ISG.EE

System Management, Training Consulting

Cluster Technologien – Was sie leisten und was nicht

Marco Reimers
IT Support Group D-ITET (ISG.EE)
ETH Zürich
CH-8092 Zürich,Switzerland
mreimers@ee.ethz.ch

Übersicht

- Cluster: Wieso, weshalb, warum ...
- Methoden und Kategorien
- Management
- Problematiken
- OpenSource Projekt

Was ist ein Cluster?

A cluster is a group of servers
and other resources that act like
a single system and enable
high availability and, in some cases,
load balancing and parallel processing.

Quelle: http://searchexchange.techtarget.com

Warum ein Cluster?

- Ausfallsicherheit / Redundanz (24x7x365)
- Lastverteilung
- Performancesteigerung
- Skalierbarkeit
- Unterbrechungsfreie Wartung / Updates
- Reduzierter Administrationsaufwand
- Kostenreduzierung

→ Reliability, Availability and Serviceability (RAS)

Wofür ein Cluster?

- Anwendungen
- Daten
- File System
- Network / SAN

Welche Nachteile gibt es?

Im Vorteil liegt auch häufig ein Nachteil!

- Administrationsaufwand → Komplexität der Administration
- Datenredundanz → Dateninkonsistenz
- Kostenreduzierung → „Single-Point-of-Failure“

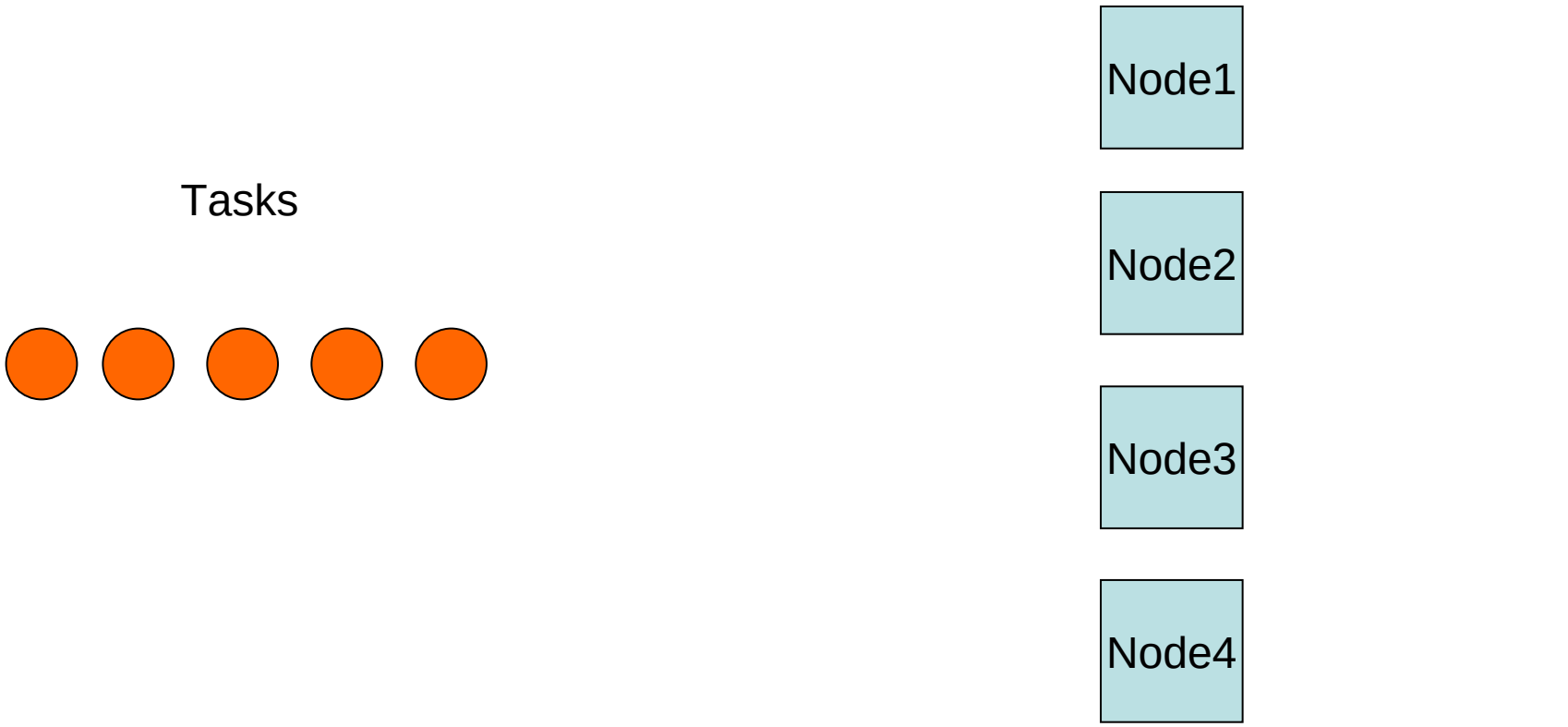
Single-Point-Of-Failure (SPOF)

- 1 Netzteil pro Rechner
- 1 LAN Anschluss
- 1 Switch
- Fehlende USV / UPS
- Keine getrennten Räumlichkeiten
- 1 oder 2 nicht unabhängige Interconnects
- 1 Daten(-anschluss)
- usw.

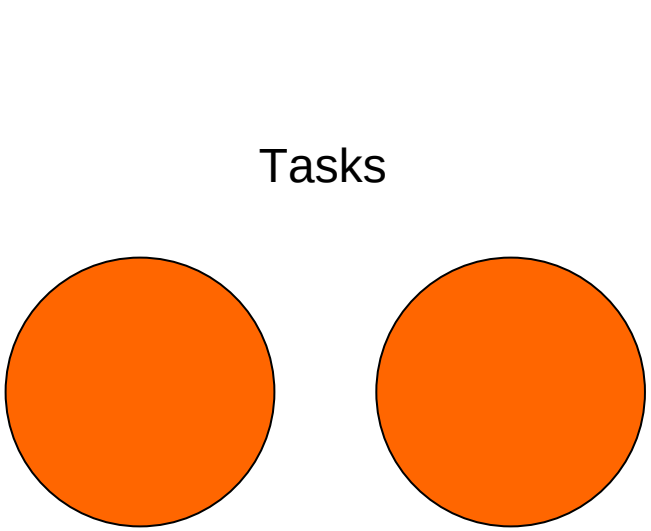
Grundsätze der Verteilung

- Throughput
- Capability
- Shared-Nothing
- Shared-All

Throughput



Capability



Node1

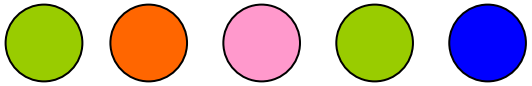
Node2

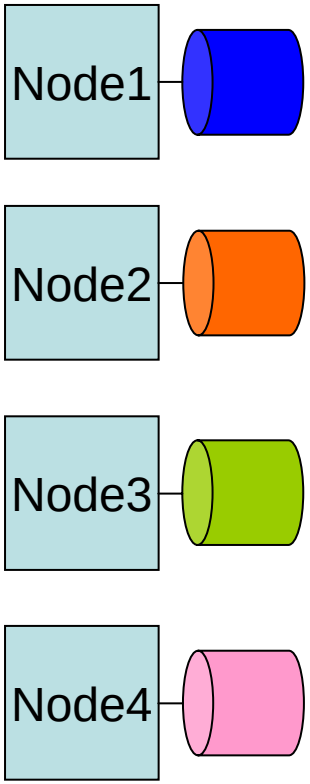
Node3

Node4

Shared-Nothing

Tasks





Cluster Kategorien

- High Performance Computing (HPC)
- High Throughput Computing (HTC)
- Load Balancing (LB)
- High Availability (HA)

- Kombination aus mehreren

High Performance Computing

- Verteilung: Capability
- 1994: Beowulf Project
- Nr. 1 der TOP500:

  IBM BlueGene/L des Lawrence Livermore National Laboratory (LLNL): 131 072 PowerPC 440 a 700 MHz max. 280,6 TFLOPS (P4 3.2GHz: max. 0.006 TFLOPS)
- Software
  - Beowulf
  - Mosix/openMosix
- Einsatzgebiet: Wissenschaftlicher Bereich

High Throughput Computing

- Verteilung: Scheduled Throughput (Capability)
- Parallele Langläufer
- Software
  - Condor
  - Grid Engine (basiert auf „Sun Grid Engine“)
- Einsatzgebiet: Wissenschaftlicher Bereich

Load Balancing

- Verteilung: Throughput
- Methoden
  - Round Robin
  - CPU Load
  - Anzahl Prozesse/Sesions (httpd)
- Software
  - mod_backhand (Apache 1.3 Modul)
  - LVS / Keepalived (Local Director)
- Einsatzgebiet: Webserver

High Availability

- Verteilung: Shared-All
- Verfügbarkeit
  - 99.9% → 8.7 Stunden p.a. nicht verfügbar
  - 99.99% → 52.6 Minuten p.a.
  - 99.999% → 5.3 Minuten p.a. („five nines“)
- Software
  - Heartbeat
  - DRBD
  - Red Hat Cluster Suite
  - Oracle Real Application Cluster (RAC)
- Einsatzgebiet: Kritische Anwendungen (Banken)

Network Balancing

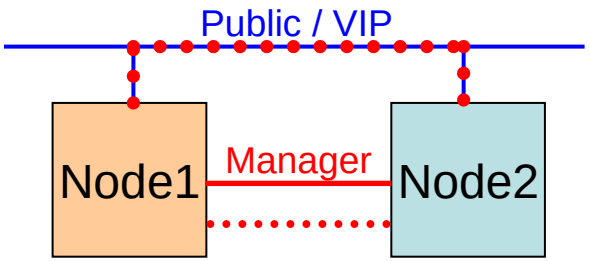
- Verteilung: Throughput
- Synonyme: Link Aggregation, Trunking
- Software
  - TEQL
  - Bonding
  - MultiPathing (Solaris)
- Einsatzgebiet: Network Throughput + TrueHA

Shared File Systems

- GFS (Global File System - RedHat)
- AFS/OpenAFS (IBM)
- Coda (Ursprung AFS2)
- OCFS2 (Oracle Cluster File System)
- NFS (Network File System)
- Lustre (Next Generation File System)

Cluster Network

- 3 IP's pro Cluster-Node
  - Public
    - Ethernet
    - 100 … 1000Mbit
  - Virtuelle IP (VIP)
    - Ethernet
    - 100 … 1000Mbit
  - Cluster Manager / Interconnect
    - redundant / shared
    - Optimal in Durchsatz und Latenz
    - Ethernet / Infiniband
    - 1 … 10 Gbit
    - Datenübertragung



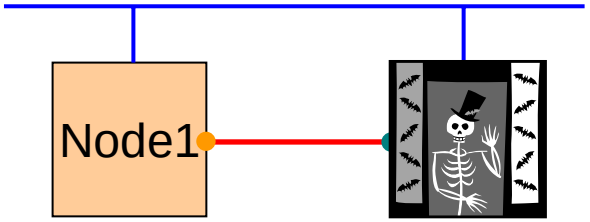
Cluster Manager / Clusterware

- Steuerzentrale
  - Initialisierung
  - Zustandsüberwachung
    - Heartbeat
    - Voting Disk (Quorum)
  - VIP Management
    - Failover
    - Failback
- Selbstregulierendes System
- Dämon Prozess auf allen Nodes

Zustandsüberwachung

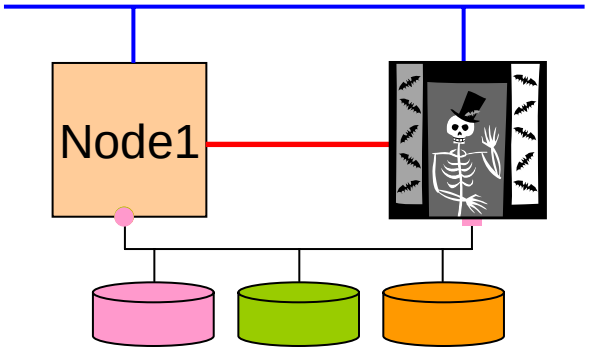
- Heartbeat
  - Ethernet / Serielles Kabel
  - Redundanz

Heartbeat



- Voting Disk (Quorum)
  - > 2 Nodes
  - Eindeutigkeit

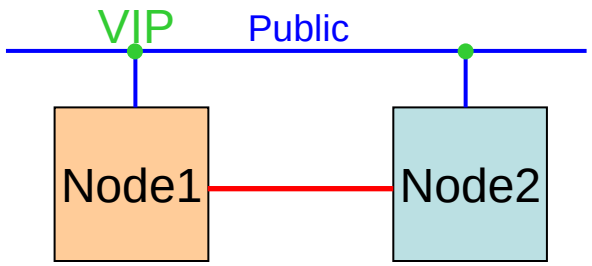
Voting Disk



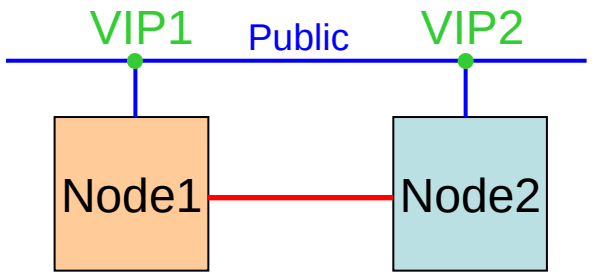
VIP Management

- Anwendungen
  - Active / Passive
  - Active / Active

Active / Passive

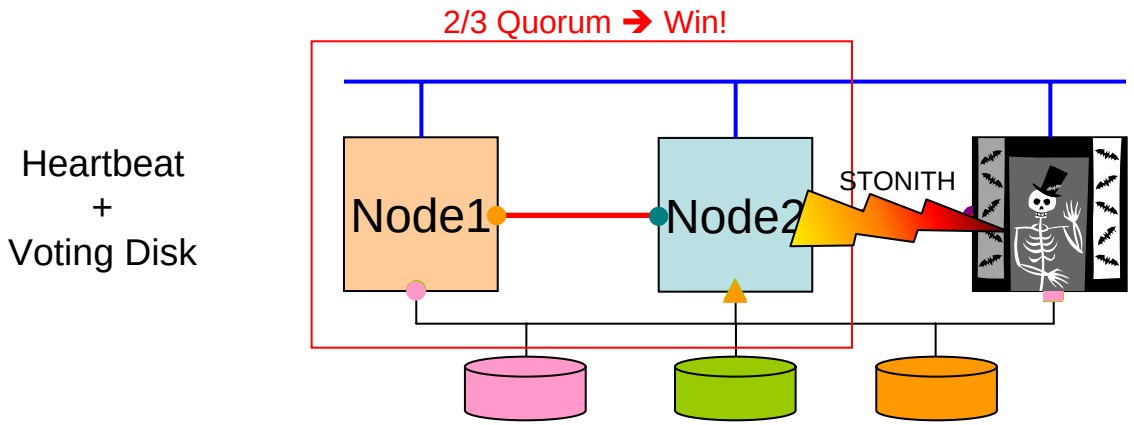


Active / Active



Problematiken I

- Split-Brain
  - Autonom agierende Clusterfragmente
  - Konkurrierendes I/O →Datenkorruption
  - Welcher Clusterteil überlebt?
  - Lösung: I/O Fencing (STONITH, Powerswitch, Zoning)



Problematiken II

- Failover
  - Transparenz bei nicht zustandslosen Anwendungen (Daten im Memory)
- Failback
  - „Neuintegration“ einer Node in das Cluster
  - Datenkonsistenz

Cluster vs. Grid

- Ein Cluster ist **kein** Grid
- Grid ist Weiterentwicklung der HPC / HTC
- Definition (Ian Foster & Carl Kesselmann)
  1. coordinates resources that are not subject to centralized control
  2. using standard, open, general-purpose protocols and interfaces
  3. to deliver nontrivial qualities of service

Quelle: http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf

ISG.EE

System Management, Training Consulting

Linux OpenSource Cluster Projekt

Ein Beispiel aus der Praxis

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

IIET
Departement Informationstechnologie
und Elektrotechnik

Cluster Komponenten

- High Availability
  - DRBD
  - HEARTBEAT
  - RAID5
- Shared File System
  - NFS
- Network Balancing
  - TEQL

DRDB

- DRBD = Distributed Replicated Block Device
- Key Features
  - RAID1 via LAN
  - Transparente Schicht über Standard-Block-Device (Lower Device)
  - Active / Passive (Primary / Secondary)
  - Intelligente Replikation (Metadaten)
  - Transaktionssicher
  - Kernel Modul: drbd
  - Eigene Major Nummer (147)
  - rc-Skript Management

DRBD

- Konfiguration
  - Protokolle
    - Performance vs. Security: write IO is reported as completed, if it has reached …

      A local disk and local tcp send buffer

      B local disk and remote buffer cache

      C both local and remote disk
  - on-io-error handler
    - if the lower level device reports io-error to the upper layers
      - pass_on
      - panic
      - detach
  - Filesystem: XFS, ext3 oder reiserfs
  - 2 x Ethernet

DRBD Konfiguration

```
resource  META {        protocol          B;
                        incon-degr-cmd \"halt -f\";
            startup {
                        wfc-timeout             10;
                        degr-wfc-timeout        60;}
            disk {      on-io-error             detach;}
            net {       sndbuf-size             1M;
                        timeout                 60;
                        connect-int             10;
                        ping-int                10;
                        max-buffers             2048;
                        max-epoch-size          2048;
                        on-disconnect           reconnect;
                        ko-count                10;}
            syncer {    rate                    70M;
                        group                   1;
                        al-extents              257;}
            on storage1 {device                 /dev/drbd0;
                        disk                    /dev/sdd1;
                        address                 100.10.10.1:7788;
                        meta-disk               internal;}
            on storage2 {device                 /dev/drbd0;
                        disk                    /dev/sde1;
                        address                 100.10.10.2:7788;
                        meta-disk               internal; }
```

NFS

- Key Features
  - Einfache Konfiguration
  - (nahezu) zustandslos

- Konfiguration
  - Version 3
  - 32 Threads
  - Export Optionen: `rw,no_root_squash,sync,no_wdelay`
  - Mount Optionen: `noauto,udp,rsize=32768,wsize=32768,intr`

TEQL

- TEQL = Traffic EQuaLizing
- Key Features
  - Basiert auf tc qdisc
    (traffic control settings / queueing discipline)
  - Kernel Modul: sch_teql
  - Eigenes Device: teql
  - rc-Skript Management
- Konfiguration
  - Identische IP-Adresse auf n-Netzwerk-Interfaces
  - Routing via TEQL Device

TEQL Konfiguration

```
#
# TEQL_DEVICES - Devices (eth0, eth1, ....) die mit teqlx verwendet werden
#
TEQL0_DEVICES="eth0 eth2"
TEQL1_DEVICES="eth3 eth5“
#
# TEQL_ADDRESS - ip-addresse von teqlx, muss identisch mit den Adressen der Netzwerkkarten sein
#
TEQL0_ADDRESS="142.168.10.2"
TEQL1_ADDRESS="142.168.100.33"
#
# TEQL_NET - Netz von teqlx, muss ebenfalls identisch sein mit allen Karten
#
TEQL0_NET="142.168.10.0"
TEQL1_NET="142.168.100.0"
#
# TEQL_NETMASK - Netzmaske von teqlx, natuerlich auch identisch
#
TEQL0_NETMASK="255.255.255.0"
TEQL1_NETMASK="255.255.255.0"
#
# TEQL_GATEWAY
#
TEQL_GATEWAY="142.168.100.1"
TEQL_DEVICE="teql1"
```

Heartbeat V1.3

- Key Features
  - 2 Node Cluster Manager (> 2 Nodes ab Version 2.x)
  - rc-Skript Management
- Konfiguration
  - Applikation
    - Active / Active
    - 2 x Ethernet
  - Storage
    - Active / Passive
    - 2 x Ethernet

Heartbeat Konfiguration

- `/etc/ha.d/ha.cf`

```
debugfile /var/log/ha-debug
logfile /var/log/ha-log
keepalive 2
deadtime 30
warntime 10
initdead 60
udpport 694
bcast eth1 eth4
auto_failback on
node app1
node app2
```

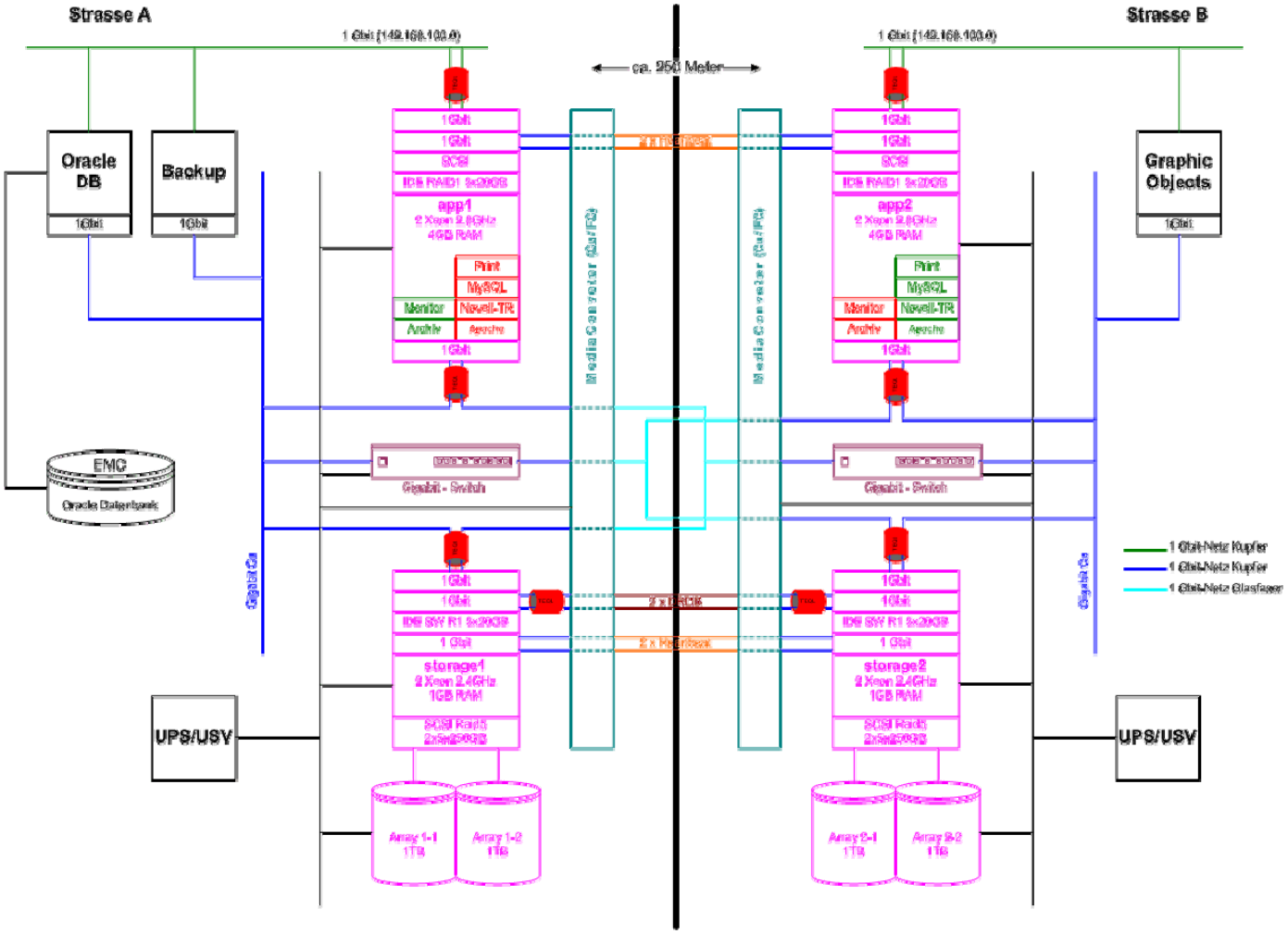
- `/etc/ha.d/haresources`
  - Storage

```
storage1 142.168.10.99 ha_notify::142.168.10.99 drbddisk nas_mount nfsserver
```

  - Applikation

```
app1 142.168.100.117 ha_notify::142.168.100.117 archivd
app1 142.168.100.110 ha_notify::142.168.100.110 monitor
app2 142.168.100.96 ha_notify::142.168.100.96 mysql printd httpd
```

Übersicht


Strasse A
Strasse B
ca. 250 Meter
Oracle DB
Backup
Graphic Objects
app1
app2
storage1
storage2
EMC
Oracle Datenbank
Gigabit - Switch
UPS/USV
UPS/USV
1 Gbit-Netz Kupfer
1 Gbit-Netz Kupfer
1 Gbit-Netz Glasfaser

Stolpersteine I

- Zeitsynchronisation (NTP)
- Heartbeat / Gratuitous ARP

```
##########################################
# New added code for teql devices
##########################################
if `echo "$TARGET_INTERFACE" | egrep -w 'teql[0-9]{1}' 1>/dev/null`; then
    . /etc/sysconfig/network/trafficequalizer.config
    NO=`echo "$TARGET_INTERFACE" | tr -d [a-z]`
    for TARGET_INTERFACE in `eval echo \\\$TEQL${NO}_DEVICES`
    do
      ha_log "info: Sending Gratuitous Arp for $BASEIP on $IF [$TARGET_INTERFACE]"
      ARGS="-i $ARP_INTERVAL_MS -r $ARP_REPEAT -p \
      $SENDARPPIDFILE $TARGET_INTERFACE $BASEIP auto $BASEIP $ARP_NETMASK“
      ha_log "info: $SENDARP $ARGS“
      $SENDARP $ARGS || ha_log "ERROR: Could not send gratuitous arps" \
      && ha_log "info: Send gratuitous arps successfully“
    done
    return
fi
##########################################
# End of: New added code for teql devices
##########################################
```

Stolpersteine II

- NFS (Failover)
  - Stale NFS Handles: Kein “unshareall” beim Stop von nfsd
  - Metadaten (`/var/lib/nfs/*`) müssen auf DRBD-Device liegen
- Crontab

```
#### Global
# MD device check
*/30 * * * * ps -eaf | grep '/sbin/mdadm --monitor /dev/md0 /dev/md1'
| grep -v 'grep' 1>/dev/null || /sbin/startproc /sbin/mdadm --monitor /dev/md0 /dev/md1

#### IP / 142.168.207.117
# start ARCHIV daemon watch script every 15 minutes
*/15 * * * * /sbin/ifconfig | grep 'inet addr:142.168.100.117' 1>/dev/null && /ARCHIV/bin/watch

####  142.168.207.96
# start PRINT daemon watch script every 15 minutes
*/15 * * * * /sbin/ifconfig | grep 'inet addr:142.168.100.96' 1>/dev/null && /PRINT/bin/watch
```

Performance

- Storage-Ebene
  - Native Device
    - Read: 72 MB/s (Char), 105 MB/s (Block)
    - Write: 67 MB/s (Char), 52 MB/s (Block)
  - DRDB Device:
    - Read: 57 MB/s (Char), 69 MB/s (Block) → 70% Native
    - Write: 56 MB/s (Char), 53 MB/s (Block) → 90% Native
  - DRBD Sync:
    - Write: 2x50 MB/s
- Application-Ebene
  - NFS
    - Read: 63 MB/s (Char), 62 MB/s (Block) → 70% Native
    - Write: 43 MB/s (Char), 44 MB/s (Block) → 70% Native

Performancemessung: bonnie++, ausser DRBD Sync: drdadm

Online Maintenance / Update

- Herauslösen des Servers (StorageX oder AppX) aus dem Clusterverbund
  - /etc/rc.d/heartbeat stop
- Unmounten der Cluster-FS (AppX)
  - /etc/rc.d/nas_nfs stop
- Stoppen der DRBD-Devices (StorageX)
  - /etc/rc.d/drbd stop
- **Servicearbeiten durchführen**
- Starten der DRBD-Devices (StorageX)
  - /etc/rc.d/drbd start
- Mounten der Cluster-FS (AppX)
  - /etc/rc.d/nas_nfs start
- Hineinbringen des Servers (StorageX oder AppX) in den Clusterverbund
  - /etc/rc.d/heartbeat start

Zusammenfassung

Pro

- Cluster sind vielseitig einsetzbar (HA,LB,HPC/HTC)
- In modernen Umgebungen und deren Erfordernissen (SLA) nicht mehr wegzudenken
- Kostenoptimierung / Skalierbarkeit
- Grosse Anzahl an OpenSource und kommerziellen Produkten
- Weiterentwicklung: Grid

Contra

- Keine magische Wunderwaffe
- Qualifiziertes Personal
- Hohe Zeit- und Arbeitsaufwände
- Kosten (SPOF)

Fragen …