Samba as Active/Active HA-Service

Dipl.-Ing. Thomas Merz

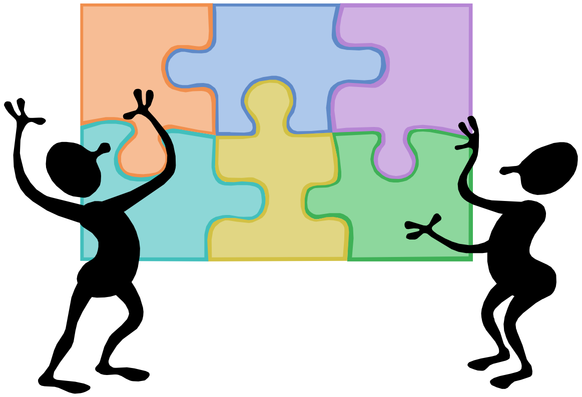
Date: 04/26/2006

merz@atix.de

- About ATIX
- HA-Cluster Basics
- Cluster Filesystem Basics
- Samba and Clusterfilesystem GFS
- Perspective

ATIX business segments

- Consulting
  - Linux in the datacentre (Cluster-solutions, HA)
  - Storage networks
  - Availability analysis / Catastrophe precaution
- Services
  - Competence Center
  - Proof of Concept
  - Project attendance
  - Installation / Production
  - Workshops

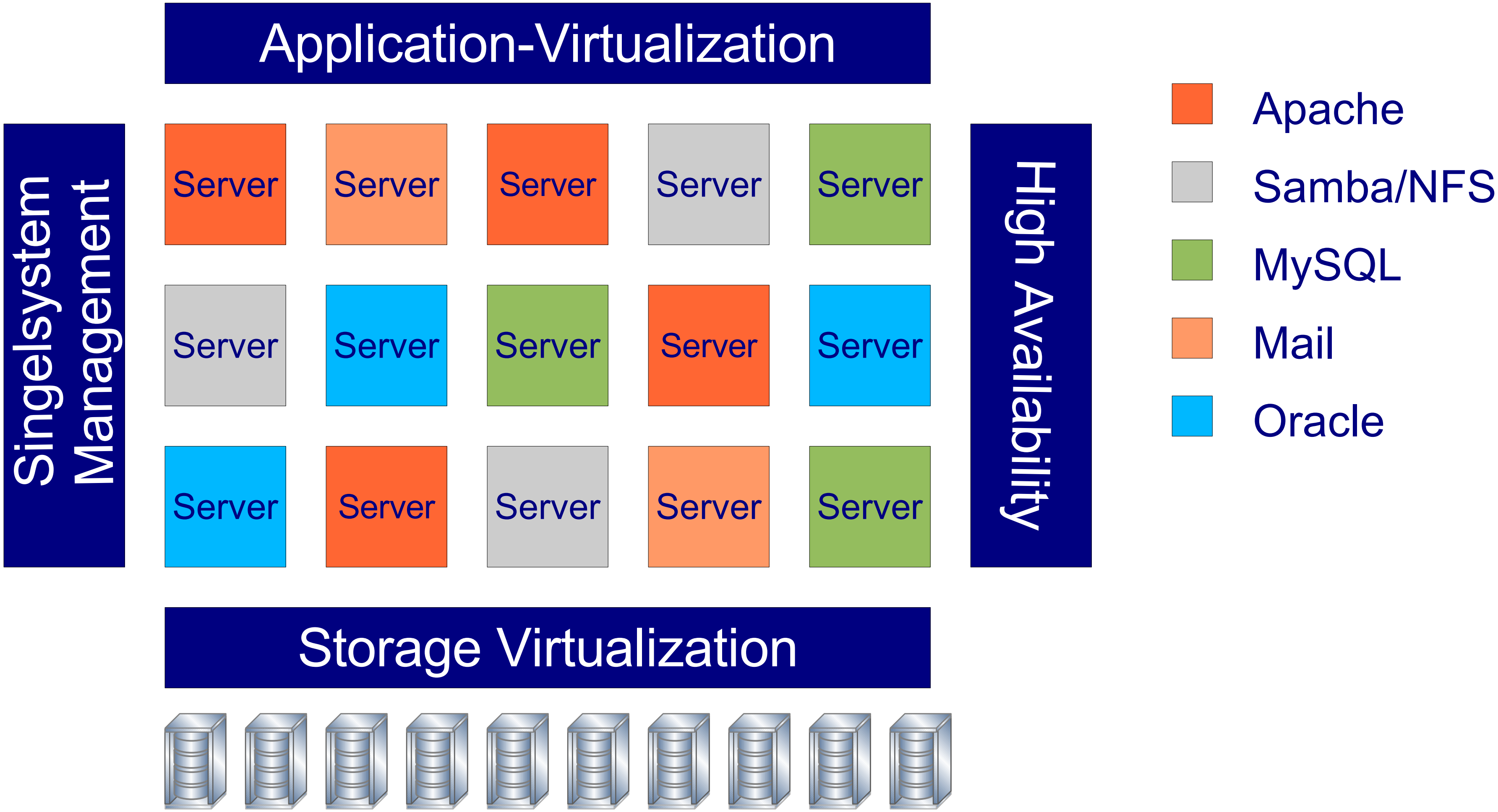


ATIX

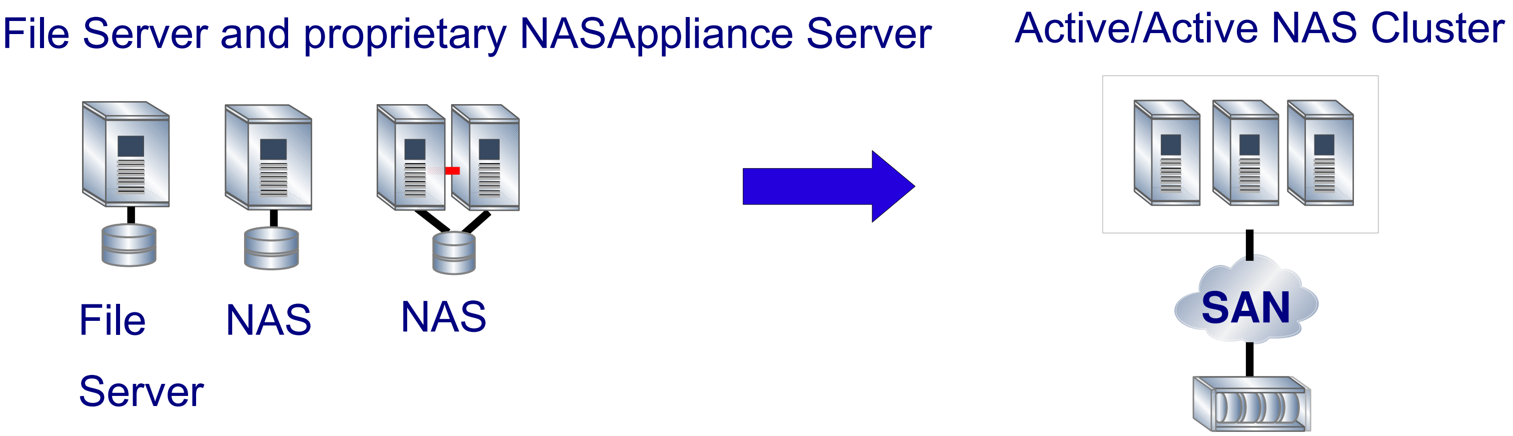
ATIX – couple of references

- Trade Fair Leipzig
  - Infrastructure for Unix/Windows user- and group data of the employees
  - High Availability platform
- IP-Tech
  - Infrastructure for Internetservice Provider (TOP 5, CH)
  - Business Continuance
- Trade Fair Munich International
  - Infrastructure for Webservices
  - High Availability platform
- Int. Pharma Group
  - Consulting for Pharma-IT Storage environment
  - Concepts for catastrophe precautions

Modular conception of *Enterprise IT-platforms*



Case Study: NAS



File Server and proprietary NASAppliance Server

Active/Active NAS Cluster

File Server

NAS

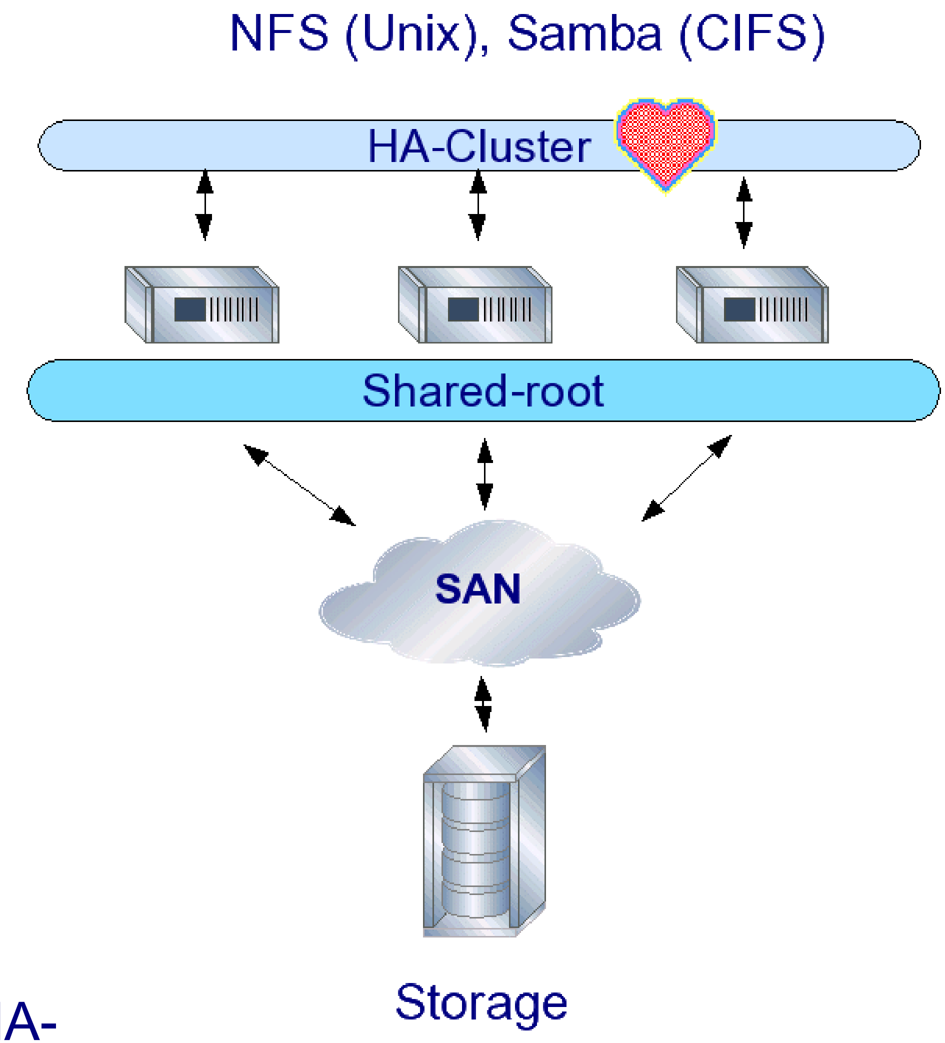
NAS

SAN

- Reduction of capital lockup
- Better utilization of Ressourcen
- Protection of investment
- Scalability as needed
- Better Availability
- Industry Standard Hardware

Example: Trade Fair Leipzig

- **Key Features:**
- Parallel NFS-Server
- Active/Active NFS, CIFS/SMB
- Active Directory Integration
- Dynamic Windows/Unix User Mapping
- ACL Support
- User, Group Quota
- ~ 300 User
- Home and Group Shares
- Replacement for a Windows HA-Cluster solution

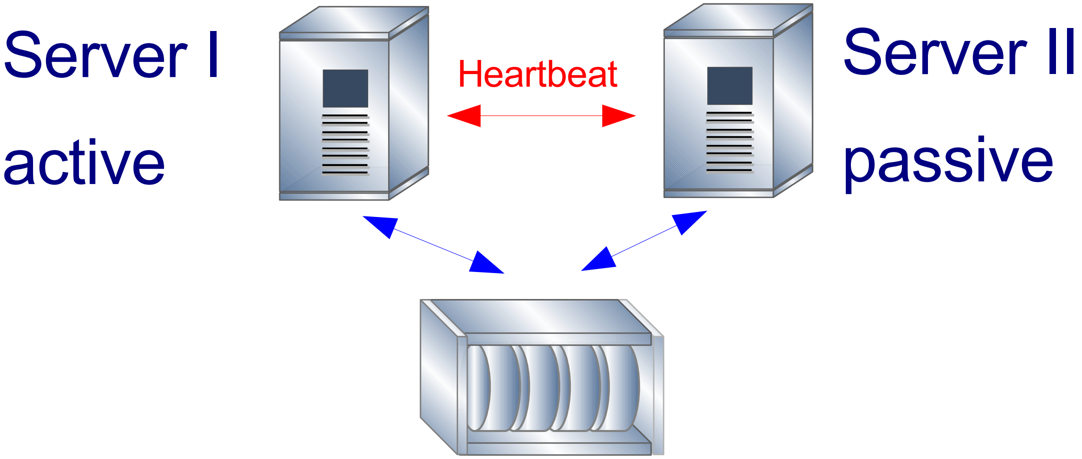


Customer's benefits: Performance-Increasement, Reduction of costs, better availabilty

ATIX

- About ATIX
- <u>HA-Cluster Basics</u>
- Cluster Filesystem Basics
- Samba and Clusterfilesystem GFS
- Perspective

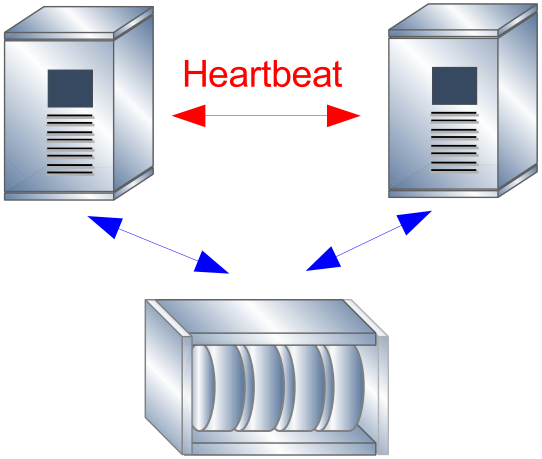
HA Cluster Active/Passive



Concept:

- Only one node active at any time
- The second node is in stand-by mode
- No performance cutbacks in case of node failure

HA Cluster Active/Active



Server I

active/(passive)

Heartbeat

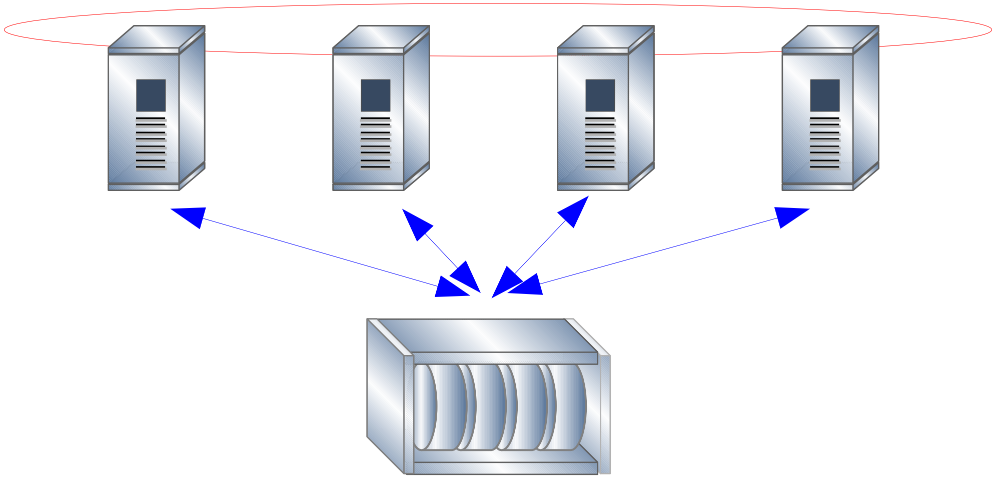
Server II

(passive)/active

Concept:

- Each node hosts different services
- Each node is active and passive
- Performance cutbacks in case of a failure

HA Cluster N+1



Heartbeat Network

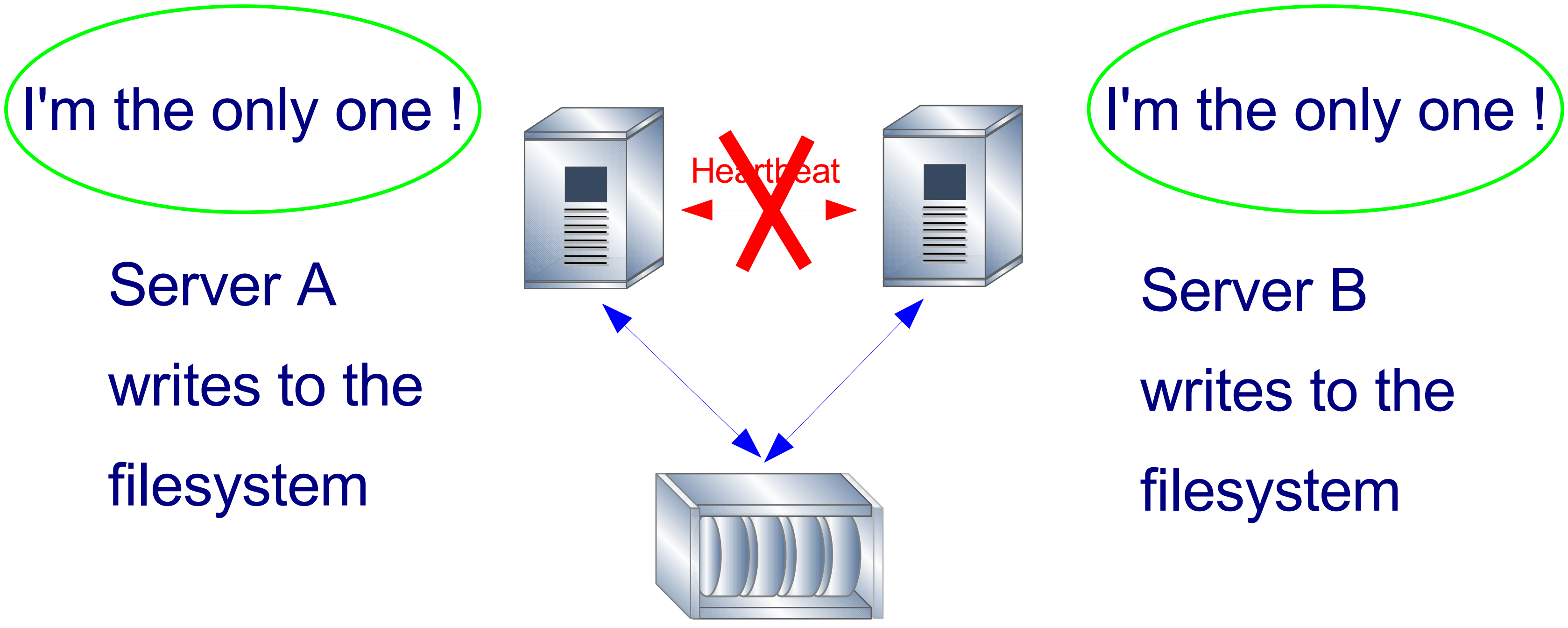
Shared Storage

Concept:

- More nodes as necessary are used (N+1)
- A node can be down without a performance cutback of services
- N+2-, N+3- concepts are possible

HA Cluster: Split Brain Problem

- Questions: „What happens, if the cluster falls apart ?“



I'm the only one !

Server A

writes to the

filesystem

I'm the only one !

Server B

writes to the

filesystem

Data corruption ?!?

Stateless and Stateful Services

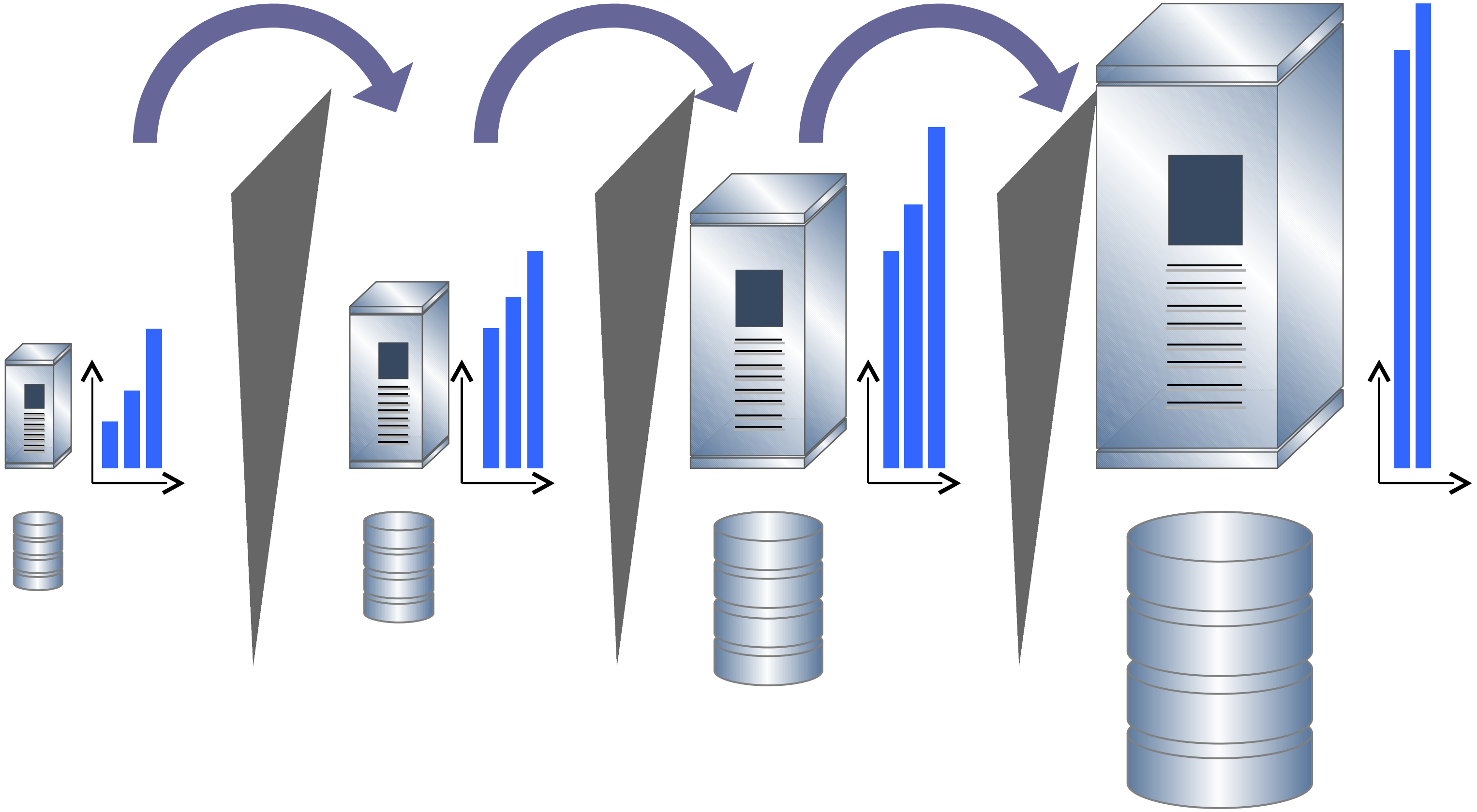
- Problem of transparent failover
- The service has to continue with the exact data states on the 2nd node as it „left“ the 1st node
- Stateless services don't have data in memory
  - => Transparent failover is no problem
- Services saving conditions in memory need a way to make them persistent

Stateless and Stateful Services

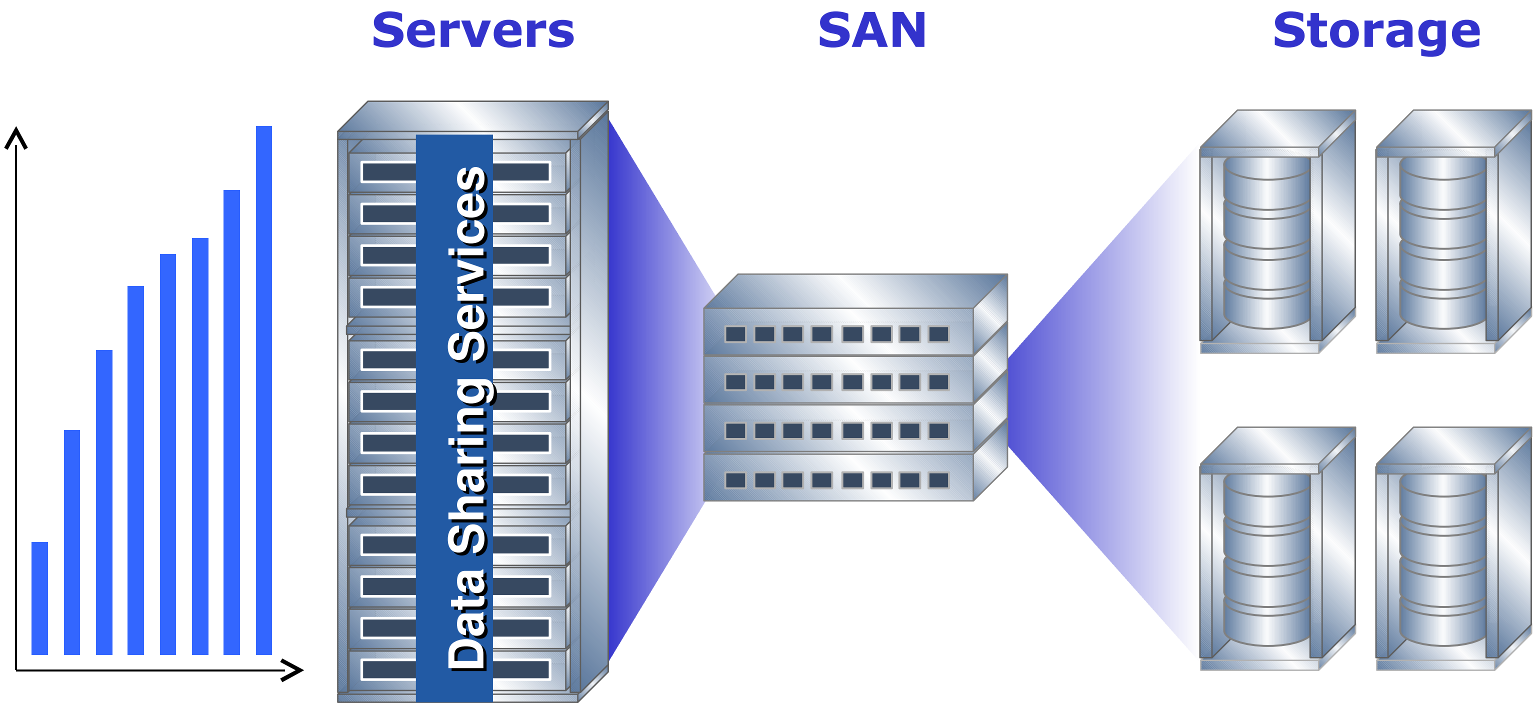
- Stateful services:
  - Perfect example: DBMS
  - Solution: Write-Logs, Redo-Logs etc.
  - Services like NFS/CIFS can be stateful
  - HA-Software needs compatibility modes to failover stateful services correctly
    - Realization via resource types
    - Data loss happens if failover mechanisms don't support stateful services

- About ATIX
- HA-Cluster Basics
- Cluster Filesystem Basics
- Samba and Clusterfilesystem GFS
- Perspective

Scalability



Scalability of a Storage-Cluster

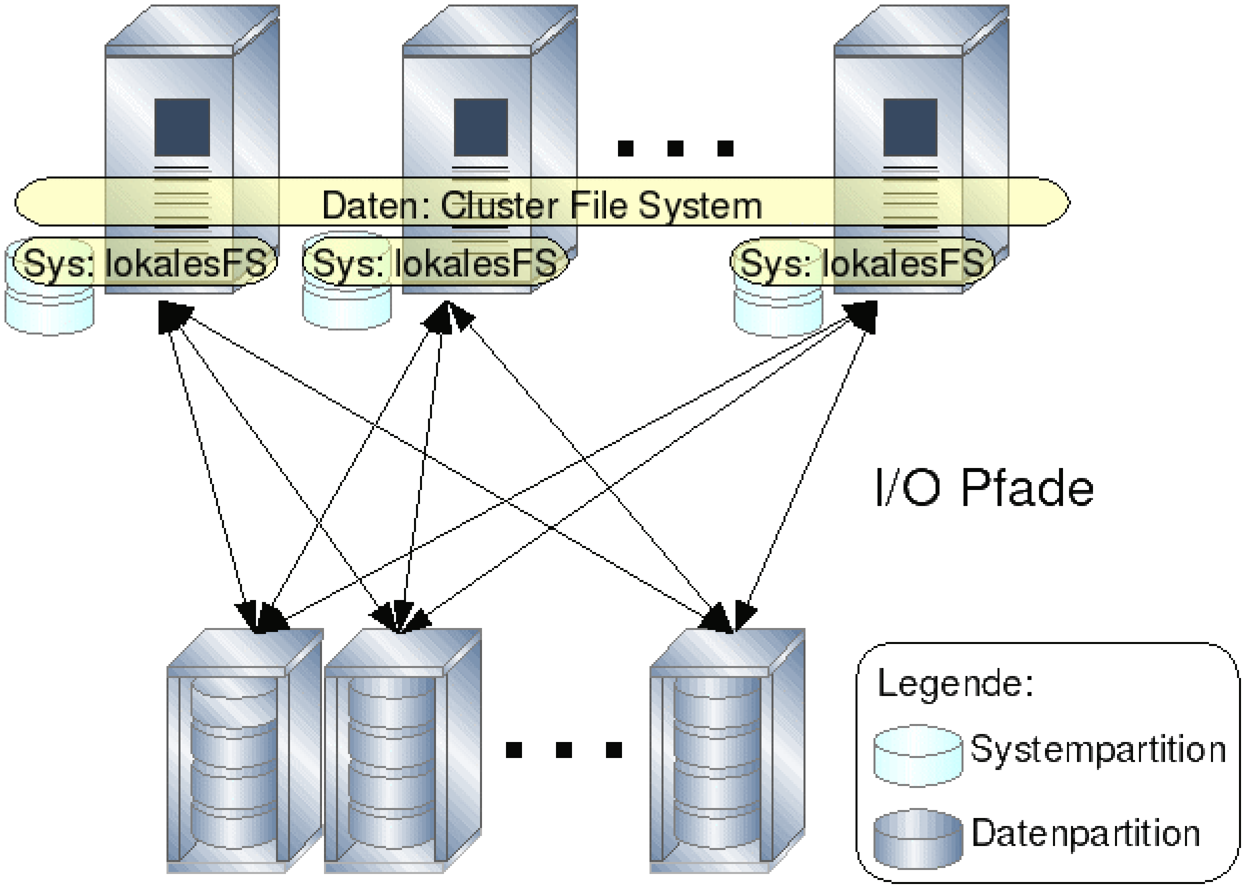


SAN + Linux + Data Sharing = Incremental Computing

- Incrementally and independently add compute, I/O and storage capacity
- Avoid architectural or application changes
- Lower cost of deployment and management

Shared Storage Cluster

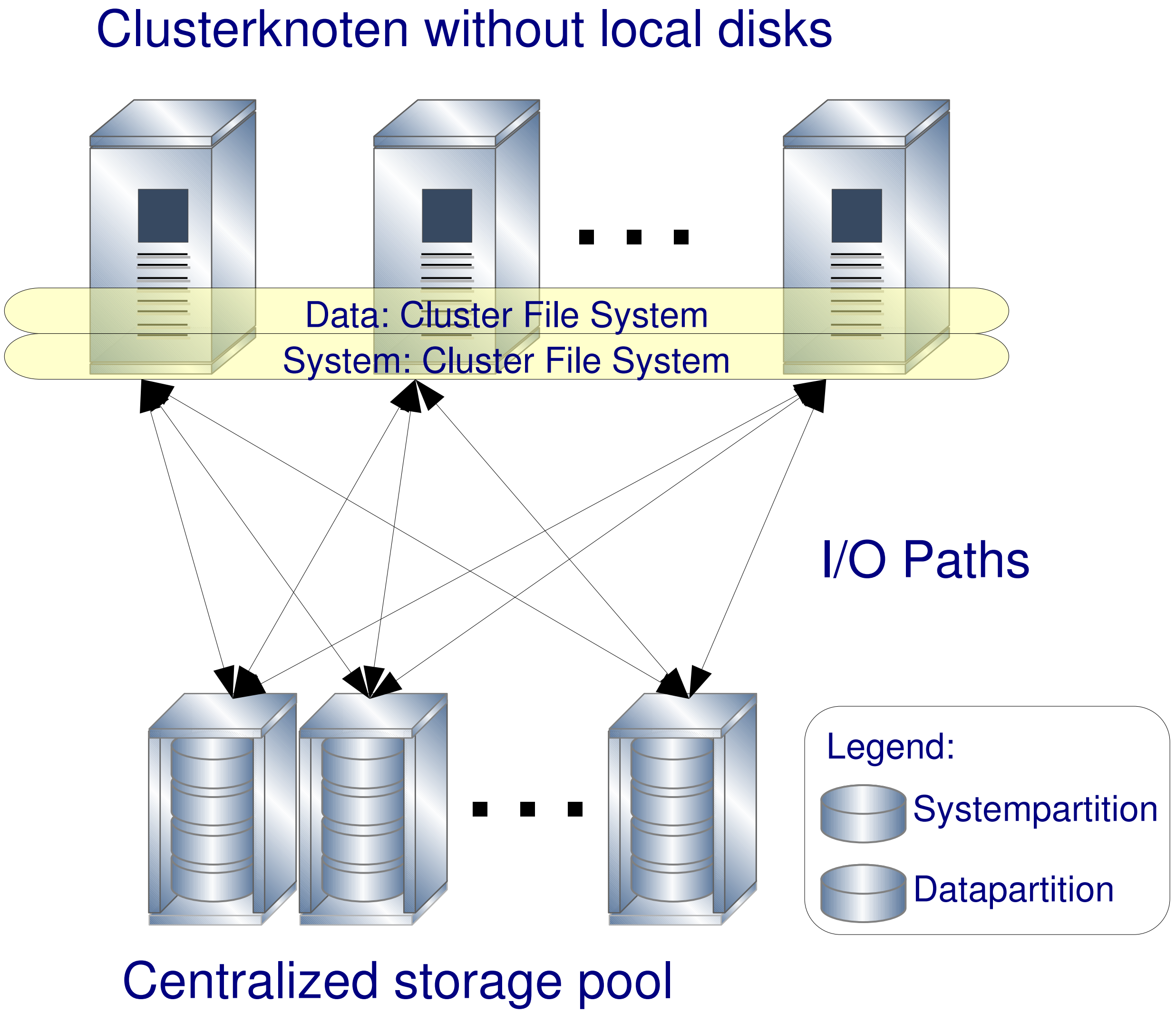
Clusternodes with local disks



Centralized storage pool

- Data Sharing
- Peer to Peer communication
- Cluster Filesystem
- Server Cluster
  - Active/Active
- Storage Cluster
  - Storage Pooling
  - Volume Manager
- Storagenetwork (FC-SAN)
- Management ??

Diskless Shared Root Cluster


Clusterknoten without local disks
. . .
Data: Cluster File System
System: Cluster File System
I/O Paths
. . .
Legend:
Systempartition
Datapartition
Centralized storage pool

- Data Sharing
- Shared Root Partition (/)
- Cluster Filesystem
  - Data
  - System
- SSI on FS Level
  - Management !!
- Scalability
- Performance
- Storage Cluster

Global Filesystem (GFS)

- Development since 1995
  - University of Minnesota - Sistina - Red Hat
  - Version 6.1
- Symmetrical cluster filesystem
- POSIX compatible filesystem
- Direct IO (Databases)
- HA Locking Server (GULM)
- Distributed Lock Manager (DLM)
- Online resizeable (no downsizing!)
- Context Dependent Path Names
- Cluster Volume Manager
- ACLs, Quotas, Multipath, ...

- About ATIX
- HA-Cluster Basics
- Cluster Filesystem Basics
- Samba and Clusterfilesystem GFS
- Perspective

Samba Challenges

Active/Active

- Servertype
  - Domain/ADS Member
  - PDC/ADS Server??
- Usermapping
  - Persistent (LDAP, RID-mapping, ..)
- Filesystem
  - ACLs
- Parallel Access???
  - Share1->Server1 Share2->Server2

Samba Challenges

Active/Active

- TDB-Files
  - GFS vs. no GFS
- Single Sign-On for Windows und Unix/Linux
  - Windows user (PDS/ADS) -> Server type
  - Unix user (passwd, yp/nis, ldap) -> Server type
- Virtual name and server name
- Config file per share vs. global config file
- Winbind failover vs no winbind failover

Samba and Cluster Filesystems

- Posix-ACLs map Windows rights to Unix filesystems
  - Windows clients differentiate between NT-ACLs and Windows 2000 ACLs
    (acl compatibility = auto|winnt|win2k)
- Temporary Samba files (tdbs) need to be host dependent in a cluster setup (CDSL)
  - They can also be stored in the RAMFS to gain better performance

Samba Active/Active

- Samba Active/Active on the same share?
  - Different servers should not export the same shares in r/w mode (ro makes sometimes sense)
  - Parallel r/w is no problem for the cluster filesystem
  - Parallel r/w is a problem for Samba itself, if the upper level application does not have its own locking mechanisms. Samba has no „cluster wide“ locking mechanism
  - With the help of a cluster filesystem, shares can be moved easily between different servers

Usermapping

- What ist mapped?
  - Windows user IDs (SIDs) to Unix user-IDs (winbind)
    - Static tables
    - Via LDAP
    - Via RID mapping
    - Persistent mapping is very important for HA-clusters
  - Unix user to windows user ID mapping is done repeatable & dynamically

Single Sign On

- Identical Users (Names, passwords and identities)
  - Linux must be able to compare against „Windows passwords“ (Kerberos)
    - PAM, Winbind
  - SIDs/RIDs need to be mapped to Unix UIDs/GIDs
    - Static mapping e.g. *Administrator=>root*
    - Automated mapping e.g.
      - *Windows User thomas => Unix User thomas*

Single Sign On

- Samba offers IDMap Backends for user authentication
  - Standard: TDB
    - Mapping is not persistent
  - Alternative: LDAP
    - Certain schema with Unix UserID and Windows SID
  - Alternative: RID mapping
    - Windows User ID has a SID part and a RID part
    - The RID part is mapped repeatable to Unix UIDs

Some Samba Pitfalls

- Servernames and VIPs
  - The servername is associated with a SID and is registered as computer within the domain
  - The virtual clustername (the virtual IP) must not be identical to the servername
- Config files
  - Standard: One config file for a virtual clustername/virtual IP setup (multiple smbd/nmbd services set up on the server, one per failover group)
  - Alternative: One config file is used for all virtual HA-Samba configurations and adjusted if failover is necessary (one smbd/nmbd service running on the server)

Winbind Failover vs. no Failover

- Each Samba service uses the same instance of winbind
  - Standard: All Samba services use this „centralized“ winbind
  - Winbind „does not failover“
- Each virtual clustername/virtual IP samba instance uses its own version of winbind
  - Winbind „fails over“ if the associated virtual clustername fails over

- About ATIX
- HA-Cluster Basics
- Cluster Filesystem Basics
- Samba and Clusterfilesystem GFS
- Perspective

Perspective

- Using GFS and sharedroot cluster configurations are changeable while the cluster is online
- CIFS and NFS are only some of the possibilities such a cluster offers
- If Samba could handle file locking on file basis in cluster compatible way, new cluster types would be possible

Any Questions?

Thank you!

Atix GmbH
Einsteinstr. 10
85716 Unterschleißheim
www.atix.de
info@atix.de