Solaris Cluster

Dipl.-Inform. Torsten Kasch
<tk@CeBiTec.Uni-Bielefeld.DE>

8. Januar 2008

Agenda

- Übersicht
- Cluster-Hardware
- Cluster-Software
- Konzepte: Data Services, Resources, Quorum
- Solaris Cluster am CeBiTec: HA-Datenbank-Server (MySQL)
- Erfahrungen

Übersicht

- Java Availability Suite
- Sun Plex
- Sun Cluster
- Solaris Cluster
- seit Juni 2007: Source Code über OpenSolaris.org verfügbar

Übersicht (cont.)

Ziele:

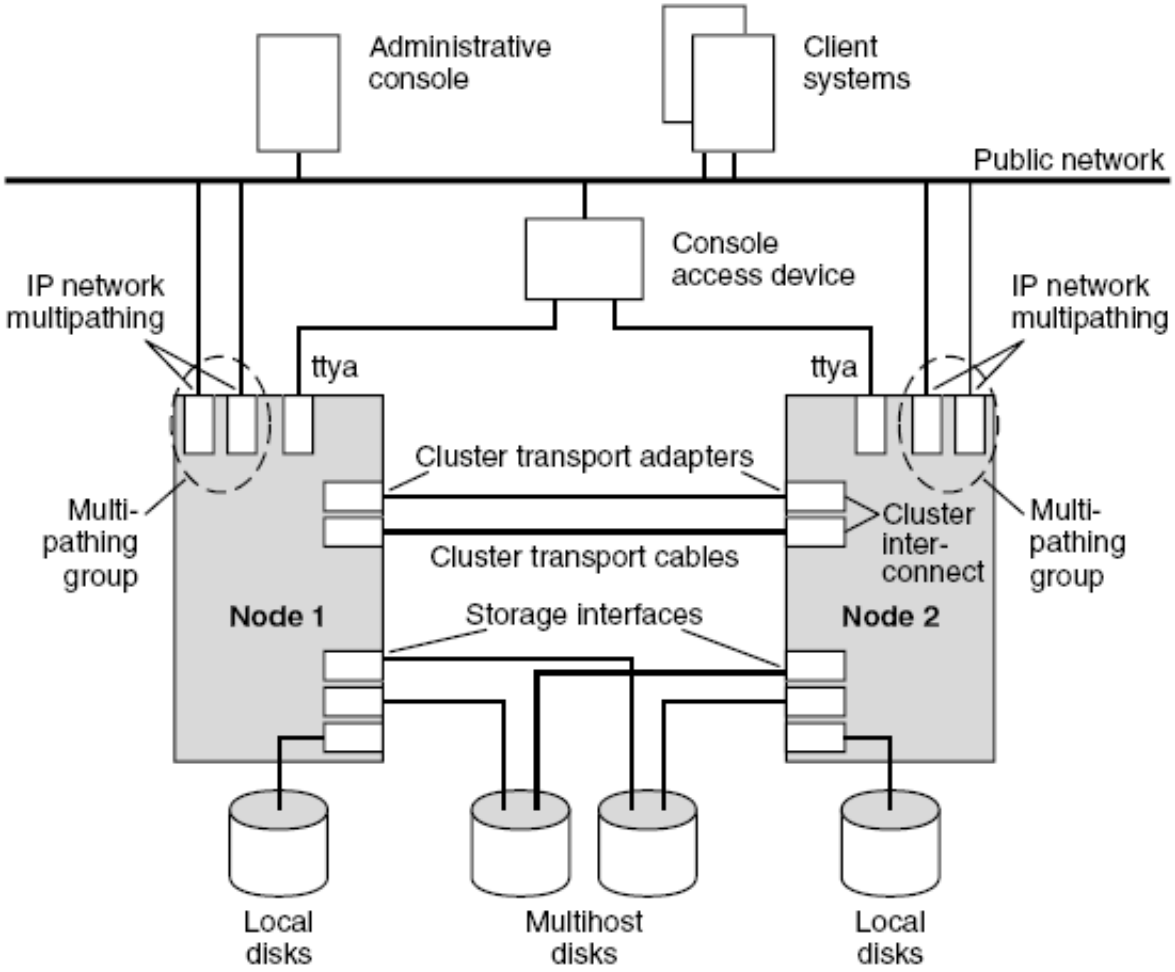
- hohe Verfügbarkeit
  von Diensten/Anwendungen
- Skalierbarkeit von Diensten/Anwendungen

Cluster-Hardware

Ein Cluster besteht aus:

- Cluster Nodes
- Cluster Interconnect
- Public Network Interfaces
- Admin Console
- Multihost Devices (Storage)

Cluster-Hardware (cont.)



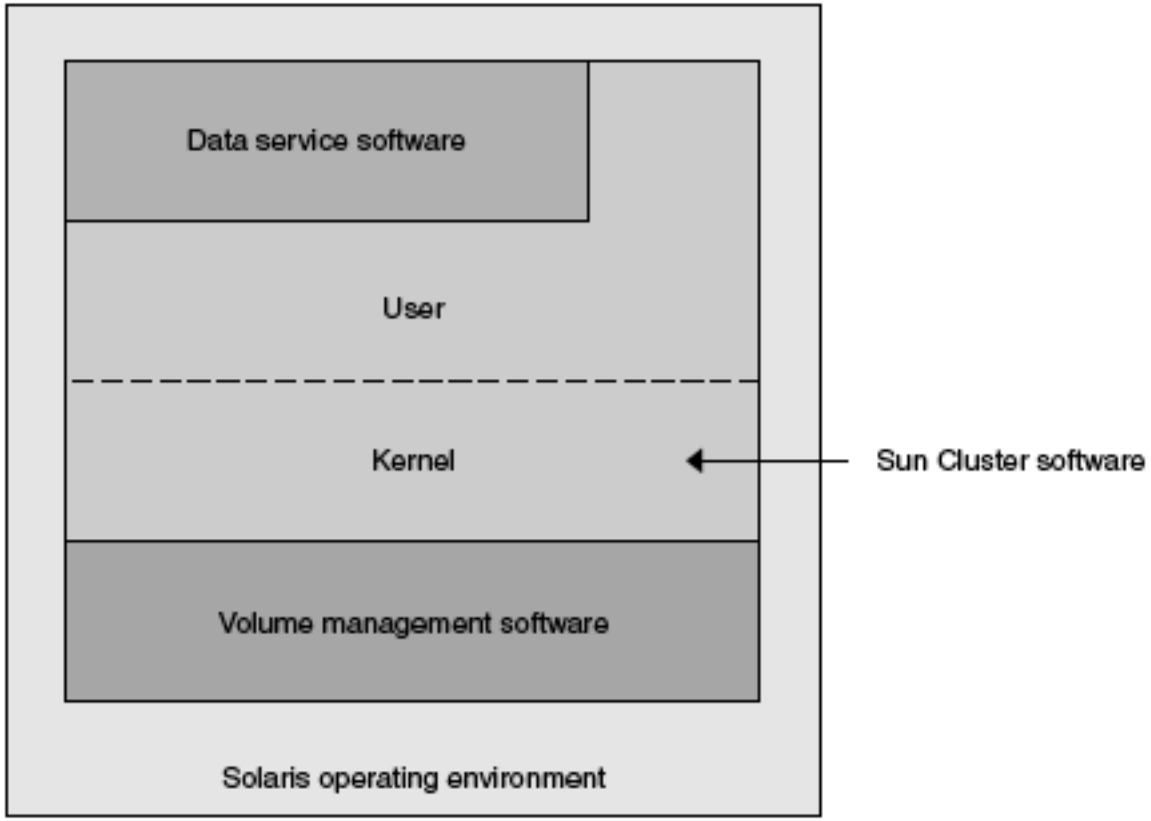
Quelle: Sun Cluster Overview for Solaris OS

Cluster-Software

auf jedem Cluster Node:

- Solaris OS
- Sun Cluster Software
- ggfs. Volume Management Software
- Data Service Application

Cluster-Software (cont.)

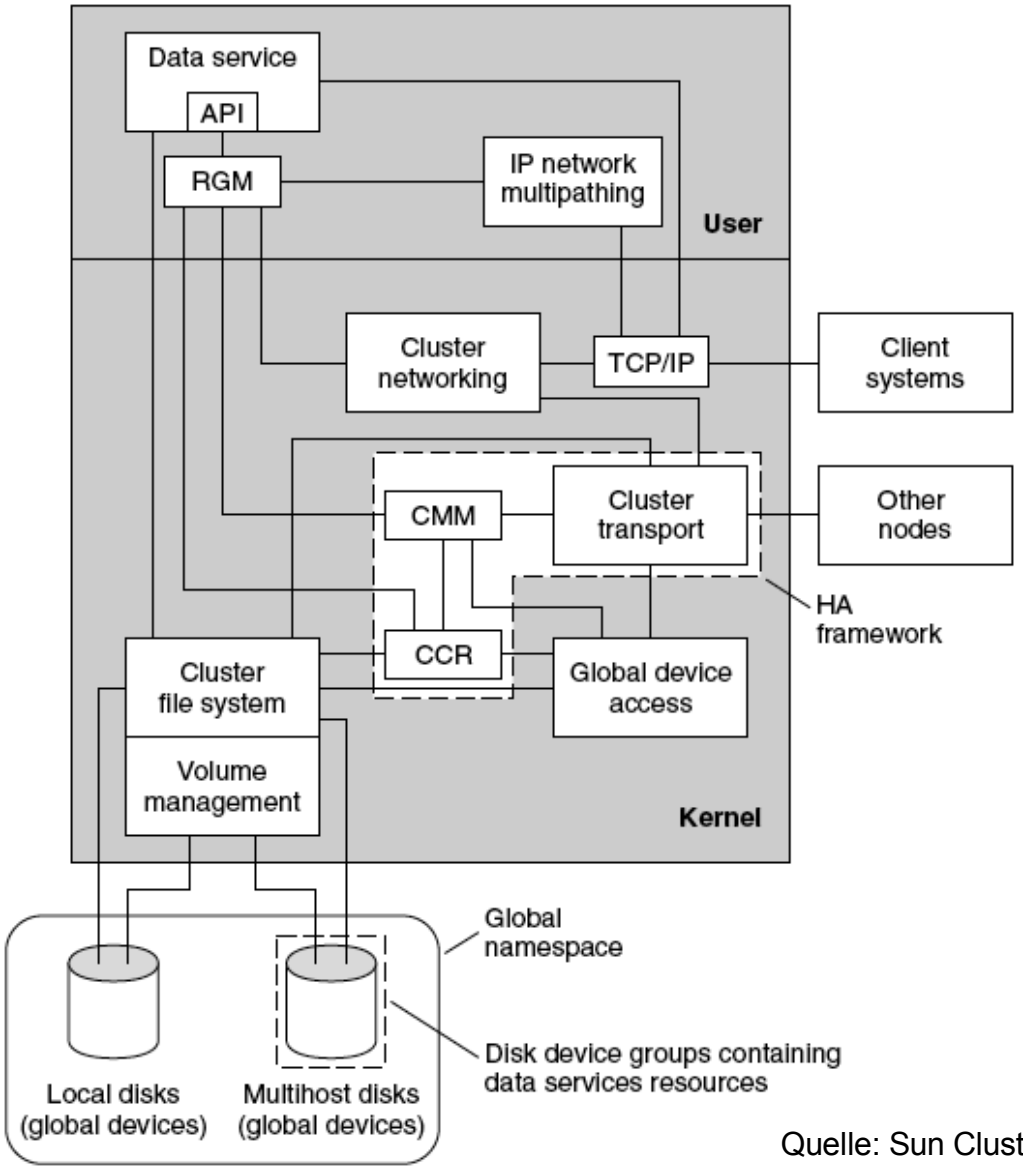


Quelle: Sun Cluster Concepts Guide

Cluster-Software (cont.)

- Cluster Membership Monitor (CMM):
  - verteilte „Agents“ auf jedem Node
  - sorgen für konsistente Sicht aller Nodes
  - deaktivieren defekte Nodes
  - verhindern Partitionierung des Clusters
- Cluster Configuration Repository (CCR)
  - Cluster-Zustand als verteilte „Datenbank“

Cluster-Software (cont.)



Quelle: Sun Cluster Overview for Solaris OS

Konzepte: Data Service Types

- Failover
  - wird automatisch migriert
  - nur eine aktive Instanz der Anwendung
- Scalable
  - mehrere Instanzen laufen gleichzeitig
  - Load Balancing durch Cluster Software
- Parallel
  - „cluster-aware“ Anwendungen (Oracle)

Konzepte: Data Services

- Instanz eines DS Types
- Container für Applikation
- erreichbar über „Logical Hostname“ oder „Shared Address“
- stellt Methoden zur Verfügung: Start, Stop, Monitoring
- Fault-Monitor:
  - Restart des DS
  - Migration des DS

Konzepte: Resource Type

- Sammlung von Attributen
- beschreibt Anwendung oder „Cluster-Objekt“
- vorgefertigte Resource Types:
  - Apache
  - Oracle
  - SAP
  - ...

Konzepte: Resource

- Instanz eines Resource Types
- mehrere Instanzen desselben Typs möglich
- typische DS Konfiguration:
  - HAStoragePlus
  - LogicalHostName

Konzepte: Resource Groups

- Gruppierung von Resource-Instanzen
- ermöglicht Verwaltung als Einheit: Resource Group Manager (RGM) migriert RGs als Ganzes im Failover Fall
- Beispiel: RTs des MySQL-Data Service
  - HAStoragePlus
  - LogicalHostname
  - GDS (Generic Data Service)

Konzepte: Global Devices

- externe „multiported“ Devices (nur Storage-Systeme)
- an mehrere Nodes angeschlossen
- von allen Nodes zugreifbar, hochverfügbar
- „Device ID Driver“ (DID): cluster-weit einheitliche Device-Namen

Konzepte: Device Groups

- Integration mit Volume Manager: Solaris Volume Manager, Veritas
- „Disk Groups“ bzw. „Disk Sets“ können importiert werden
- bei Multipathing: HA-Volumes

Konzepte: Cluster Filesystem

- Abstraktion vom physikalischen FS
- zwischen Kernel/FS auf einem Node und (anderem) „Storage-Node“
- transparenter Zugriff von allen Nodes aus
- hochverfügbar bei Multipath-Anbindung
- Unterstützung von fcntl(2) Advisory Locking

Konzepte: Quorum

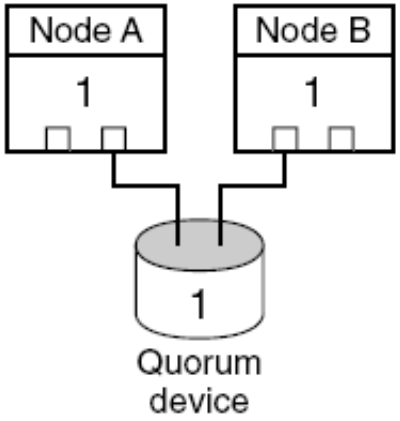
- Redundanz führt zu Problemen:
  - „Split Brain“
    Partitionierung führt zu Datenkorruption
  - „Amnesia“
    inkonsistente Konfiguration der Nodes
- Quorum-Konzept schafft Abhilfe:
  - Welche Nodes formen neuen Cluster nach Partitionierung?

Konzepte: Quorum (cont.)

- Shared Disk* (an min. 2 Nodes)
- „Voting System“
  - jeder aktive Member Node: 1
  - jedes Quorum-Device: N-1
    (N: Anzahl der angeschlossenen Nodes)
- Nodes mit Mehrheit an Votes
  bilden neuen Cluster nach Partitionierung

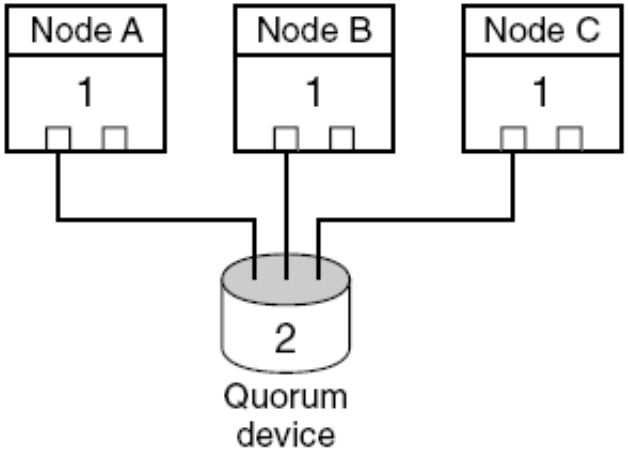
* seit SunCluster 3.2 auch als „Quorum Server“ möglich

Konzepte: Quorum (cont.)



Total configured votes: 3

Votes required for quorum: 2



Total configured votes: 5

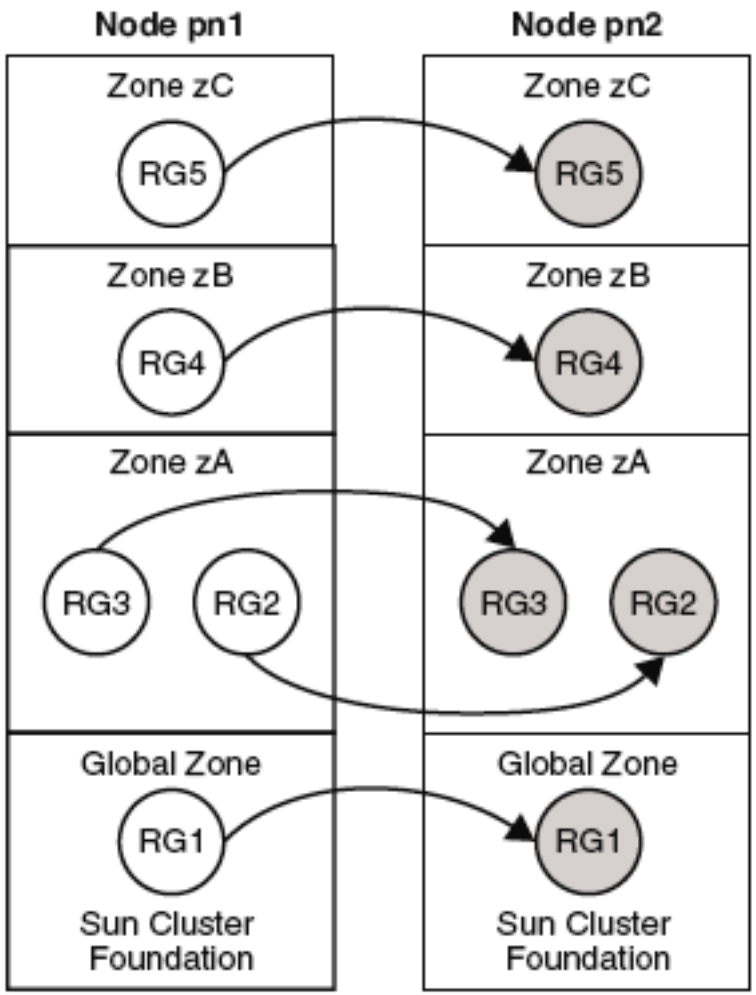
Votes required for quorum: 3

Quelle: Sun Cluster Overview for Solaris OS

Konzepte: Failure Fencing

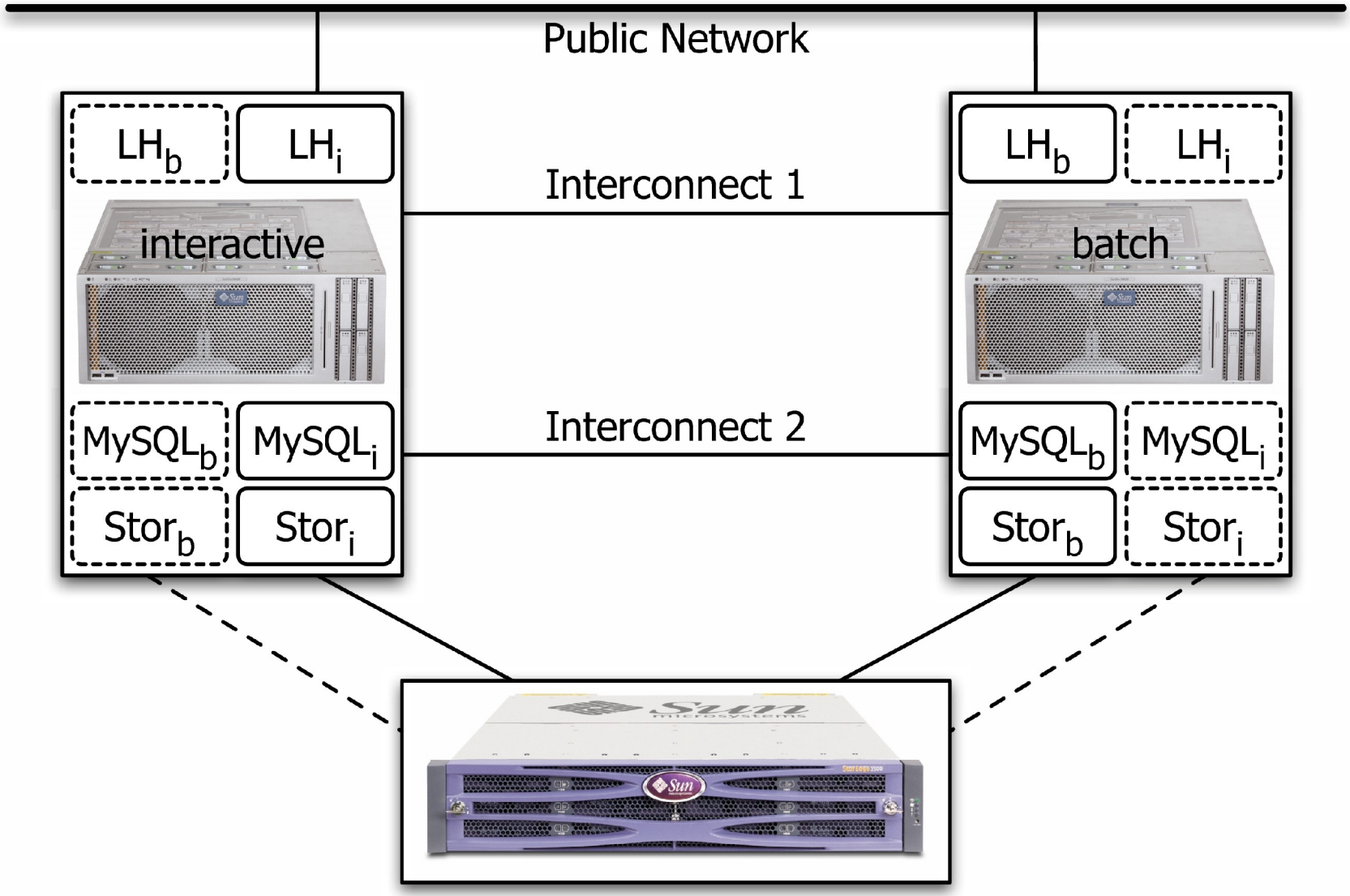
- hindert defekte Nodes,
  auf Multihost-Storage zuzugreifen
- implementiert über SCSI Reservations
- Zugriffsversuche führen zu Panic des OS
- „FailFast“ Mechanismus

Konzepte: Zone-Failover



Quelle: Sun Cluster Overview for Solaris OS

HA-MySQL Service am CeBiTec



HA-MySQL Service am CeBiTec

Demo

Erfahrungen...

- seit 08/2007 in Produktion
- bisher einziges Problem:

```
[...]
${MYSQL_MYISAMCHK} -c -s ${MYSQL_DATADIR}/*/*.MYI
[...]
```

Erfahrungen...

- seit 08/2007 in Produktion
- bisher einziges Problem:

```
[...]
${MYSQL_MYISAMCHK} -c -s ${MYSQL_DATADIR}/*/*.MYI
[...]
```

```
root@zed-batch # echo ${MYSQL_DATADIR}/*/*.MYI | wc -c
3072553
```

Vielen Dank

für Eure Aufmerksamkeit!