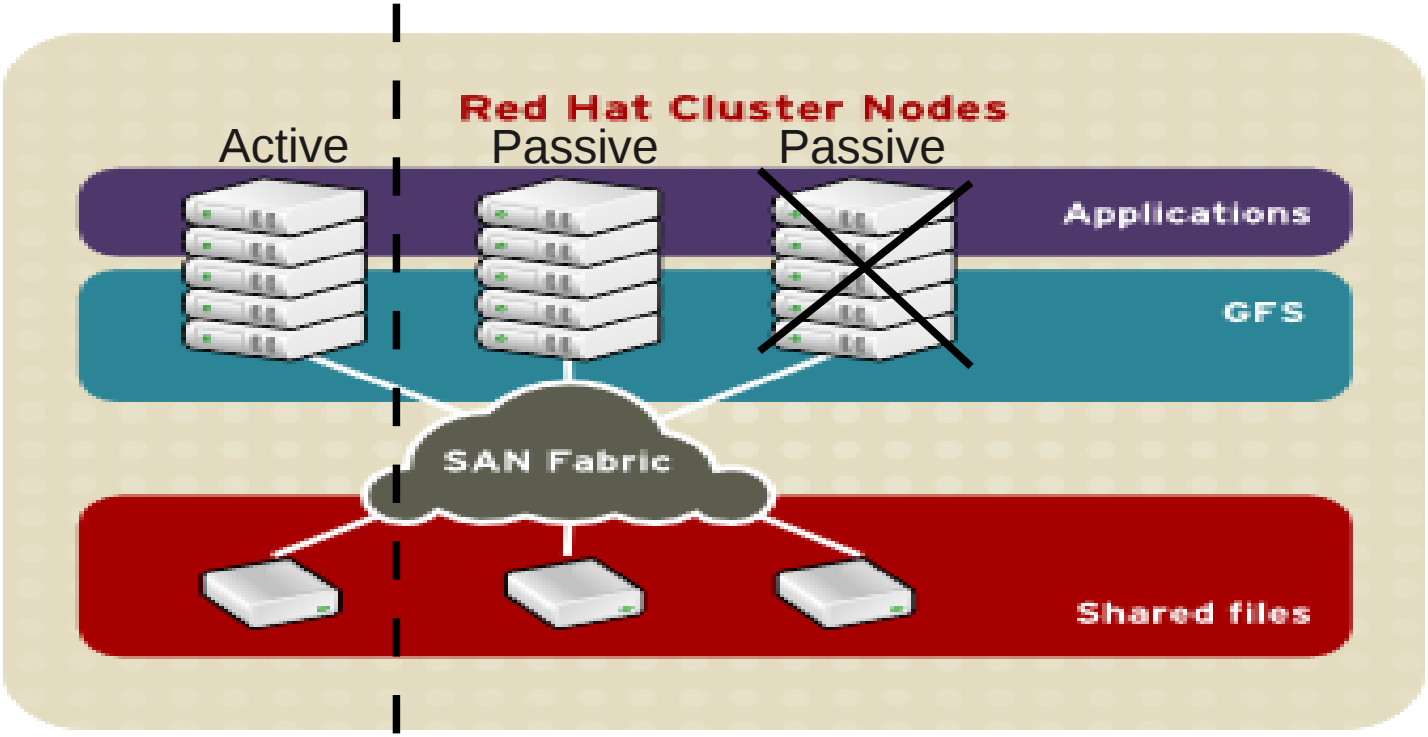
- nur ein Knoten ist aktiv, GFS ist daher unnötig aber möglich
- CLVM und DLM nötig für LVM im Cluster
- LVM immer nur auf einem Knoten aktiv
- lvm.conf: `locking_type = 3` und `fallback_* = 0`
- cluster.conf: `<lvm vg=...>` als Hinweis für den Cluster

Stretched HA-Cluster



- getrennt mit Medienbruch (Interconnect für FC und GB) und mehr als 200m Entfernung
- Paketlaufzeiten und Latenzen wie lokal erwartet (<=2ms RTT)
- ist immer ein HA-Setup mit max. 2 Sites & 16 Knoten, d.h. Active/Passive
- Einschränkungen: kein GFS2, kein CLVM, kein DLM

Stretched HA-Cluster



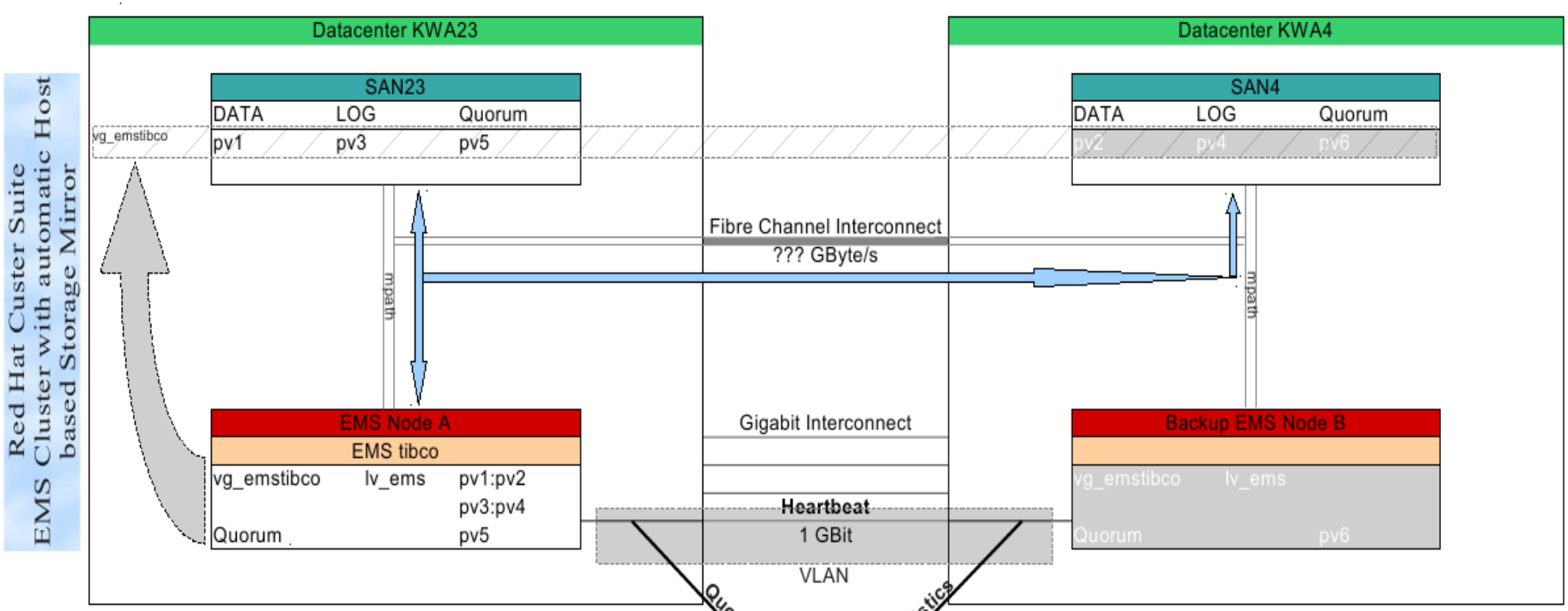
- lvm.conf: `locking_type = 1` und Tagging der Volumes mit `volume_list = vg00, @node`
- cluster.conf: `<lvm vg=...>` als Hinweis für den Cluster
- `mkinitrd` notwendig, weil diese beim Booten das LVM prüfen muss (kein DLM!), um es nötigenfalls zu aktivieren

Problemfall: Storage

LIS AG
Linux Information Systems

- Storage meist nicht ausfallsicher ausgelegt
- Replikationsmöglichkeiten
  - Metro Cluster (seeehr teuer)
  - proprietäre Lösung des SAN (meist nur passiv)
  - LVM-Mirror
  - DRBD
- Storage Cluster
  - DRBD
  - NFS

Stretched LVM-Mirroring Setup



Characteristics
- Active-Passive setup
- non shared
- host base mirror
- stretched

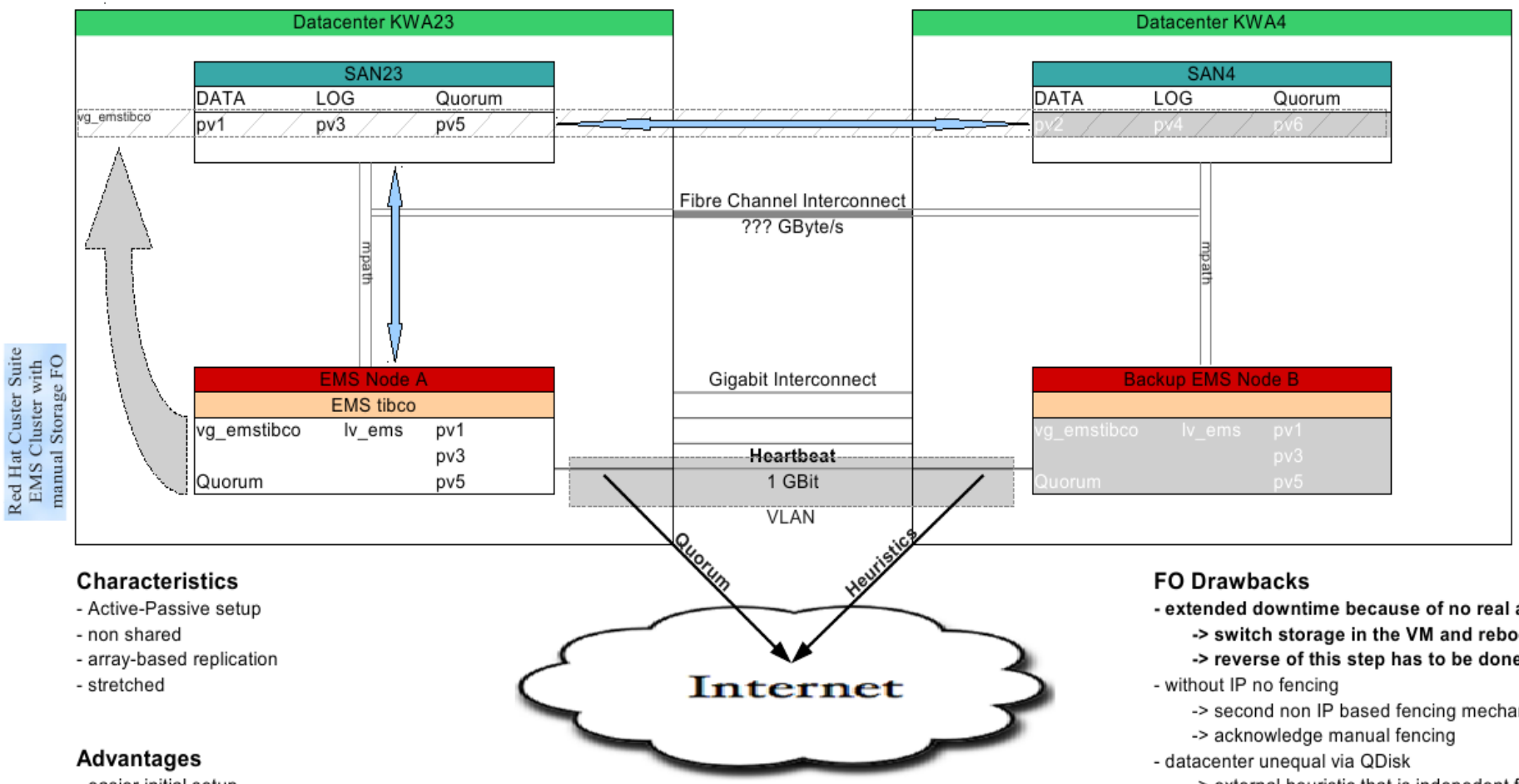
Advantages
- automatic FO in almost all desaster scenarios

Disadvantages
- not extendable - pure 2-node setup
- complex to setup, so failures need deeper knowledge
- complex to test and document properly

FO Drawbacks
- without IP no fencing
  - -> second non IP based fencing mechanism (e.g. SCSI fencing)
  - -> acknowledge manual fencing
- datacenter unequal via QDisk
  - -> external heuristic that is indepedent from Interconnect
- **Datacenter split leads to fence stall**
  - **-> acknowledge manual fencing**
- *downtime after mirror split*
  - *-> planned downtime needed*

Bold = service interruption until manual intervention is done
Italic = planned service interruption to regain initial cluster status

Stretched Array-Mirroring Setup



Characteristics
- Active-Passive setup
- non shared
- array-based replication
- stretched

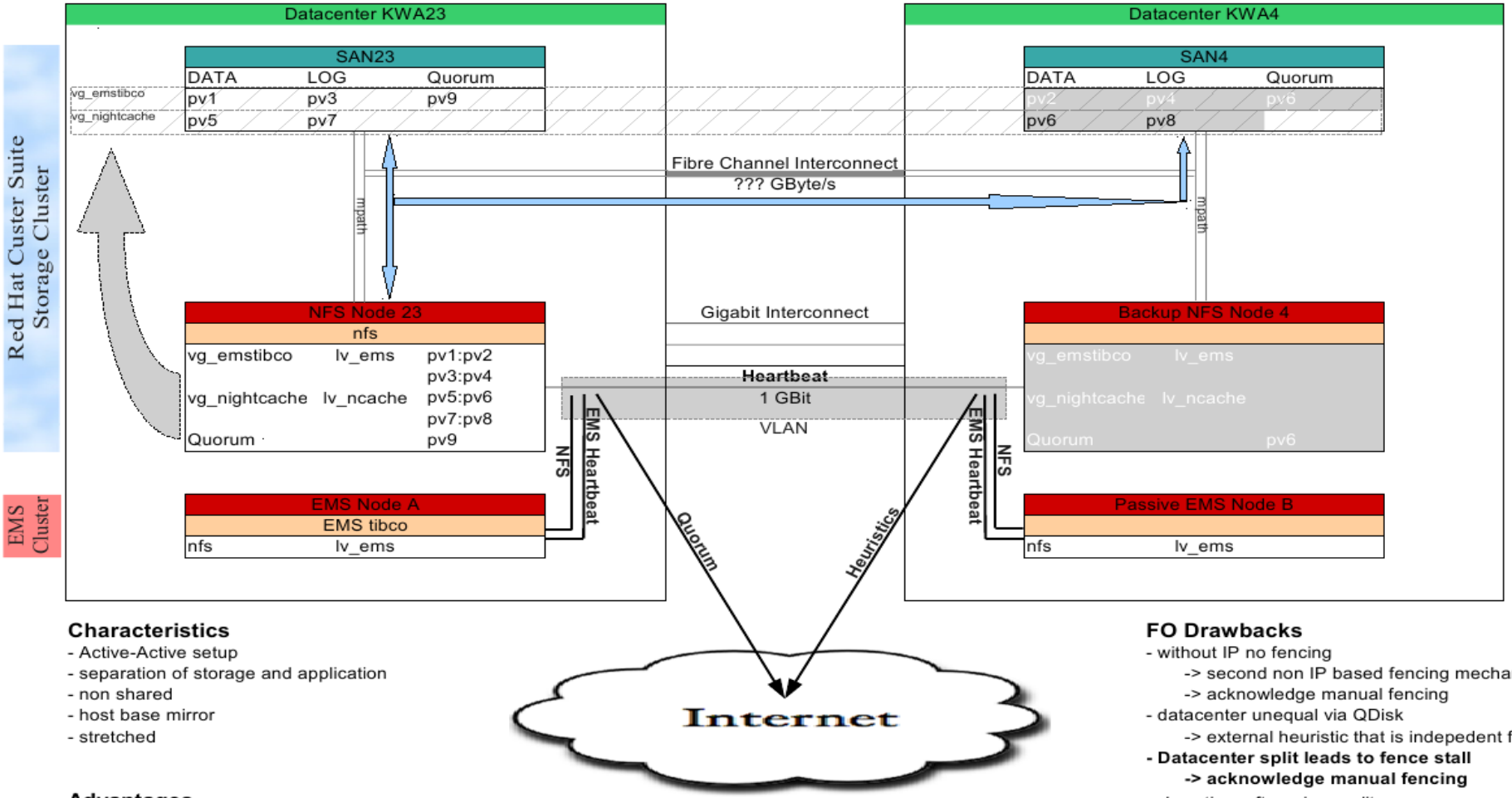
Advantages
- easier initial setup

Disadvantages
- no true HA setup
- manual intervention needed during site split
- extended downtime

FO Drawbacks
- **extended downtime because of no real active mirror**
  - **-> switch storage in the VM and reboot the nodes**
  - **-> reverse of this step has to be done in the same manner**
- without IP no fencing
  - -> second non IP based fencing mechanism (e.g. SCSI fencing)
  - -> acknowledge manual fencing
- datacenter unequal via QDisk
  - -> external heuristic that is indepedent from Interconnect
- **Datacenter split leads to fence stall**
  - **-> acknowledge manual fencing**
- *downtime after mirror split*
  - *-> planned downtime needed*

Bold = service interruption until manual intervention is done
Italic = planned service interruption to regain initial cluster status

Stretched HA-NFS Storage Setup



Characteristics
- Active-Active setup
- separation of storage and application
- non shared
- host base mirror
- stretched

Advantages
- clear separation of application and storage layer
- application can use internal FO machanisms (mostly faster)
- easily expandable to more than two nodes and/or a LB setup
- possibility to have the Storage Cluster in Bare Metal

Disadvantages
- initially more machines/VMs
- initially more work

FO Drawbacks
- without IP no fencing
  -> second non IP based fencing mechanism (e.g. SCSI fencing)
  -> acknowledge manual fencing
- datacenter unequal via QDisk
  -> external heuristic that is indepedent from Interconnect
- **Datacenter split leads to fence stall**
  -> acknowledge manual fencing
- *downtime after mirror split*
  -> planned downtime needed

Bold = service interruption until manual intervention is done
Italic = planned service interruption to regain initial cluster status

Störung: Partitionierung

LIS AG
Linux Information Systems

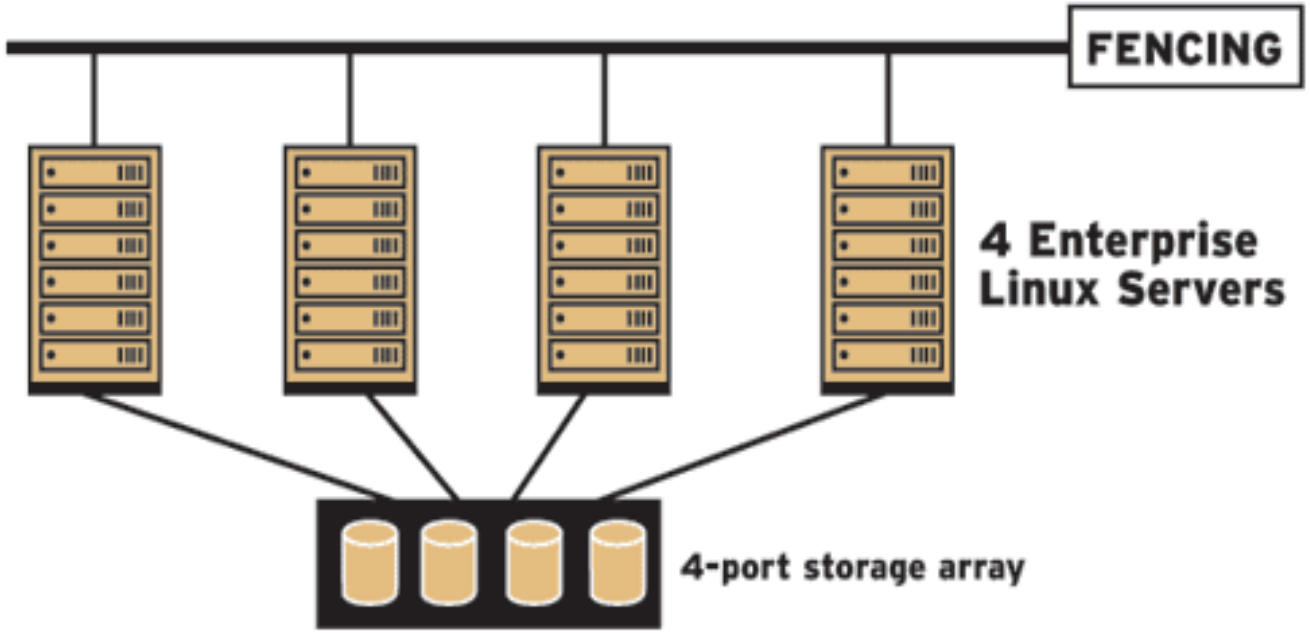
- Trennung des Clusters
  - fünf Knoten
  - Switch defekt
  - ein Teil mit drei und einer mit zwei Knoten bleibt übrig
- wir brauchen ein Quorum!
  - ein Knoten, der kein Quorum sieht, schaltet ab!
  - Achtung! das kann die Verfügbarkeit einschränken
- im Stretched Setup (2 Knoten) benötigen wir eine dritte Instanz
  - > Quorum Disk!

Quorum Disk

LIS AG
Linux Information Systems

- Shared Disk
- jeder Knoten hat eigenen Bereich, auf dem er seinen Heartbeat legt
- kann ein Knoten nicht Schreiben, schaltet er ab
- sieht ein anderer Knoten keine Schreibtzugriffe mehr, so schaltet er den betroffenen Knoten ab

Shared-Cluster Fencing



- Datenintegrität auch bei „auskreisenden“ Noten
- Knoten werden kooperativ oder nicht kooperativ gefenced
- der Cluster bleibt funktionstüchtig

Konfiguration mit Conga

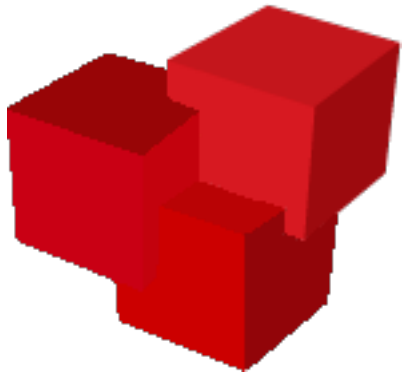
- Web-GUI
- schreibt in die /etc/cluster/cluster.conf der Knoten
- liest diese auch live ein und interpretiert sie
- daher kooperativ mit dem CLI-Admin
- unnötig

Konfiguration für SLAC-Admins

LIS AG
Linux Information Systems

- Conga ist unnötig
- /etc/cluster/cluster.conf kann direkt editiert werden
- Konfiguration validieren: `ccs_config_validate`
- Konfiguration synchronisieren:
  `ccs_sync -f /etc/cluster/cluster.conf`
- Status: `clustat`
- Ressourcen Management: `clusvcadm`
- Fencing: `fence_node -n <Name a.d. cluster.conf>`

Fragen?



SLAC 2013 – Joachim v. Thadden

LIS AG
Linux Information Systems

Vielen Dank
für Eure Aufmerksamkeit!