



Herzlich Willkommen
zum Vortrag

Active/Active HA Clustering mit einem Cluster-Filesystem am Beispiel von Samba

Referent: Marc Grimme

Datum: 04.03.2006 / Chemnitzer Linux-Tage

grimme@atix.de



Inhalt

- Über ATIX
- HA-Cluster Grundlagen
- Cluster Filesystem Grundlagen
- Samba und das Clusterfilesystem GFS
- Ausblick



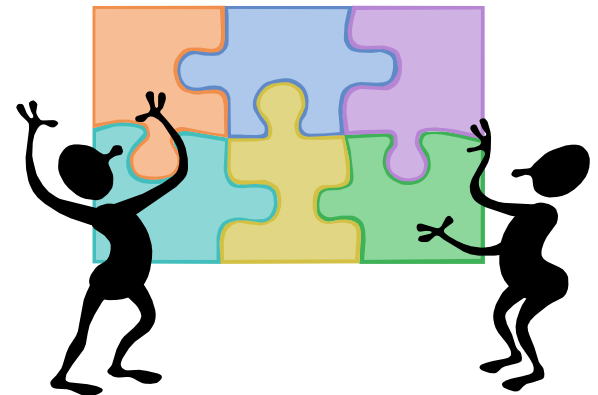
ATIX Geschäftsfelder

- Beratung

- Linux im Datacenter (Cluster-Lösungen, HA)
- Speichernetzwerke
- Verfügbarkeitsanalyse / Katastrophenvorsorge

- Services

- Competence Center
- Proof of Concept
- Projektbegleitung
- Installation / Produktion
- Workshops / Schulungen



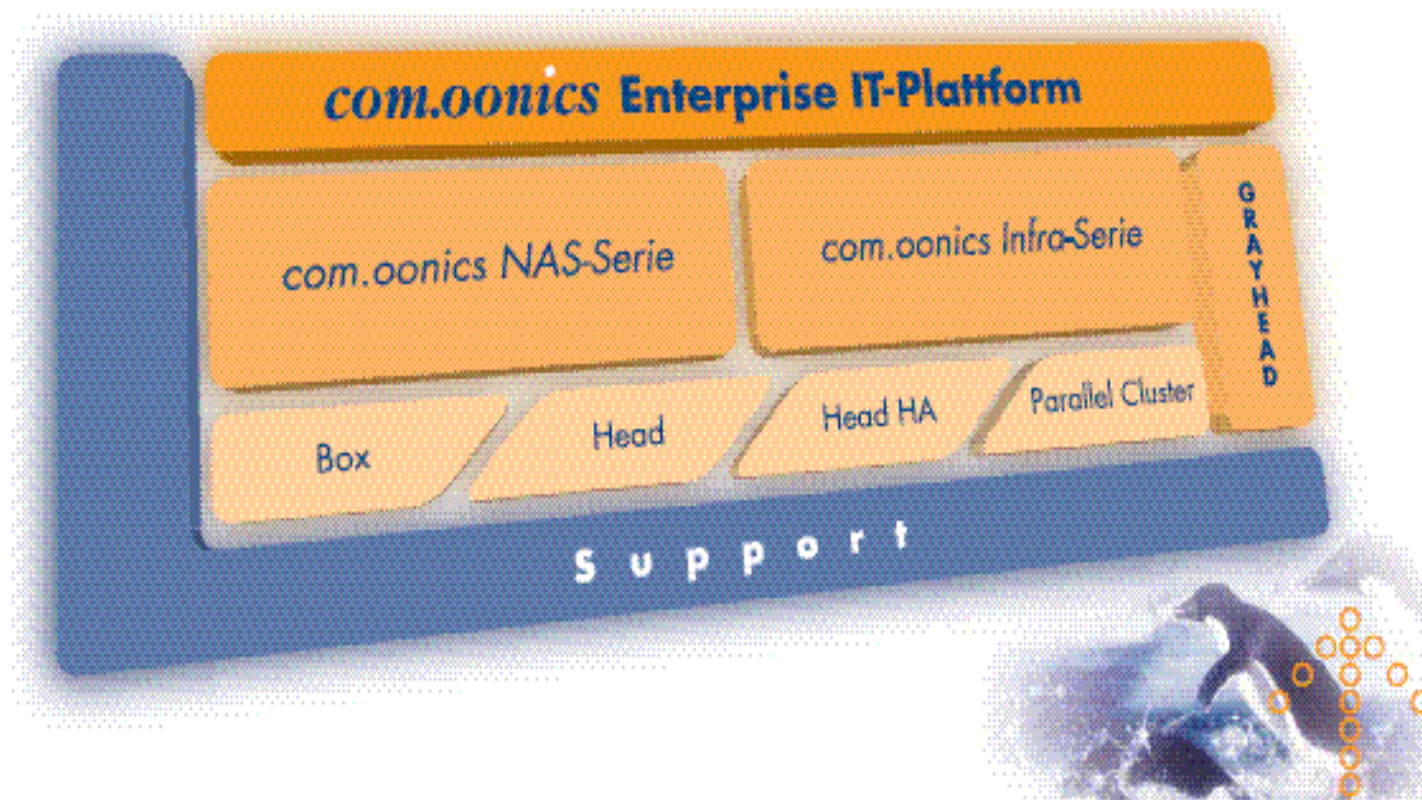


ATIX – ausgewählte Referenzprojekte

- Com.oonics Enterprise IT-Plattform
 - Leipziger-Messe
 - Infrastruktur für die Unix/Windows Benutzer- und Gruppendaten der Mitarbeiter der Leipziger Messe
 - Hochverfügbarkeitsplattform
 - IP-Tech
 - Infrastruktur für Internetservice Provider (TOP 5, Schweiz)
 - Business Continuanace
 - Messe München International
 - Infrastruktur für die Webservices der Messe
 - Hochverfügbarkeitsplattform
- Int. Pharma Konzern
 - Consulting für Pharma-IT Speicherlandschaft
 - Konzept zur Katastrophenvorsorge



com.oonics Enterprise IT- *Plattform*

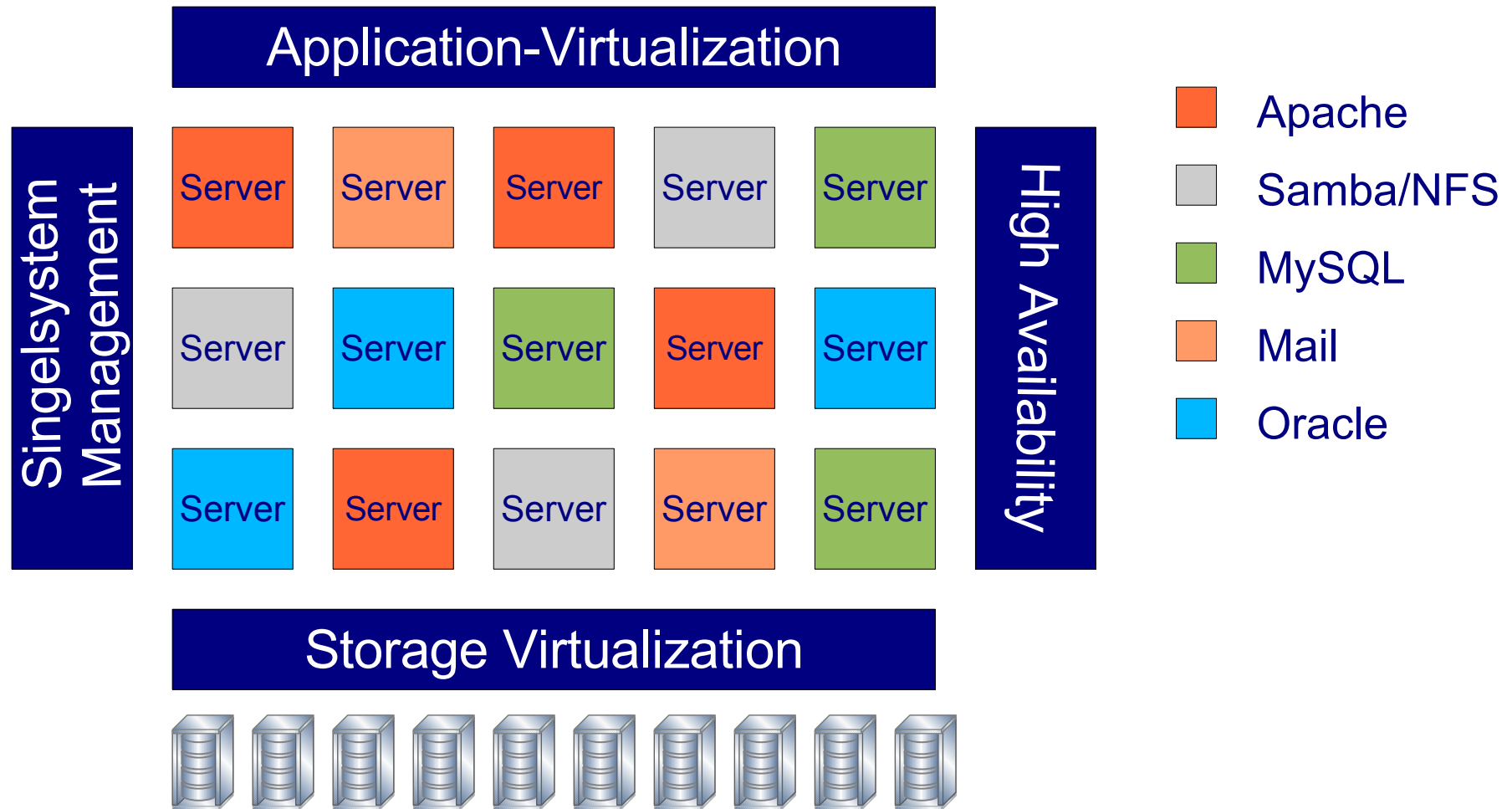




com.oonics Enterprise IT-Plattform

- Modulare Linux IT-Plattform
 - com.oonics NAS-Serie
 - com.oonics NASBox, NASHead, NASHead HA, NASCluster
 - ▶ Maßgeschneiderte NAS-Lösung für NAS-Applikationen
 - com.oonics Infra-Serie
 - com.oonics InfraBox, InfraHead, InfraHead HA, InfraCluster
 - ▶ Multipurpose Server für verschiedenste Einsatzgebiete (z.B. Datenbanken, Web-Server)

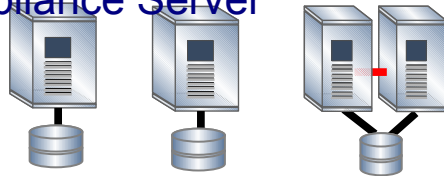
Das Baukastensystem der **nb-team** *com.oonics* Enterprise IT-Plattform



Case Study: NAS

File Server und proprietäre NAS

Appliance Server



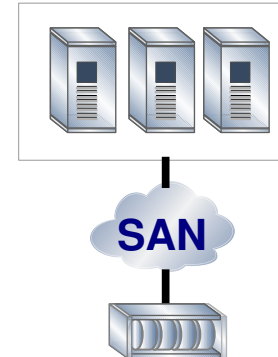
File
Server

NAS

NAS



Active/Active NAS Cluster



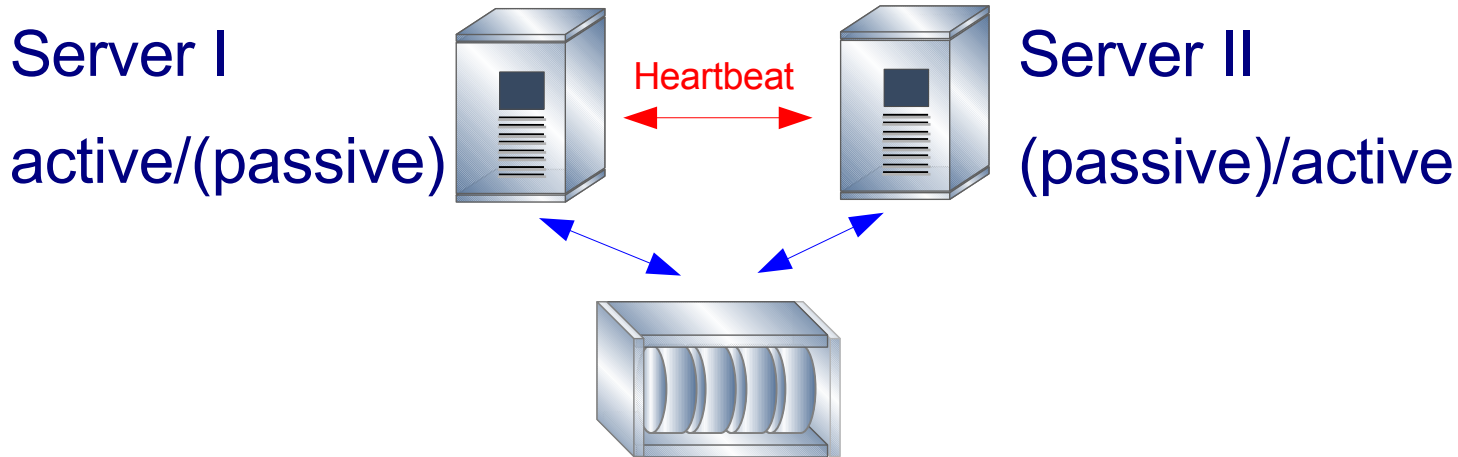
- Reduzierung der Kapitalbindung
- Bessere Ausnutzung der Ressourcen
- Investitionsschutz
- Bei Bedarf Skalierbarkeit
- Höhere Verfügbarkeit
- Industrie Standard Hardware



Inhalt

- Über ATIX
- HA-Cluster Grundlagen
- Cluster Filesystem Grundlagen
- Samba und das Clusterfilesystem GFS
- Ausblick

HA Cluster Active/Active

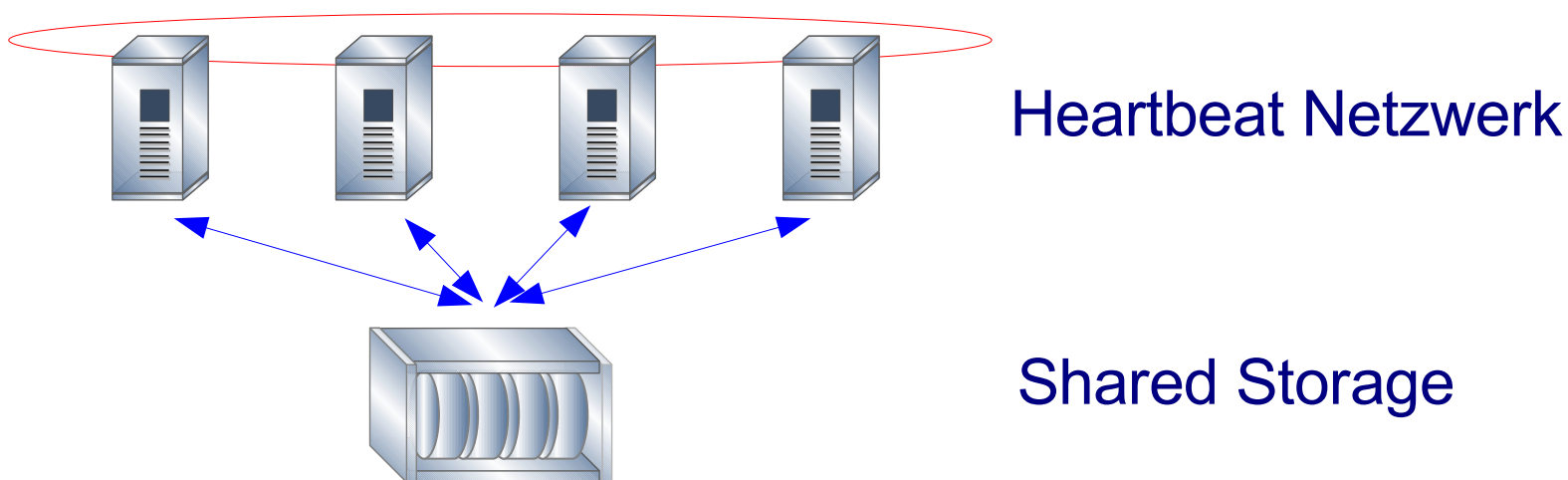


Konzept:

- Auf jeden Knoten laufen unterschiedliche Dienste
- Dadurch ist jeder Knoten aktiv und passiv zugleich
- Im Fehlerfall gibt es aber Performanceeinbussen
- An sonst gilt das gleiche wie für Active/Passive Cluster



HA Cluster N+1



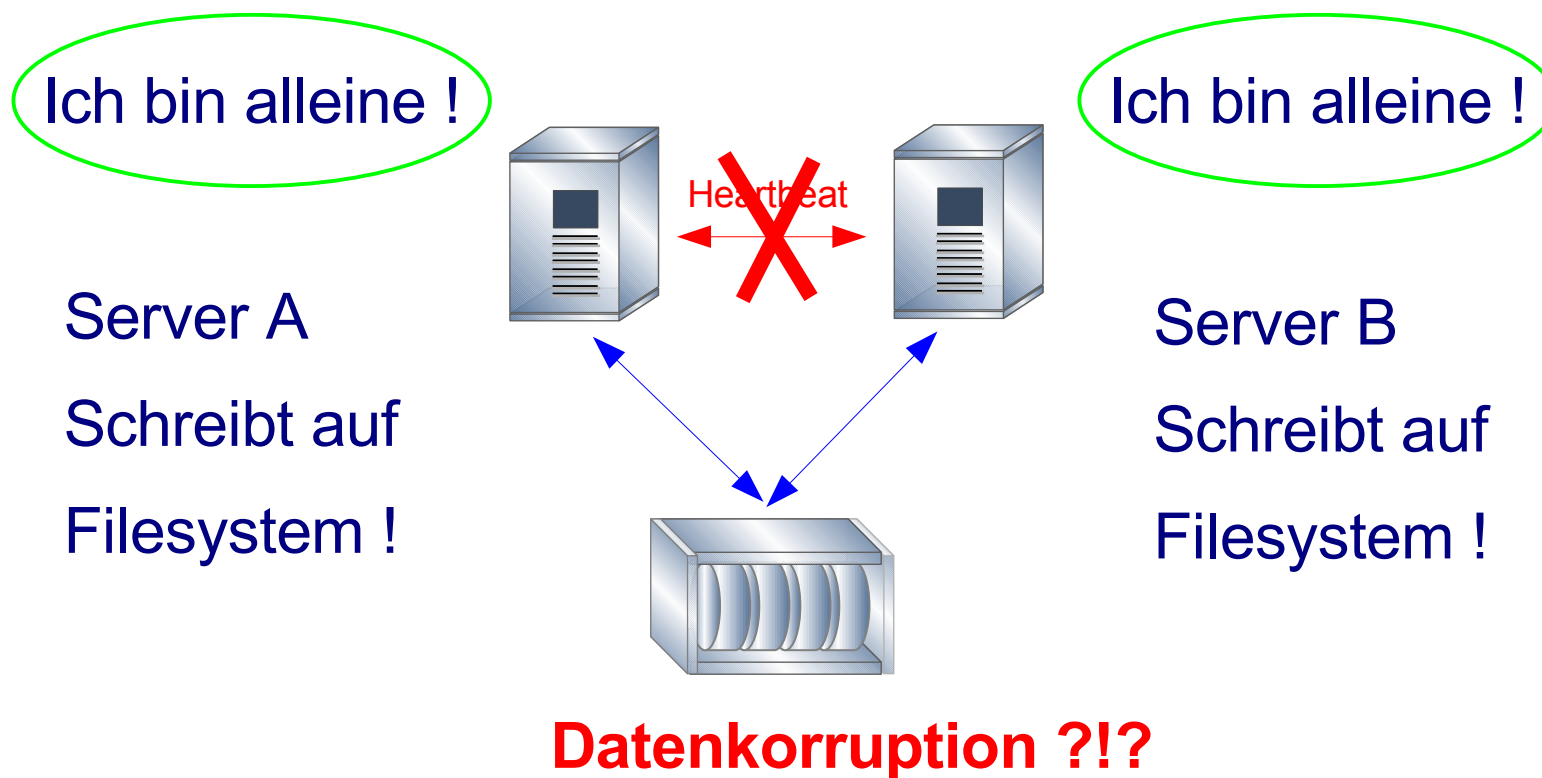
Konzept:

- Es werden um einen mehr Server verwendet als notwendig (N+1)
- Das heisst der Cluster kann den Ausfall eines Knoten ohne Performanceeinbussen hinnehmen
- N+2, N+3 für grössere Ns auch möglich
- Typischerweise unterstützen HA-Software maximal ca. 16 Server



HA Cluster: Split Brain Problem

- Frage: „Was passiert, wenn sich der Cluster sich teilt?“





Stateless und Statefull Services

- Problem des transparenten Failovers
- Theoretisch muss der Dienst in genau dem gleichen Zustand auf dem 2. Knoten wieder laufen
- Dienste, die stateless sind, speichern im Memory keine clientabhängigen Zustände ab
 - => Transparenter Failover ist kein Problem
- Dienste, die Zustände im Memory haben, müssen die Möglichkeit bieten diese persistent zu machen



Stateless und Statefull Services

- Statefull Services:
 - Perfektes Beispiel: DBMS
 - Lösung: Write-Logs, Redo-Logs etc.
 - Aber auch Dienste wie NFS/CIFS können statefull sein
 - Es ist in diesem Fall immer notwendig, dass solche Dienste auch von der HA-Software unterstützt werden
 - Meist geschieht das über spezielle Resource-Types
 - Werden statefull Services ohne entsprechenden Support gefailover, kann es zu Datenverlust kommen

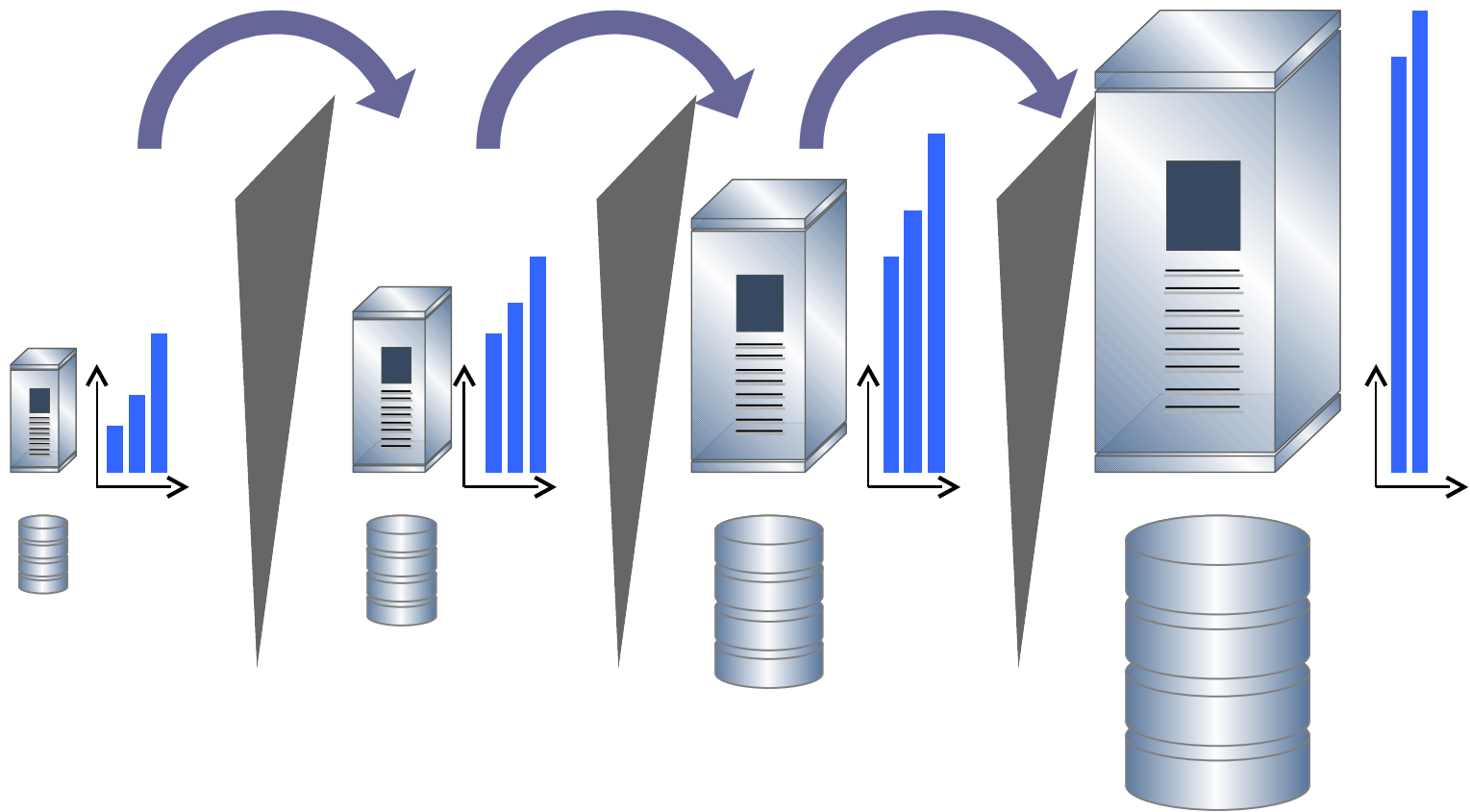


Inhalt

- Über ATIX
- HA-Cluster Grundlagen
- Cluster Filesystem Grundlagen
- Samba und das Clusterfilesystem GFS
- Ausblick

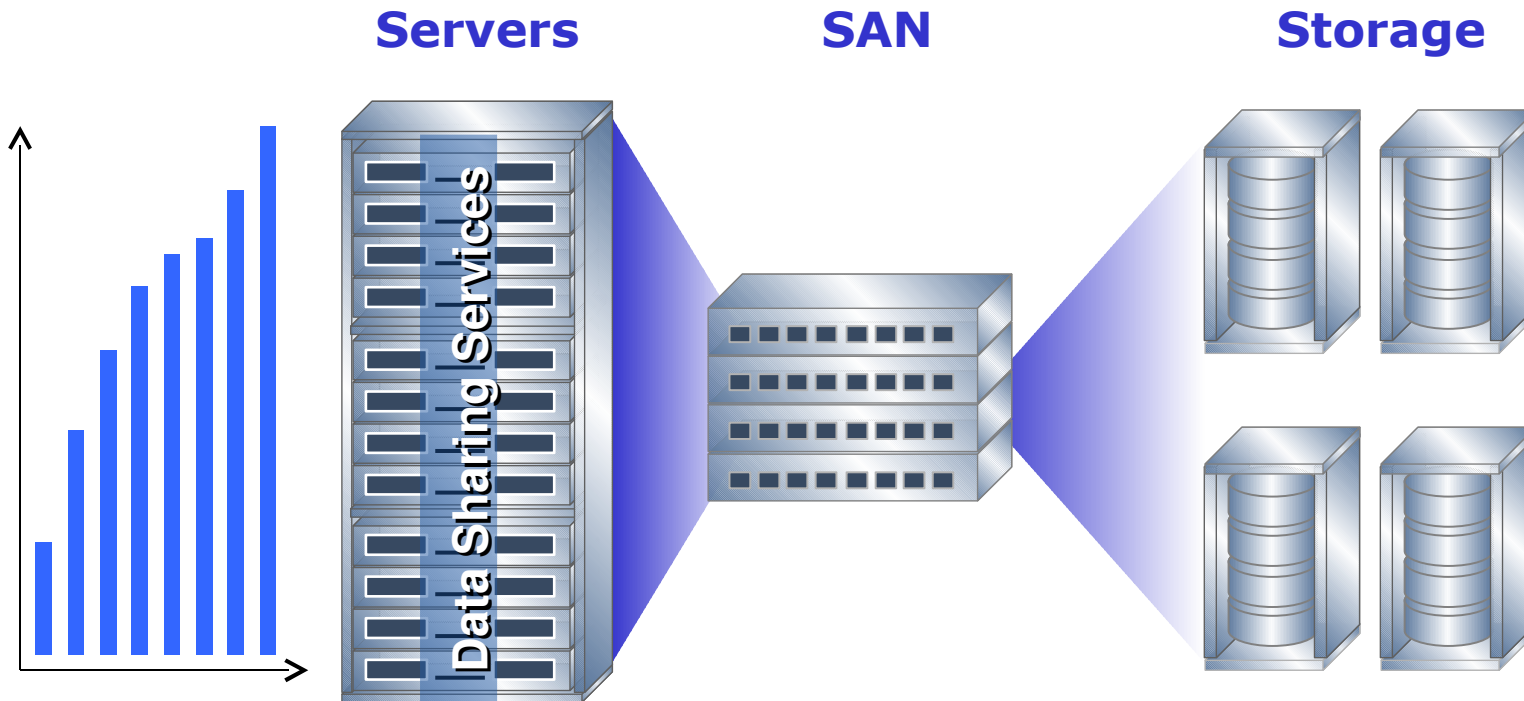


Skalierbarkeit





Skalierbarkeit Storage-Cluster



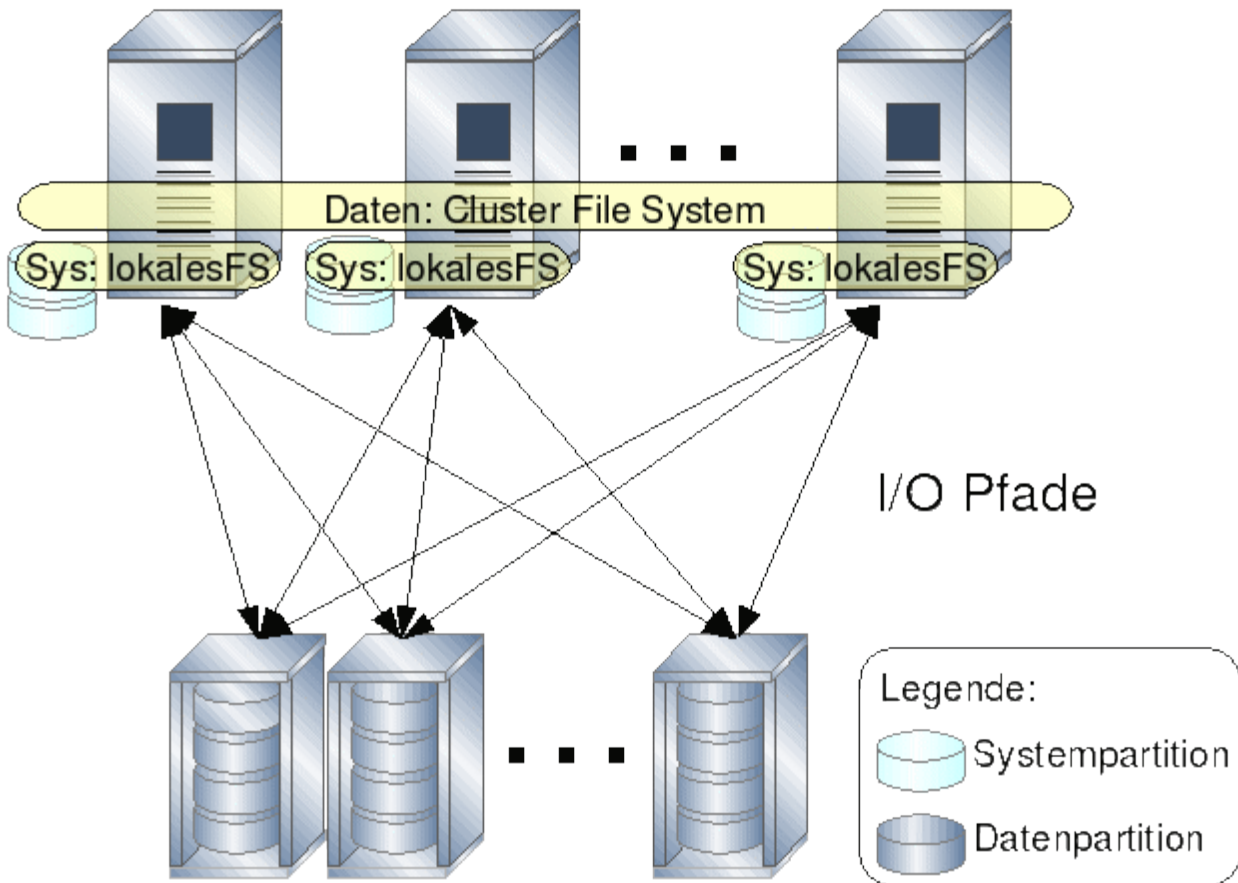
SAN + Linux + Data Sharing = Incremental Computing

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



Shared Storage Cluster

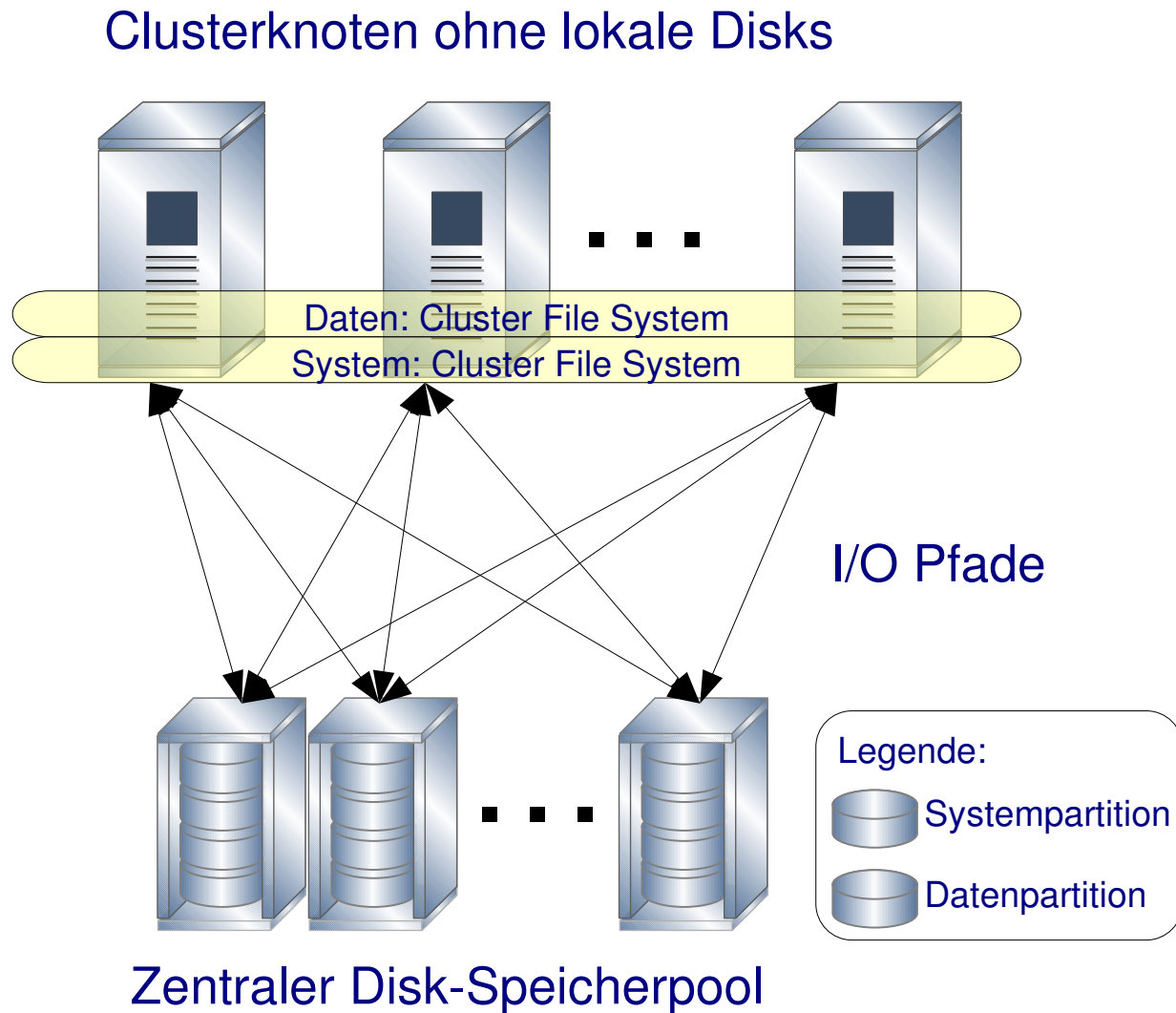
Clusterknoten mit lokalen Disks



Zentraler Disk-Speicherpool

- Data Sharing
- Peer to Peer Kommunikation
- Cluster Filesystem
- Server Cluster
 - Active/Active
- Storage Cluster
 - Storage Pooling
 - Volume Manager
- Hochleistungs Speichernetzwerk
- **Management ??**

Diskless Shared Root Cluster



- Data Sharing
- Shared Root Partition (/)
- Cluster Filesystem
 - Daten
 - System
- SSI auf FS Ebene
 - Management !!
- Skalierbarkeit
- Performance
- Storage Cluster



Global Filesystem (GFS)

- Entwicklung seit 1995
 - University of Minnesota - Sistina - Red Hat
 - Version 6.1
- Symmetrisches Cluster Dateisystem
- Vollwertiges POSIX Dateisystem
- Direct IO (Datenbanken)
- HA Locking Server (GULM)
- Distributed Lock Manager (DLM)
- Online vergrößerbar
- Context Dependent Path Names
- Cluster Volume Manager
- Weiteres: ACLs, Quotas, Multipath



Inhalt

- Über ATIX
- HA-Cluster Grundlagen
- Cluster Filesystem Grundlagen
- Samba und das Clusterfilesystem GFS
- Ausblick



Herausforderungen Samba

Active/Active

- **Servertyp**
 - Domain/ADS Member
 - PDC/ADS Server??
- **Usermapping**
 - Persistent (LDAP, RID-mapping, ..)
- **Filesystem**
 - ACLs
- **Paralleler Zugriff???**
 - Share1->Server1 Share2->Server2



Herausforderungen Samba

Active/Active

- TDB-Dateien
 - GFS vs. Non GFS
- Single Sign-On für Windows und Unix/Linux
 - Windows User (PDS/ADS) -> Server Typ
 - Unix User (passwd, yp/nis, ldap) -> Server Typ
- Virtueller Name und Server Name
- Konfigurationsdateien pro Share vs. globale Konfigurationsdatei
- Winbind Failover vs No Failover



Browsing und Samba Servertyp

- Browsing
 - Domain Master Browser
 - Local Master Browser
 - Der Parameter OS Level
- Servertyp/Security: ADS
 - ADS: als ADS Domain Member
 - User: Server authentifiziert nur lokale Benutzer
 - Server: Erst wird versucht gegen einen anderen NT-Server zu authentifizieren dann lokale User
 - Domain: speziell für NT-Domain Members



Usermapping

- Was wird gemappt?
 - Windows User Ids (SIDs) auf Unix User-Ids (winbind)
 - Statische Tabellen
 - Via LDAP
 - Via RID-Mappings
 - Wichtig für HA-Cluster ist ein persistentes Mapping
 - Unix User auf Windowsuser IDs erfolgt dynamisch



Samba und Cluster Filesysteme

- Posix-ACLs bilden Windows Berechtigungen auf Unix Filesysteme ab
 - Windows Clients unterscheiden zwischen NT-ACLs und Windows 2000 ACLs
(Parameter `acl compatibility = auto|winnt|win2k`)
- Temporäre Samba Dateien (tdbs) müssen im Active/Active Cluster Hostabhängig gemacht werden (CDSL)
 - Teilweise können diese auch im RAMFS gespeichert werden (bessere Performance)



Samba Active/Active

- Was wird unter Samba Active/Active zu verstehen?
 - Unterschiedliche Server können nicht die gleichen Shares exportieren (selbst mit Hilfe eines Cluster Filesystemes)
 - Dafür müsste Samba diesen Mechanismus anbieten und die Clients diesen unterstützen.
 - ABER: mit Hilfe eines CFS können Shares flexibel „im Betrieb“ verschoben werden
 - Trotzdem gelten im Bezug auf Zugriff und Locking die gleichen Regeln wie immer



TDB-Dateien

- TDB Dateien sind Samba-eigene Datenbank Dateien für
 - Locking Informationen (gültig nur während der Laufzeit)
 - User-/Groupmappings (Cache)
 - SIDs und sonstige Sicherheitsrelevante Infos
 - ...
- TDB Dateien sind Prozess abhängig und können via smb.conf in entsprechende Verzeichnisse gelegt werden (lock directory)



Single Sign On

- Gleiche Benutzer (Namen, Passwörter und Identitäten)
 - D.h. Linux muss Windowspasswörter vergleichen können (Kerberos)
 - PAM, Winbind
 - SIDs/RIDs müssen auf Unix UIDs/GIDs gemappt werden können
 - Das kann statisch erfolgen *Administrator=>root*
 - Das kann automatisch erfolgen *Windows User marc=> Unix User marc*



Single Sign On

- Für das Mapping bietet Samba sogenannte IDMap Backends an
 - Standard: TDB (nicht persistent)
 - Neuerdings: `net idmap dump <full_path_and_tdb_filename> > dumpfile.txt`
 - Möglichkeit: LDAP
 - Spezielles Schema mit Unix UserID und Windows SID
 - Möglichkeit: RID Mapping
 - Windows User ID besteht aus SID-Teil und RID-Teil
 - Der RID Teil wird reproduzierbar auf Unix UIDs gemappt



Sonstige Samba Pidfalls

- Servernamen und VIPs
 - Servername wird einer SID zugeordnet und in der Domain als Computer registriert
 - Aber der virtuelle Clustername bzw. die VIP muss und sollte nicht immer dem Servernamen entsprechen (netbios name)
- Konfigurationsdateien
 - Es kann pro Samba Server-VIP Kombination eine Konfigurationsdatei geben (üblicher Ansatz)
 - Oder eine globale Konfigurationsdatei wird beim Failover angepasst (HUP Signal)



Winbind Failover vs. no Failover

- Entweder kann auf jedem Server eine eigene Instanz von Winbind laufen
 - Alle auf dem Server laufenden Samba Prozesse verbinden sich direkt darauf
 - Der übliche Ansatz
- .. oder es wird ein Winbind Prozess an jede Samba-VIP gebunden



Inhalt

- Über ATIX
- HA-Cluster Grundlagen
- Cluster Filesystem Grundlagen
- Samba und das Clusterfilesystem GFS
- Ausblick



Ausblick

- Durch GFS und open-sharedroot lassen sich flexible Cluster bilden die im Betrieb veränderbar sind
- CIFS und NFS sind genauso optimal Anwendungen wie Apache, Mail etc.
- Koppelt man Samba mit HA-Software auf einem sharedroot Cluster entsteht eine ideale Plattform



Noch Fragen?



Vielen Dank für Ihre
Aufmerksamkeit!



Atix GmbH

Einsteinstr. 10

85716 Unterschleißheim

www.atix.de

info@atix.de

ATIX

Chemnitzer
Linux-Tage