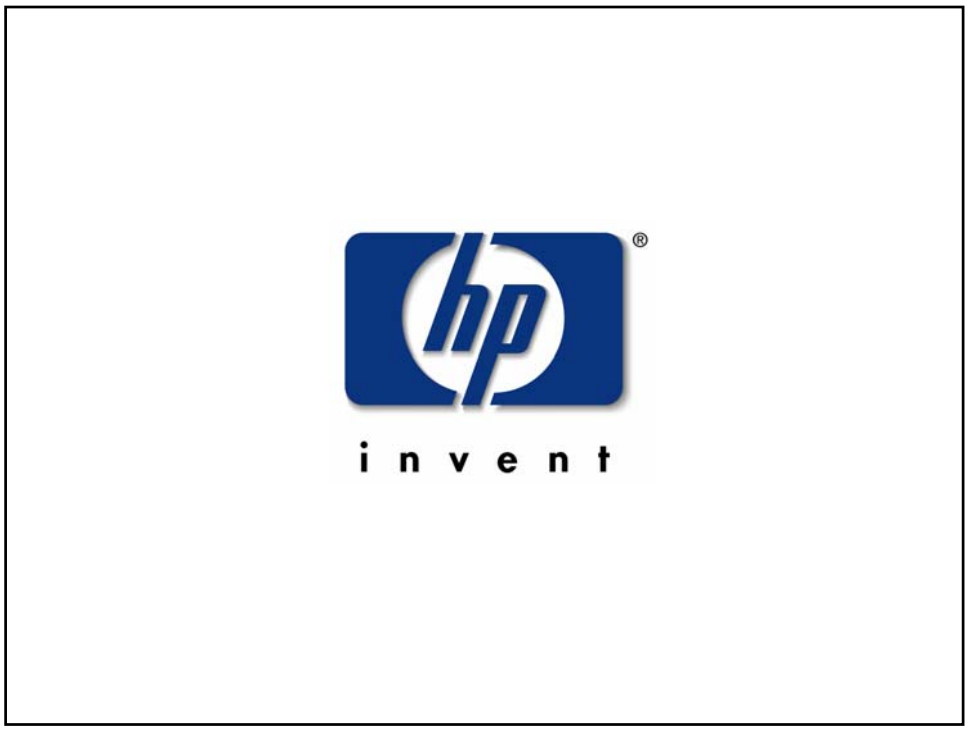

invent


invent
Hochverfügbarkeit
mit Linux
Dr. Sebastian Hausmann
Technical Consultant, HP
© 2004 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice
hp

Agenda:

Warum Hochverfügbarkeit?

HP ServiceGuard für Linux

Beispiele für Bedrohungen

Was kostet Ausfallzeit?

Industry	Business Operations	Average Cost per Hour Downtime
Financial	*Brokerage operations*	*$6.45M*
Financial	*Credit card*	*$2.6M*
Media	*Pay-per-view*	*$150K*
Retail	*Home catalog sales*	*$90K*
Transportation	*Airline reservations*	*$89K*
Media	*Telesales*	*$69K*
Healthcare	*Patient record*	*Loss of life*

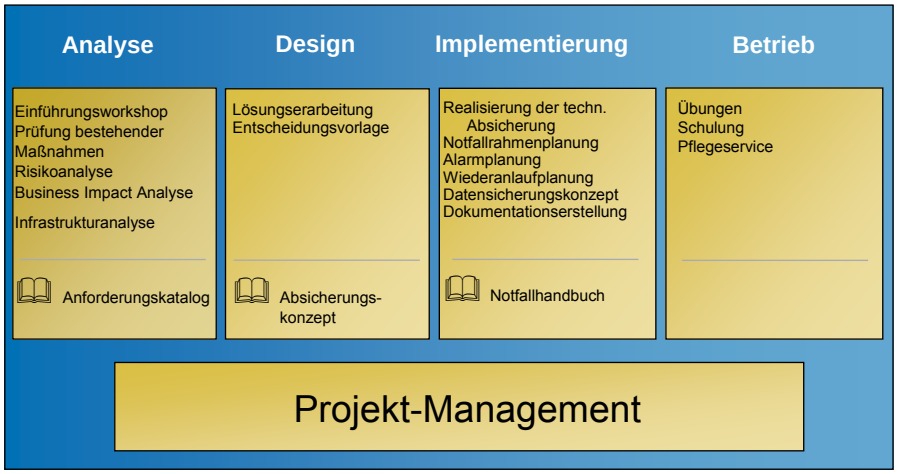
Source: Strategic Research (www.sresearch.com)

Beispiel: Kosten: 1000 € pro Stunde
Ausfallzeit: Ø 12 Std. (15 Min. – 72 Std.) pro Jahr
Quelle: Handelskammer Hamburg
→ **Kosten pro Jahr: Ø 12000 € (250 – 72000 €)**

Weitere Gründe für Hochverfügbarkeit

- Haftung (nach KonTraG 1998, AktG, GmbHG)
  ggfs. persönliche Haftung
- Basel II (ab 2006)
  Differenzierter, ausgefeilter Risikoansatz führt zur Belohnung in Form von niedriger EK-Zuweisung bei Kreditvergabe
- Versicherungsindustrie (Prämien & Versicherbarkeit)
- Kundenanforderungen (z.B. Ausschreibungen)

HP Business Continuity Prozessmodell

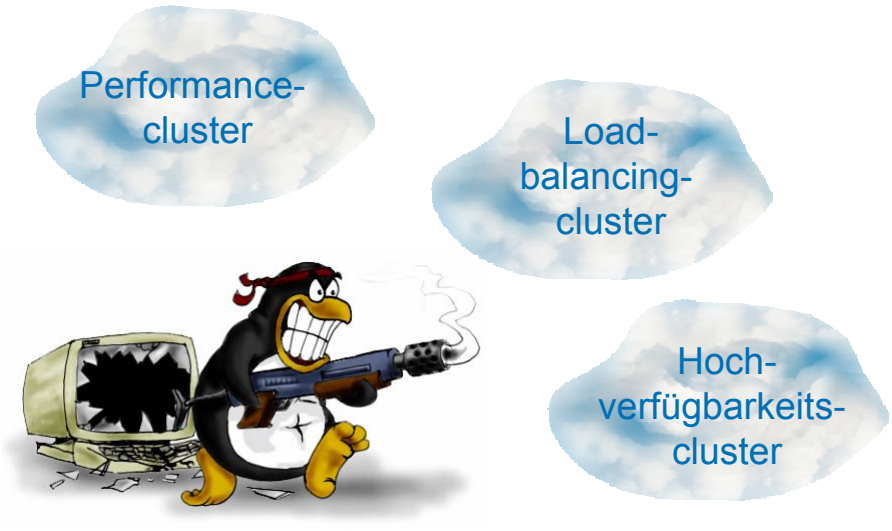


Klassifizierung der Verfügbarkeit

Availability Level	Mögliche Ausfälle	Beispiele
AL4: Fault Tolerant	Keine Unterbrechung bei Fehlern und Reparaturen	Assured Availability HP NSK
AL3: Fault Resilient	Verlust von Transaktionen möglich	Oracle Fail Safe
AL2: High Availability	Re-Logon, Neustart, Performanceverlust	HP ServiceGuard MS Cluster
AL1: Data Availability	Shutdown, System nicht verfügbar, Daten	RAID, Datenspiegelung
AL0: Conventional	Server stoppt, Datenverlust, Dateien zerstört	Jeder Server
AL+: Desastertoleranz	Wiederherstellung nach Ausfall eines kompletten Rechenzentrums	Cluster-Extension

Source: Harvard Research Group

Linux und Cluster



Aberdeen White Paper, January 2001

today (2001):

SteelEye	Lifekeeper
Mission Critical Linux	★ Convolo
Legato	Legato Cluster

the next 12 months:

PolyServer	★ Understudy & LocalCluster
SGI	FailSafe
Turbo Linux	★ Cluster Server 6
Veritas	Cluster Server for Linux
Hewlett-Packard	**MC/ServiceGuard = Serviceguard**
Apptime	Watchdog
Red Hat	★ High Availability Server

★ open source

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

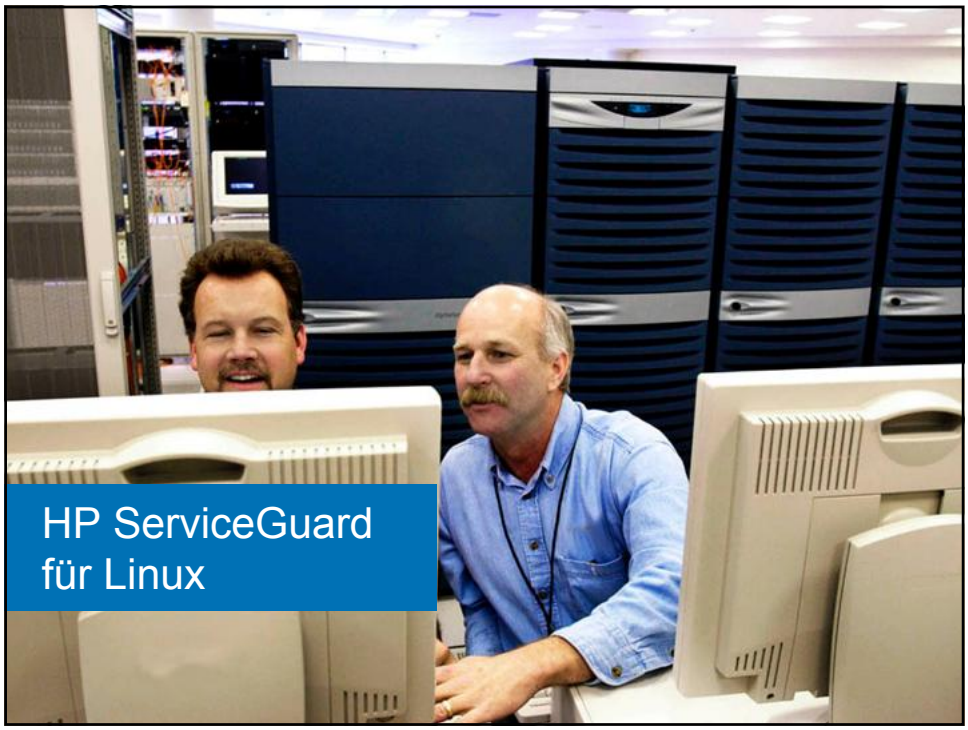
... what's going on outside ?

HighAvailability

High Availability Linux Project

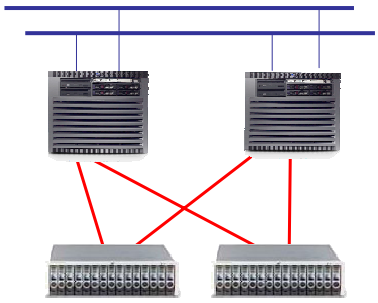
see http://linux-ha.org/

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005



HP ServiceGuard

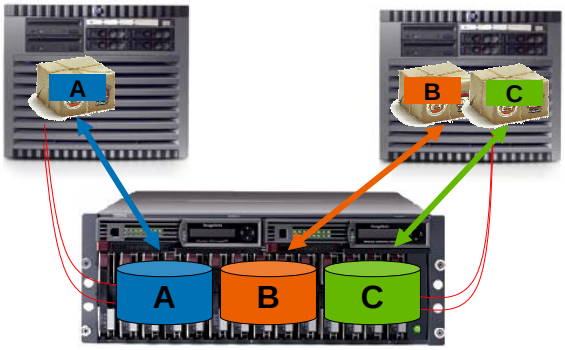
- Nicht fehlertolerant, kein Loadbalancing
- Überwacht Hardware und Software
- Erfordert redundante Hardware (no SPOF)
- Läuft auf Standard-Hardware und -Betriebssystem
- Integration von Applikationen ohne Modifikationen
- Mehr als 80.000 Lizenzen verkauft (HP-UX & Linux)
- Unterstützt lokale, Campus-, Metropolitan- und kontinentale Konfigurationen



Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

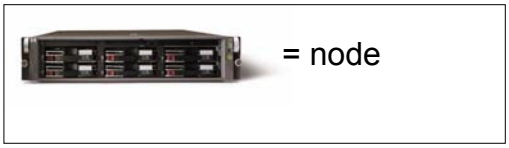
„Shared Storage“

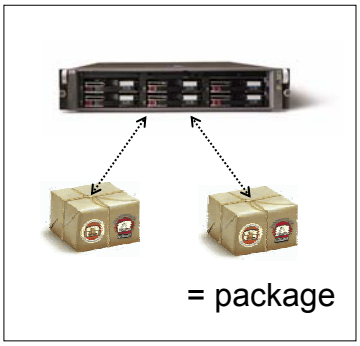
Die **Knoten** eines **Clusters** “sharen” gemeinsame Plattenlaufwerke über Multi-Initiator-SCSI-Busse oder über FC-basierende SANs.
(Eigentlich „Shared-Nothing“-Prinzip)

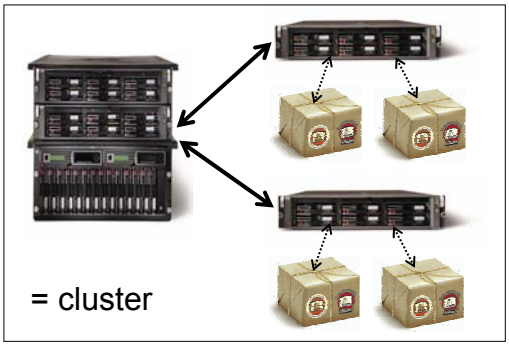


Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Begriffe

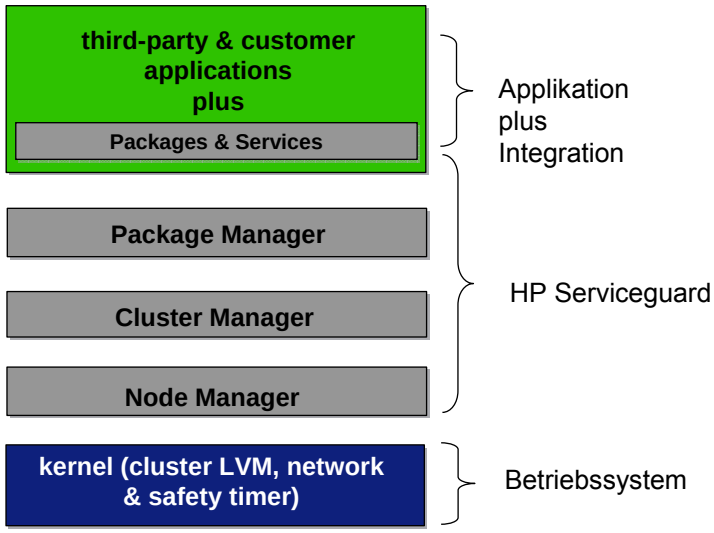






Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

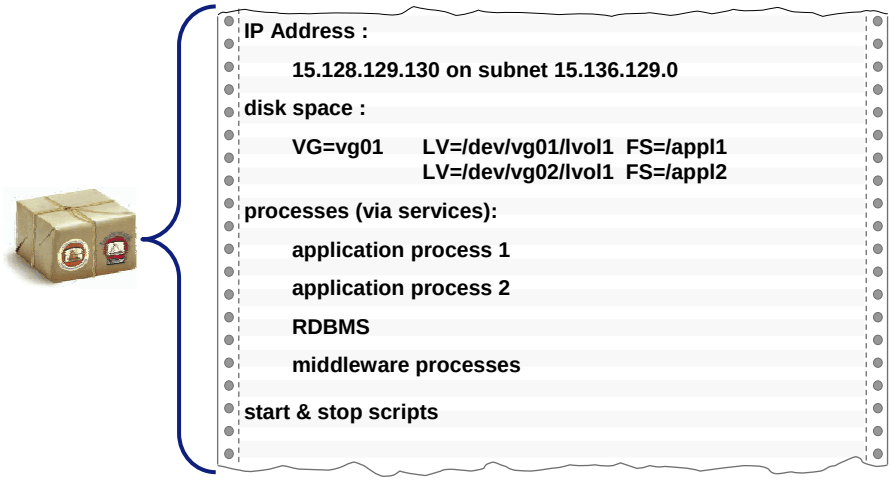
Layer-Architektur



Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Paket

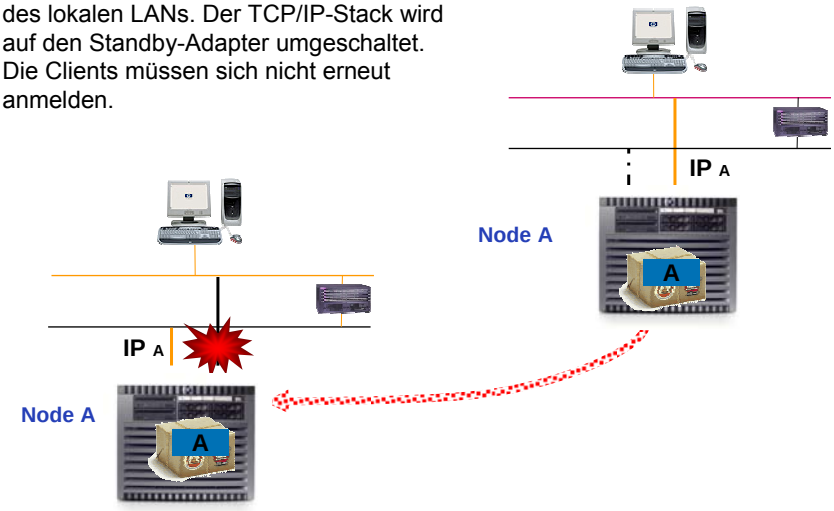
Ein Paket definiert alle Resourcen, die zum Betrieb einer Applikation auf einem Cluster-Knoten notwendig sind.



Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Lokaler LAN-Ausfall

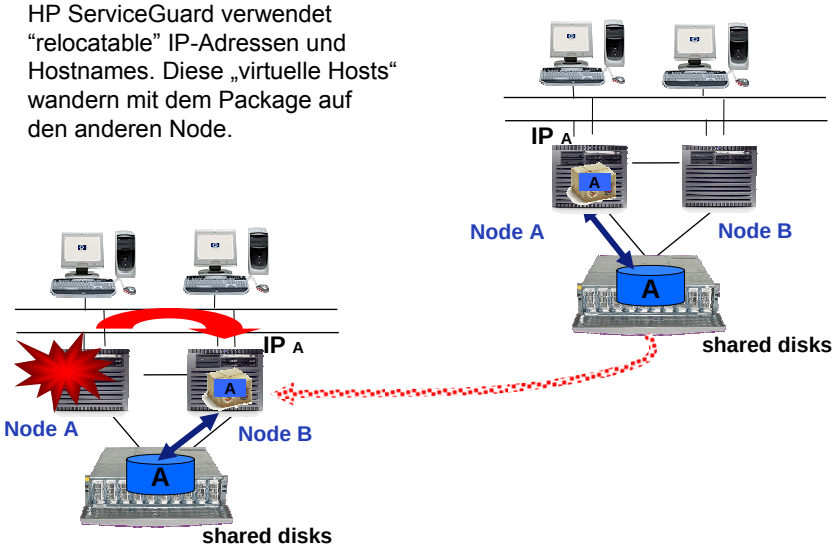
Transparentes und schnelles Umschalten des lokalen LANs. Der TCP/IP-Stack wird auf den Standby-Adapter umgeschaltet. Die Clients müssen sich nicht erneut anmelden.



Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Paket-Umschaltung

HP ServiceGuard verwendet “relocatable” IP-Adressen und Hostnames. Diese „virtuelle Hosts“ wandern mit dem Package auf den anderen Node.

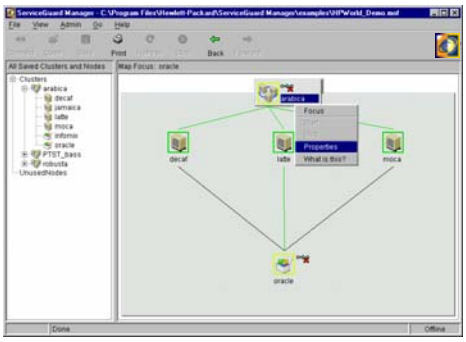


Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Management

Serviceguard-Manager:

- Java GUI
- uses cluster object manager (no SNMP)
- supports many platforms
- monitoring and **controlling**



command line interface:

```
cmapplyconf, cmcheckconf, cmdeleteconf, cmgetconf,
cmhaltcl, cmhaltnode, cmhaltpkg, cmhaltserv,
cmmakepkg, cmmodnet, cmmodpkg, cmquerycl, cmreadlog,
cmreadlog, cmruncl, cmrunnode, cmrunpkg, cmrunserv,
cmscancl, cmviewcl, cmviewconf
```

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

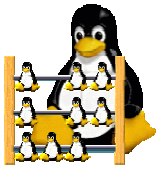
Designziele

- ein Sourcecode für alle HP SG-Varianten
- langfristig Funktionalität wie HP-UX-Version
- Standard HP-Softwaresupport
- Keine Kernel-Anpassungen
  oder
  Anpassungen unter GPL
- Nutzung vorhandener Open-Source-Software:
  - Linux LVM
  - Linux Software RAID
  - ReiserFS
  - ...

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Eigenschaften der Version A.11.15

- IA-32- und IA-64-Architektur
- 2 SCSI-Nodes oder bis zu 16 FC-Nodes
- Active / Active
- maximal 150 Packages im Cluster
- maximal 900 Services je Cluster
- maximal 200 virtuelle IP-Adressen
- maximal 7 Heartbeat LANs
- “shared” FC/SCSI Plattenlaufwerke



- SUSE Linux Enterprise 8 (UL 1.0)
- Red Hat Enterprise Linux AS 3

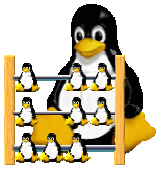
neu

- CLI und SGMgr Unterstützung
- Online Rekonfiguration von Nodes, Packages, VGs
- Quorum Server Support für 100 Nodes / 50 Cluster

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Eigenschaften der Version A.11.16

- IA-32- und IA-64-Architektur
- 2 SCSI-Nodes oder bis zu 16 FC-Nodes
- Active / Active
- maximal 150 Packages im Cluster
- maximal 900 Services je Cluster
- maximal 200 virtuelle IP-Adressen
- maximal 7 Heartbeat LANs
- “shared” FC/SCSI Plattenlaufwerke



- SUSE Linux Enterprise 9
- Red Hat Enterprise Linux AS 3

Kernelversion beachten!

- Kontrolle und Überwachung über GUI

neu

- Non-Root-Access
- CLI und SGMgr Unterstützung
- Online Rekonfiguration von Nodes, Packages, VGs
- Quorum Server Support für 100 Nodes / 50 Cluster

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Unterstützte Distributionen

A.11.14:

- RedHat AS 2.1 with kernel 2.4.9-e3/e25/e27 (SCSI & FC)
- Redhat 7.3 with kernel 2.4.18 (SCSI only)
- SuSE ES8 UL 1.0 with kernel 2.4.19 (SCSI & FC)
- Linux LVM for Kernel (patches supplied for RedHat)
- ext2, ext3 Filesystem and ReiserFS

A.11.15:

- RedHat EL AS 3.0 (SCSI & FC) April 2004
- SuSE ES8 UL 1.0 with kernel 2.4.19 of SP2a (SCSI & FC)
- SuSE ES8 UL 1.0 with kernel 2.4.21 of SP3 (IPF & FC only)
- Linux LVM for kernel
- ext2, ext3 Filesystem and ReiserFS

Unterstützte Distributionen

A.11.16:

- RedHat EL 3 AS/ES 2.4.21 (U3) -20.EL
- RedHat EL 3 AS/ES 2.4.21 (U4)-27.EL
- SUSE SLES 9 2.6.5 (SP1)-7.139
- SUSE SLES 9 2.6.5 (SP1)-7.145
- Linux LVM for kernel
- ext2, ext3 Filesystem and ReiserFS

Unterstützte Hardware

Server:

HP ProLiant
- DL360 G3/G4
- DL380 G2-G4 (auch Packaged-Cluster)
- DL560, DL580 G2, DL585
- DL740, DL760 G2
- ML350 G3/G4, ML 370 G3/G4
- BL20p G2, BL25p, BL40P

HP Integrity
- rx1600, rx2600
- rx4640, rx5670
- rx7620, rx8620, Superdome

Storage:
- XP48, XP128, XP512, XP1024, XP12000
- EVA3000/5000
- VA7xx0
- MSA1000, MSA1500cs
- MSA500 G2 (nur ProLiants)

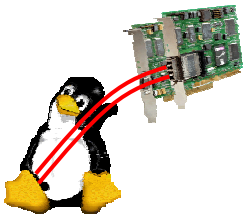
HBAs:
- supported Proliant SCSI HBAs
- A6826A für IPF
- FCA2214(DC) für IA32

Config-Guide beachten!
ftp://ftp.compaq.com/pub/solutions/enterprise/ha/linux/svcguard-certmatrix.pdf

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

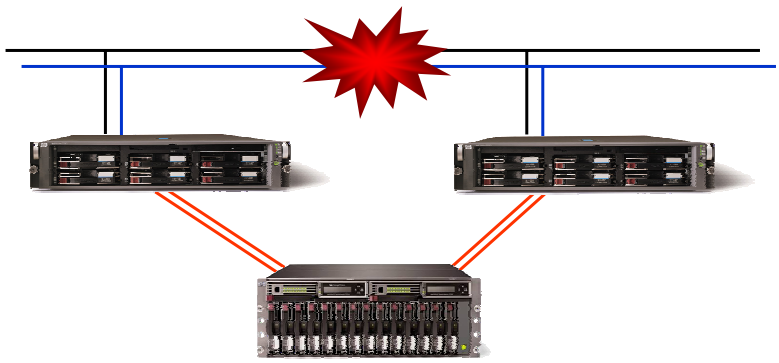
multipathing support

- hp Autopath for Linux for kernel 2.4.2-2 only (A.11.13)
- raidtools multipathing is supported for
  - VA
  - XP
- support for SecurePath 3.0A with A.11.14.02
  - EVA3000 (2 nodes only)
  - EVA5000 (2 nodes only)
  - MSA (2 nodes only)
- new Qlogic multipath driver for
  - XP disk arrays
  - VA disk arrays



Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

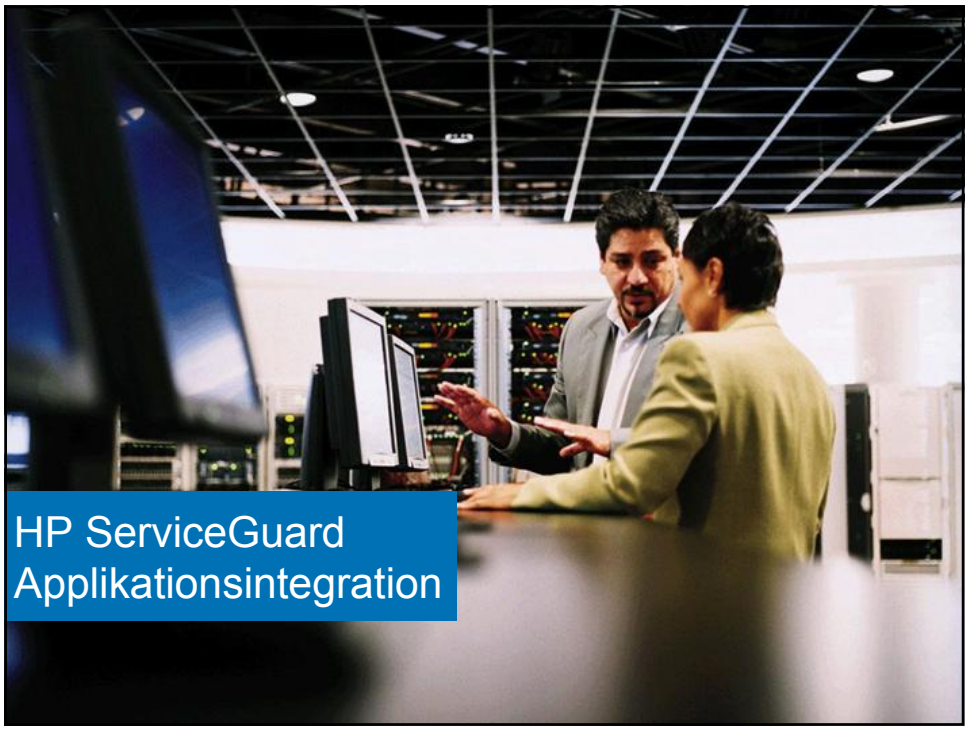
„Split Brain“-Situation



Die Clusterservices können nicht mehr miteinander kommunizieren
→ beide Knoten versuchen, alle Cluster-Pakete zu übernehmen

Split-Brain → Quorum-Server

- Beim Start und im Fehlerfall regelt der QS die Initialisierung des Clusters.
- Der QS läuft nicht auf einem Cluster-Knoten.
- QS benutzt TCP/IP und wartet mit Port 1238 auf Anfragen von ServiceGuard-Knoten.
- Nur eine IP-Adresse für den QS möglich.
- Nur ein QS pro Cluster möglich.
- Auf QS Maschinen können andere Applikationen laufen.
- QS kann als ServiceGuard-Paket auf einem anderen Cluster laufen.
- Ein QS kann für mehrere Cluster genutzt werden:
  Maximal 50 Cluster und maximal 100 Knoten.
- Der QS wird für die Cluster-Konfiguration benötigt (cmapplyconf).
- Der QS wird für die Cluster-Neubildung benötigt, falls verbleibende Knoten keinen Quorum bilden können (≤ 50% der Knoten).
- Also ist ein QS notwendig für einen 2-Knoten-Cluster und optional für 3 bis 16 Knoten.
- Ab Version **A.11.15** gibt es auch eine „Cluster-Lock-LUN“ für Cluster mit **2** bis **4** Knoten.
- Die Cluster-Lock-LUN ist eine „shared LUN“ mit einer festen Partitionsgröße von 100kB.
- **kein SPOF**



Applikations-Integrationen

- drei Arten von Applikations-Integrationen:
  - → offizielle “Produkte”
  - → getestete Script-Lösungen
  - → Whitepapers

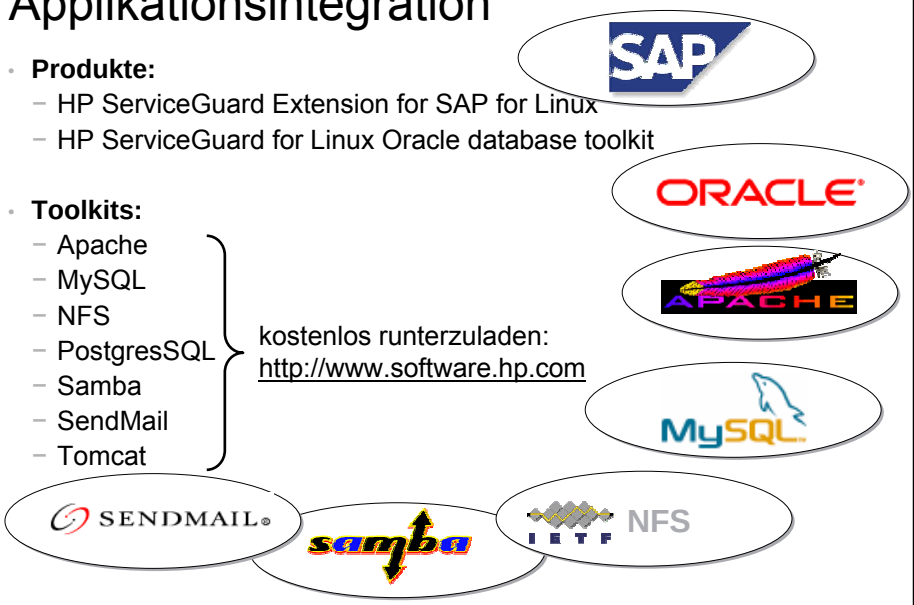
- Voraussetzungen (Auszug):
  - → Reboot/Powerfail-Resistenz
  - → automatische Start-/Stop-Prozeduren
  - → keine Abhängigkeiten zu CPU-IDs / MAC-Adressen
  - → NFS-Locks vermeiden
  - → kein Binding an den Host-Namen
  - → feste TCP-Ports und DNS verwenden
  - → möglichst keine Daten im Root-Bereich
  - → keine lokale Peripherie verwenden
  - → Clients brauchen ein Reconnect-Verfahren

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Applikationsintegration

- **Produkte:**
  - HP ServiceGuard Extension for SAP for Linux
  - HP ServiceGuard for Linux Oracle database toolkit
- **Toolkits:**
  - Apache
  - MySQL
  - NFS
  - PostgresSQL
  - Samba
  - SendMail
  - Tomcat

kostenlos runterzuladen:
http://www.software.hp.com

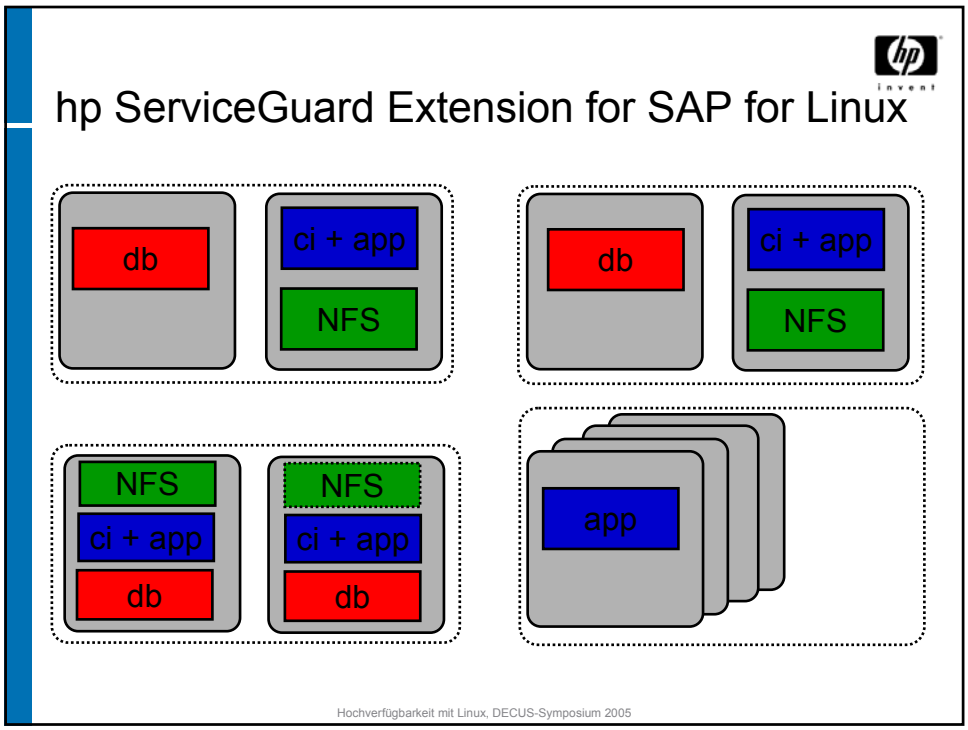


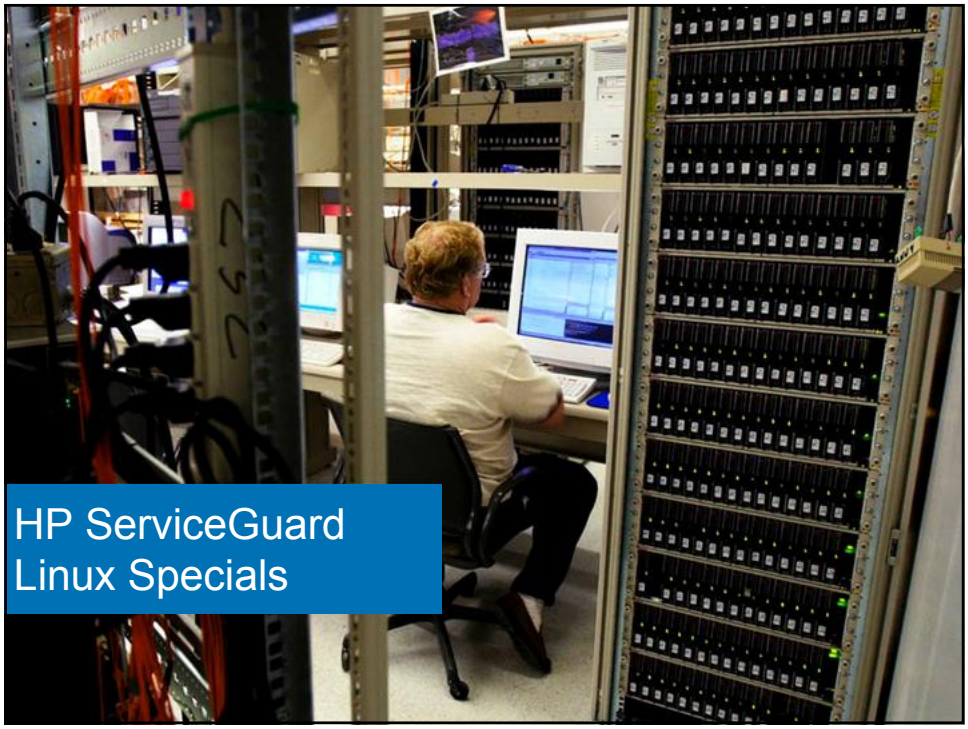
Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

hp ServiceGuard Extension for SAP for Linux

- **what is it, what does it do ?**
  - based on scripts and configuration files
  - smoothly integrated into SG package control script architecture
  - protects database (**db**) and central instance (**ci**)
  - implements NFS services via single HA NFS package
  - supports single and dual package configs for **db** and **ci**
  - uses one relocatable ip address for **ci** and one for **db**
  - supports multiple SAP instances per cluster
  - simple interface to SAP instance reconfigure during cluster reforming
- **what is not provided ?**
  - any type of monitoring (except HA NFS)
  - application server clustering
  - load balancing
  - Cluster Consistency Monitor (currently HP-UX only)

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

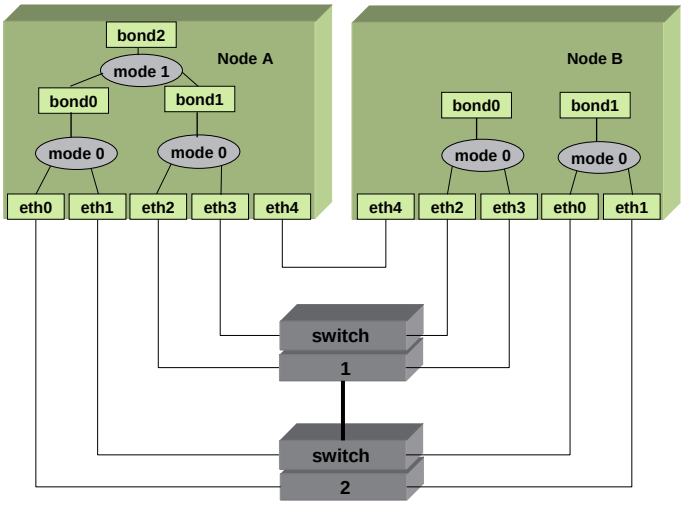

hp ServiceGuard Extension for SAP for Linux
db
ci + app
NFS
db
ci + app
NFS
NFS
ci + app
db
NFS
ci + app
db
app
Hochverfügbarkeit mit Linux, DECUS-Symposium 2005


HP ServiceGuard
Linux Specials

bonding driver

- the bonding driver originally came from Donald Becker's beowulf patches
- its basic purpose is the bundling of network interfaces like HP's Autoport Aggregation
- bonding is part of the official kernel distribution since kernel 2.2.18.
- it's recommended to load it as a module in order to be able to pass parameters to the driver
- supports HIGH AVAILABILITY (=1) and LOAD BALANCING (=0) mode
- supports any type of ethernet interface, a bond can use different cards at different speed
- HIGH AVAILABILITY requires **`MII`** state reporting supported by the card and the driver
- LOAD BALANCING requires switches with trunking capability
- more than 2 physical slave cards can be combined in a bond

„Bonding Driver“



multipathing – RAID-Tools

- „undocumented“ RAID personality of RAID-Tools
- included in current RAID-Tools rpms
- supports failover mode only
- supports SCSI, FC, IDE/ATA, ...
- up to **256** virtual devices (**/dev/md0**...**md255**)
- a maximum of **64** redundant LUN pathes
- no automatic discovery of redundant pathes
- no persistent bindings between virtual and physical devices, no deterministic preferred path
- no automatic failback of pathes
- path managment via RAID disk add/remove commands

multipathing – hp Securepath

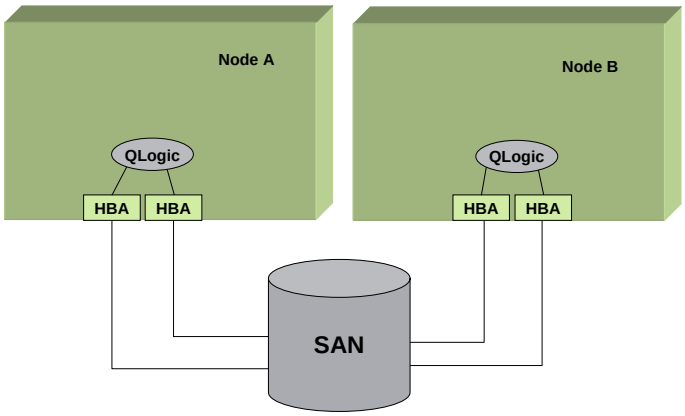
- required for **MSA1000** and **EVA5000/3000** arrays
- supported HBAs:
  - **FCA2214**, **FCA2214DC**
  - **A6826A** (for integrity servers)
- supports up to **8** paths per LUN
- typically max **32** LUNs in EVA configurations
- supports load balancing and failover mode
- supports persistence of physical and virtual devices
- requires a reboot for adding LUNs (not for extending LUNs)
- 2 node support for Serviceguard only
- only CLI interface, no GUI support like for Windows
- for details look at: http://www.hp.com/go/securepath

multipathing – Qlogic driver

- built-in multipathing features for disk arrays following simple FC configuration rules (like VA7xxx)
- added XP specific redundant path discovery with driver version **6.06.50** → still open source
- supports failover and „static“ loadbalancing mode
- requires SANSurfer software for enhanced management (preferred pathes, load balancing, persistent binding)
- adding and removing LUNs can be done „online“
- supports **128** LUNs
- will support all HP online storage platforms in the future
- driver is XP-aware and still open source

Multipath im SAN

- QLogic-Treiber ab Version 7.00.03
- kein Securepath mehr



LVM implementation

- developed by Heinz Mauelshagen during an IBM project at German Telekom in Darmstadt
- implementation is more than 90% compatible with HP-UX plus some enhancements and some IBM LVM features

- the CLI is a superset of HP-UX LVM with similar option parameters
- Linux LVM uses the same directory and naming scheme as HP-UX

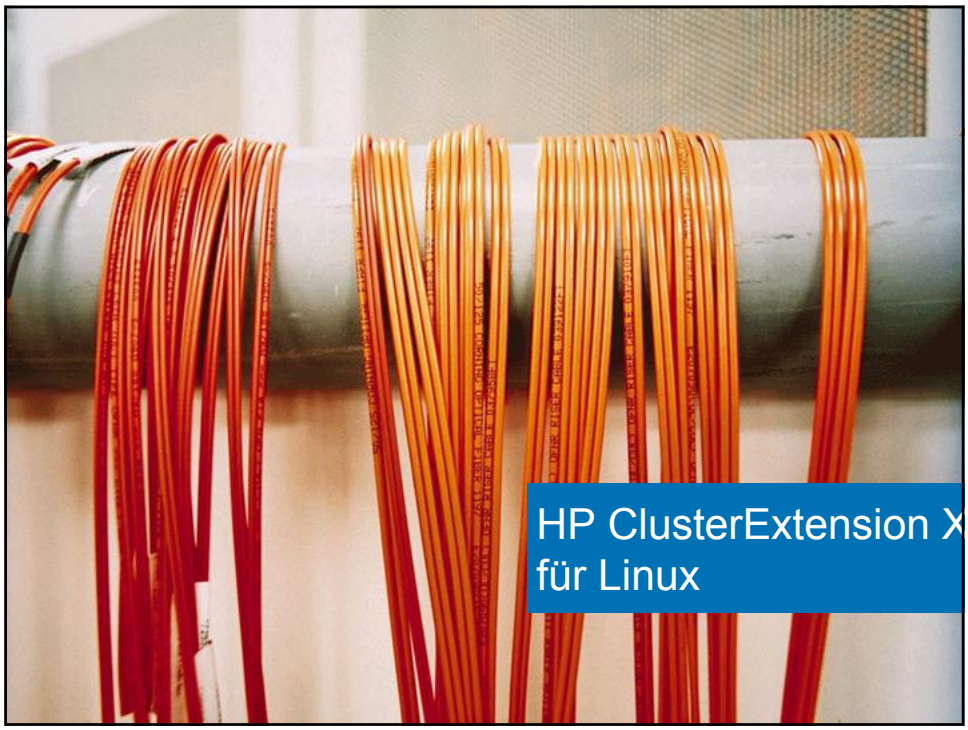
- since kernel version 2.4 official part of the kernel distribution (SUSE)
- the latest version is 1.0.9 (still developed by Heinz Mauelshagen)
- the LVM project found a technical home at http://www.sistina.com

LVM implementation (continued)

- needs to be compiled into the kernel or loaded as module
- kernel is able to boot from LVM based root disks by using a specific RAM disk layout for the boot process
- physical volumes can be of any type of disk partition
- LVM can be combined with Linux software RAID
- Linux LVM uses the **/proc** filesystem as interface for the CLI

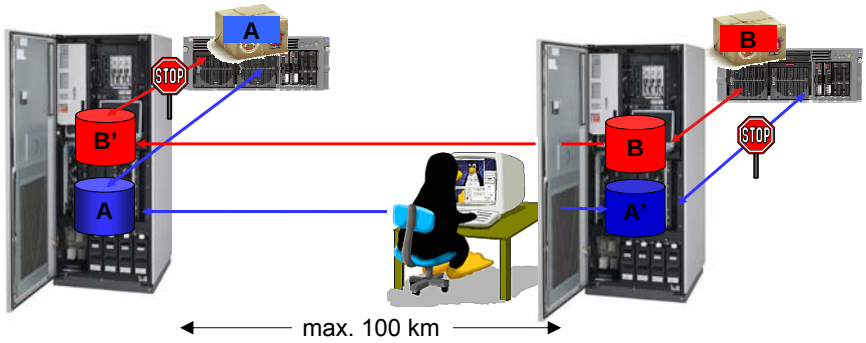
Limitations:

- no PV links
- no LVM based mirroring
- no cluster aware volume groups



Continuous Access XP

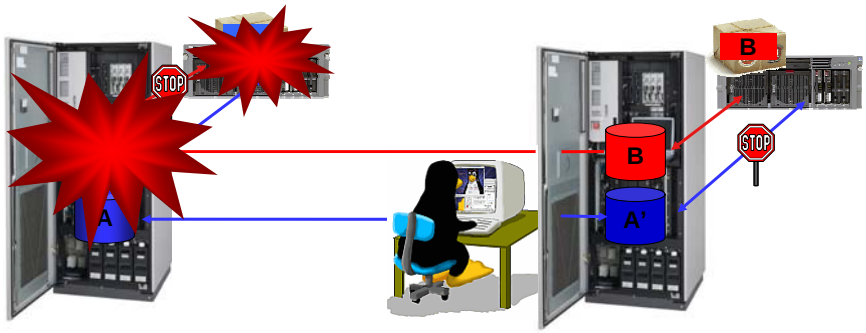
- Unterstützt synchronen und asynchronen CA-Betrieb
- Aktiv-Aktiv-Konfiguration
- Datenspiegelung durch Festplattensystem
  → transparent für Cluster-Package-Manager



Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Continuous Access XP

- Selbst nach Zerstörung eines Rechenzentrums laufen die Applikationen weiter.



Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Die Vorteile von HP Cluster Extension

- im Prinzip keine Distanz-Limitationen
- asynchroner Modus für lange Distanzen (keine Signal Latency)
- Load Balancing zwischen den XP Systemen
- bessere Performance als OS Mirroring beim Schreiben
- die Resynchronisation erfolgt immer auf Basis von Deltas (Tracks)
- implementiert ein Quorum für die Verfügbarkeit der Daten
- der Spiegel ist immer “schreibgeschützt”
- man kann einen “schreibbaren” Snapshot der Produktionsdaten erzeugen (z.B. für Wartungsarbeiten)
- eine Technologie für AIX, HP-UX, Windows2000 und Linux-Cluster

Hochverfügbarkeit mit Linux, DECUS-Symposium 2005

Award at LinuxWorld 2005

HP's Virtual Server Environment for Linux was named Best Clustering Solution in the LinuxWorld Products Excellence Awards program.

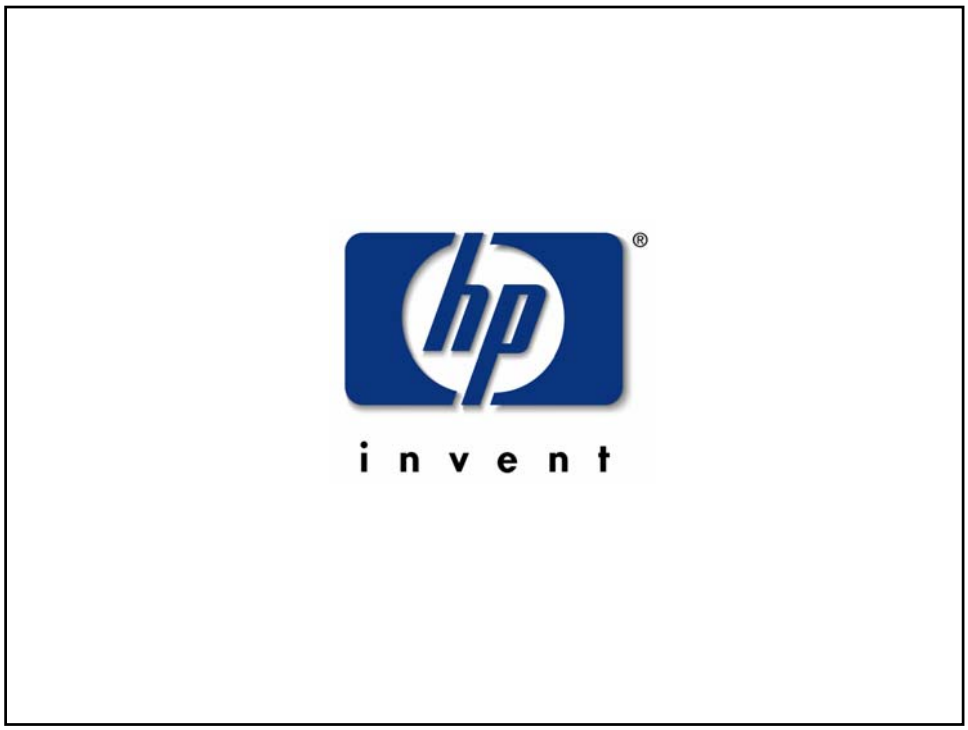
The awards, which recognize important innovations in Linux and Open Source technologies, were given out Feb. 16 during the LinuxWorld Conference and Expo in Boston.

HP VSE is an integrated server offering that provides a flexible computing environment. VSE is part of the HP virtualization portfolio. HP virtualization solutions let business pool and share IT resources so utilization is optimized and supply automatically meets demand.

HP released the first version of VSE for Linux in February with the availability of Global Workload Manager and HP Serviceguard for Linux clustering on the 2.6 kernel. HP gWLM provides the policy engine to allocate virtual server resources in a Linux operating system. HP Serviceguard for Linux is the high availability clustering component of the solution. It maximizes service uptime and data integrity and minimizes planned downtime, providing advanced mission-critical capabilities.

Ausblick

- Unterstützung von Redhat 4
- Unterstützung von WBEM
- Cluster-Extension für EVA
- Weitere Toolkits


invent