



Herzlich Willkommen!
...zum Vortrag:
Überblick über
Clustertechnologien



 **HP User Society**
DECUS München e.V. 



Ein Baustein für
“Adaptive
Enterprise”
Business
continuity:

Clustertechnologien
im Vergleich

Dr. Christoph Balbach
Manager Presales Nord-Ost


 **HP User Society**
Vortrag 1K06  DECUS München e.V.

 **Ein Baustein für
“Adaptive Enterprise”
Business
continuity:**


**Hochverfügbarkeit bei
Servern und Storage**



© 2004 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice

 **Agenda**

- Warum business continuity Lösungen?
- Technische Details:
 - Single/Multi System Image
 - Shared Root
 - System Management
 - Cluster Alias
 - Cluster File System
 - Direct and Direct Access I/O
 - Device Naming
 - Distributed Lock Manager
 - Konfigurationsfragen
 - Interconnect
 - Quorum
 - Applikation Support
 - Special coding for clusters?
 - Failover scripting
 - Auswirkungen im Fehlerfall
 - Data Replication
 - Disaster Tolerance





Sind Sie hinreichend abgesichert ?

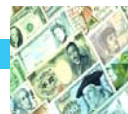


Geschäftsprozesse

- Welche Kosten entstehen, wenn Ihr Unternehmen eine Stunde, ein Tag nicht arbeiten kann?
- Welche Ressourcen sind notwendig, damit Ihr Unternehmen im Notfall überleben kann?

Kunden & Partner

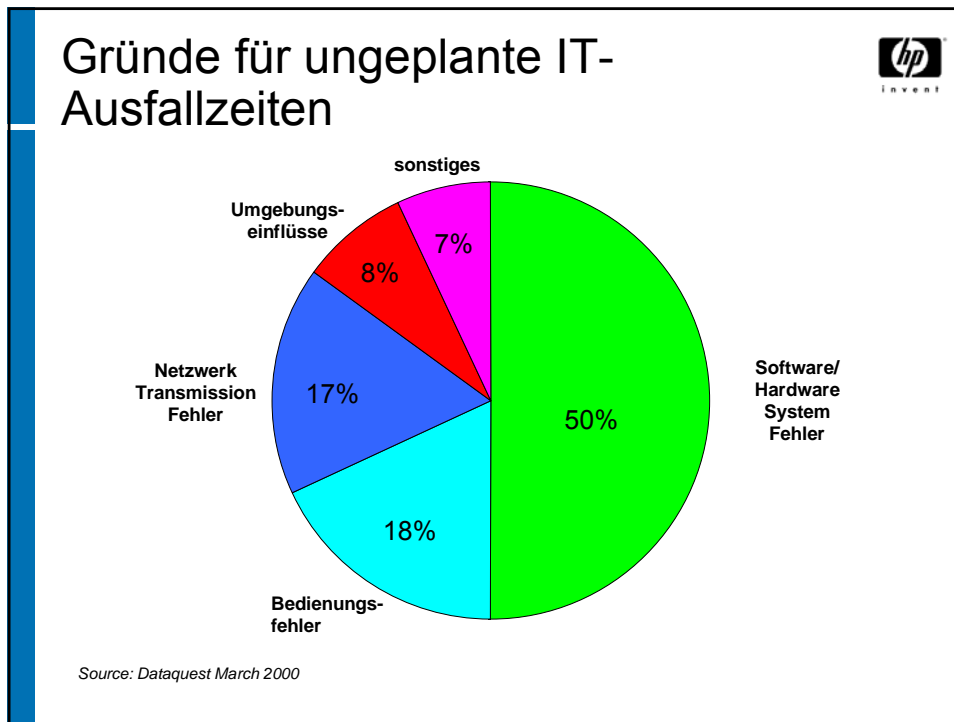
- Wie werden Ihre Kunden reagieren, wenn Ihr Geschäftsbetrieb unterbrochen ist ?
- Wie werden sich Ihre Partner verhalten, wenn Ihr Geschäftsbetrieb unterbrochen ist?



IT-Umgebung

- Haben Sie einen Notfallplan/Wiederanlaufkonzept ?
- Haben Sie jemals eine Wiederanlaufübung gemacht?



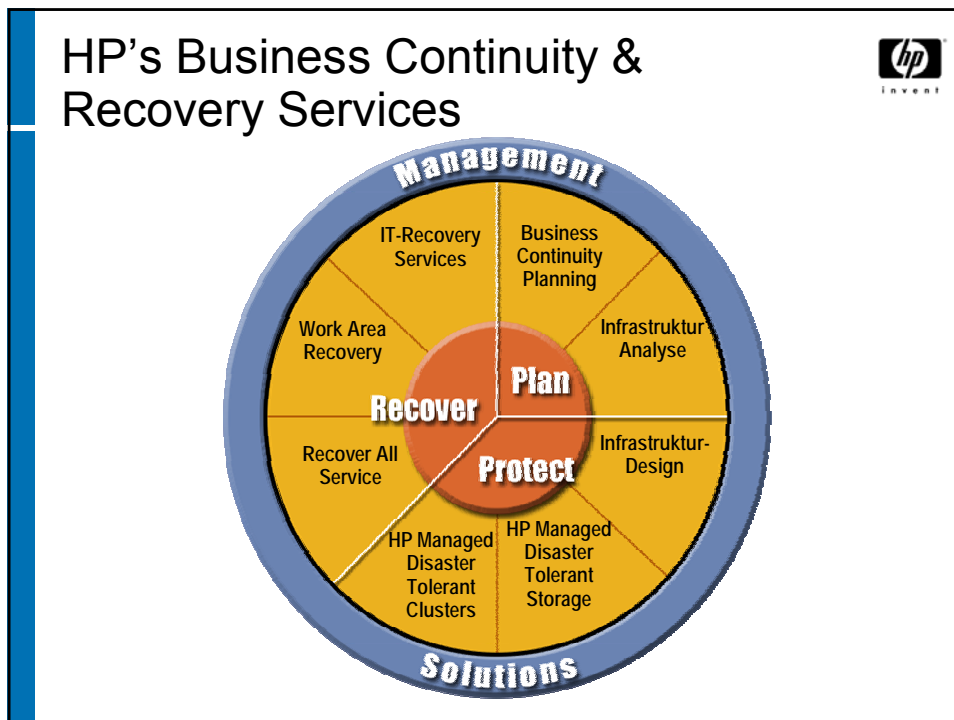



- ## Gründe für Business Continuity
-
- Haftung (nach KonTraG 1998, AktG, GmbHG)
ggfs. pers. Haftung
 - Basel II (ab 2006)
Differenzierter, ausgefeilter Risikoansatz führt zur Belohnung in Form von niedriger EK-Zuweisung bei Kreditvergabe
 - Versicherungsindustrie (Prämien & Versicherbarkeit)
 - Kundenanforderungen (z.B. Ausschreibungen)
 - immaterielle Schäden (Imageverlust, Verlust von Marktanteilen usw.)

Kosten pro Stunde Ausfallzeit

industry	application	average cost per hour of downtime (€)
financial	brokerage operations	€ 7,840,000
financial	credit card sales	€ 3,160,000
media	pay-per-view home shopping	€ 183,000
retail	(television)	€ 137,000
retail	catalog sales	€ 109,000
transportation	airline reservations	€ 108,000
entertainment	tele-ticket sales	€ 83,000
shipping	package shipping	€ 34,000

source: contingency planning research





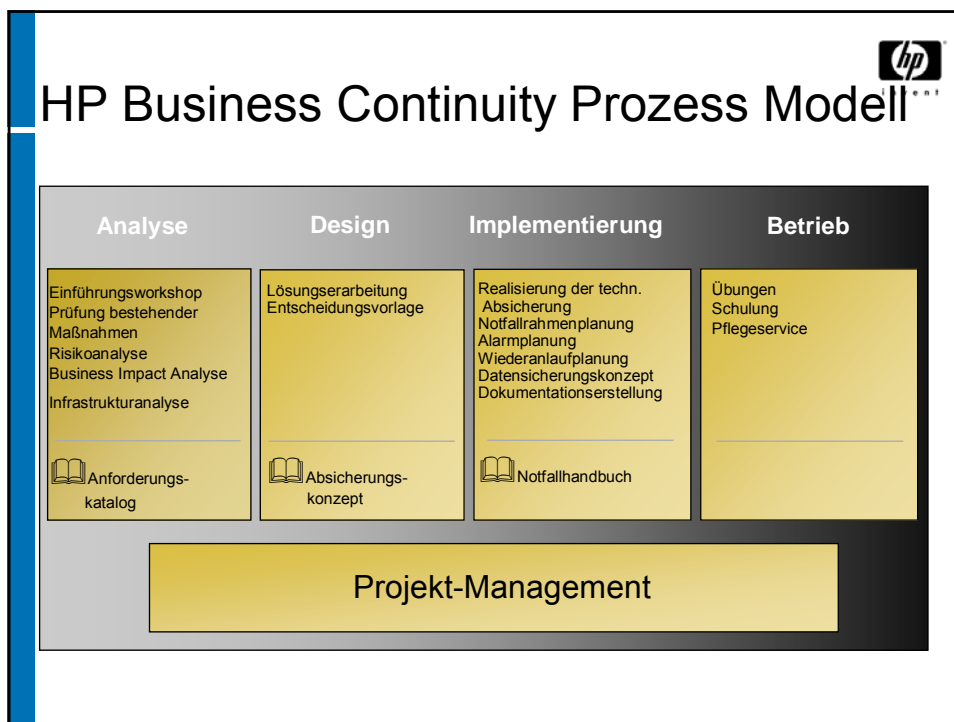
Business Continuity - Plan

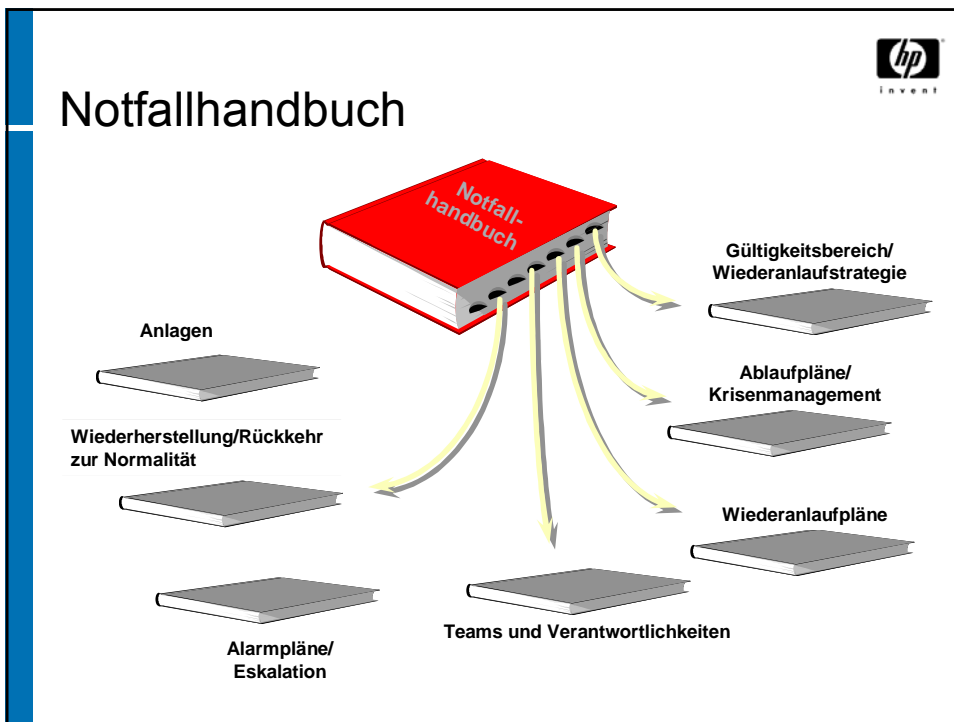
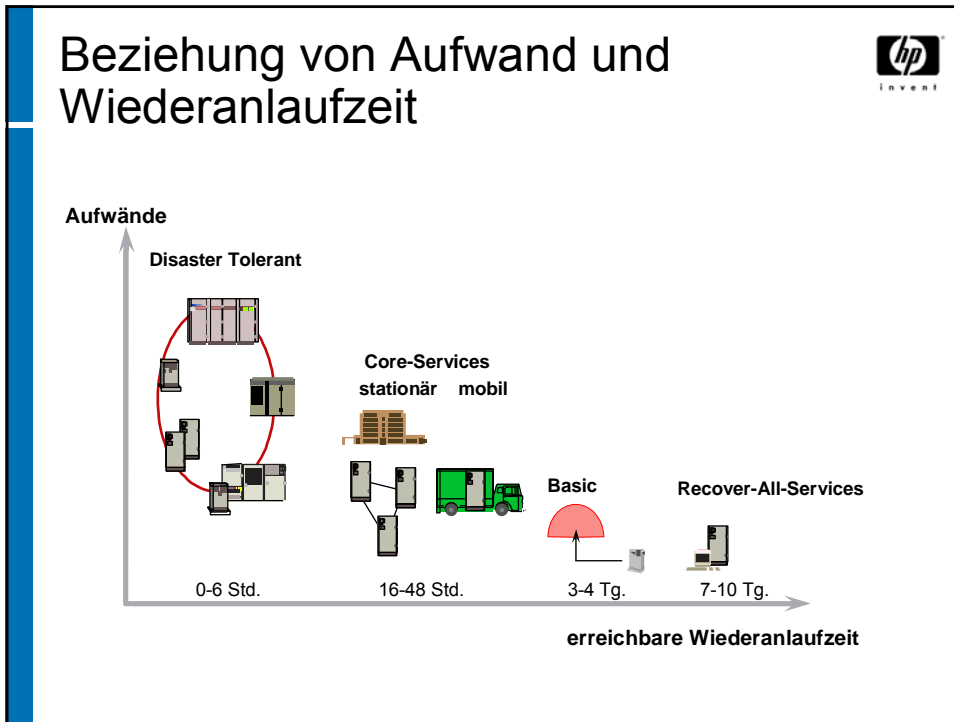
Ihre Anforderungen

- Bestimmen der Wiederherstellungszeiten entsprechend der Kritikalität Ihrer Geschäftsprozesse
- Einschätzung und Reduzierung der Risiken
- Überprüfung und Optimierung der bestehenden Wiederanlaufprozeduren
- Integration von Mitarbeitern, Geschäftsprozessen und Technologie in den gesamten Business Continuity Prozess

HP Lösungen

<ul style="list-style-type: none"> • Business Impact Analyse • Risikoanalyse • Dokumentation möglicher Wiederanlaufstrategien 	<ul style="list-style-type: none"> • Wiederanlaufverfahren • Infrastrukturanalyse/Sicherheitsbegehung • Audits • Notfallhandbuch
--	--





Hochverfügbarkeit: Gartner/IDC Studie



“HP is by far the leading vendor for servers used in clusters with 45% of mentions. In fact, this percentage is far greater than HP’s 2002 server market share, where they shipped 31% of the server units WW. This data suggests that the Digital and Compaq legacy is alive and well within the new HP.”

Source: IDC Server Clustering 2003, End-User Survey Results, April 2003

The closest Unix cluster competitors are IBM and Sun with 17% and 14% respectively. Even when considering just the clustering Unix-based software, HP has the leading 14% share while Veritas follows with 9%, IBM with 8% and Sun with 4%.

Agenda



Adaptive Enterprise
business continuity

Cluster Positionierung:
Cluster Details...



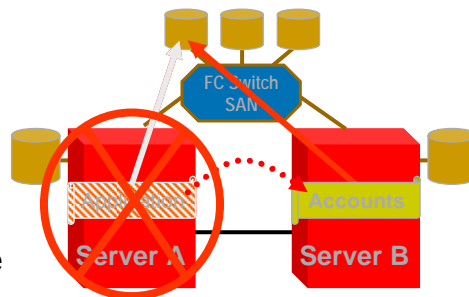
Agenda

- Technische Details:
 - Single/Multi System Image
 - Shared Root
 - System Management
 - Cluster Alias
 - Cluster File System
 - Direct and Direct Access I/O
 - Device Naming
 - Distributed Lock Manager
 - Konfigurationsfragen
 - Interconnect
 - Quorum
 - Applikation Support
 - Special coding for clusters?
 - Failover scripting
 - Auswirkungen im Fehlerfall
 - Data Replication
 - Disaster Tolerance



Multi System Image Cluster

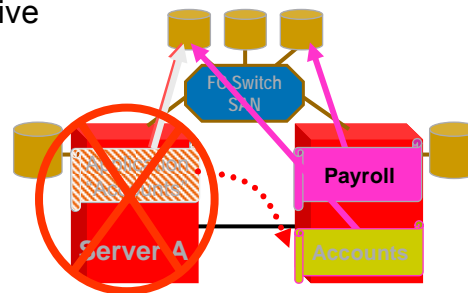
- Systeme sind relativ unabhängige Einheiten
- Platten sind physisch mit mehreren Systemen verkabelt, aber nur erreichbar für ein System beim Zugriff
- Deshalb gibt es keinen simultanen Datenzugriff von mehreren Systemen
- sieht nur Applikation-Failover Möglichkeit vor
- Die Systeme agieren unterschiedlich
- Die Systeme werden unabhängig gemanaged
- Das nennt man: active-passive



Multi System Image Cluster (continued)



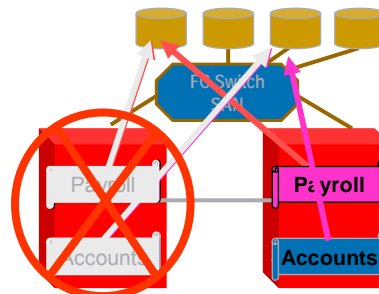
- Man kann verschiedene Applikationen oder verschiedene Instanzen der gleichen Applikation auf verschiedenen Systemen des Clusters laufen lassen
- Die Systeme agieren unterschiedlich
- Die Systeme werden unabhängig gemanaged
- Das nennt man: active-active




Single System Image Clusters




- Systeme kooperieren sehr eng miteinander
- Platten sind physisch mit allen Systemen verkabelt und erreichbar für alle Systeme jederzeit
- Deshalb ist simultaner Datenzugriff einfach
- Erlaubt beides: Applikations-Failover und simultane Ausführung
- Die Systeme agieren gleich
- Die Systeme werden gemanaged als eine Einheit
- Das nennt man: active-active





Multi oder Single System View (1 of 2)

	Multi System View	Single System View	Shared Root
HACMP AIX, Linux	Yes	No	No
LifeKeeper Linux, Windows	Yes	No	No Each
NonStop Kernel	Yes	Yes	node (1-16 CPUs)
OpenVMS Cluster Software	Yes	Yes	Yes
Oracle 9i RAC Many O/S's	Yes	Yes	Yes (effectively)
PolyServe Matrix HA Linux, Windows	Yes	No	No



Multi or Single System View (2 of 2)

	Multi System View	Single System View	Shared Root
Serviceguard HP-UX, Linux	Yes	No	No
SunCluster Solaris	Yes	No	No
TruCluster Tru64 UNIX	Yes	Yes	Yes
Veritas Cluster Server AIX, HP-UX, Linux, Solaris, Windows	Yes	No	No
Windows 2000/2003 Cluster Service	Yes	No	No



SSI und Cluster Alias

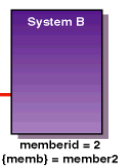
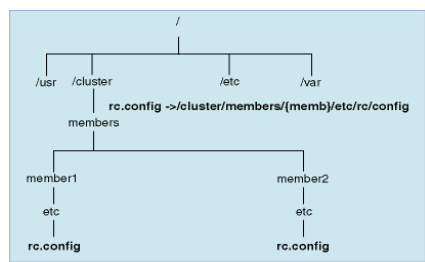
- SSI
 - Single Management und security domain
 - Systeme können rebooten ohne das Cluster auszuschalten
 - Connection Manager überwacht den Cluster Status
 - Rolling Upgrades sind voll supported, incl. mehrfacher Versionen und Patch Level
 - jeder Server ist in den Clustern zugelassen
- Cluster Alias:
 - erlaubt multiple Namen und IP Adressen als Cluster Aliases
 - erlaubt einem oder mehreren Systemen ein routing System zu sein (ein ARP Master), und setzt BIND 8.1 ein

CDSL & Search List Logical Path Resolution



Trucluster

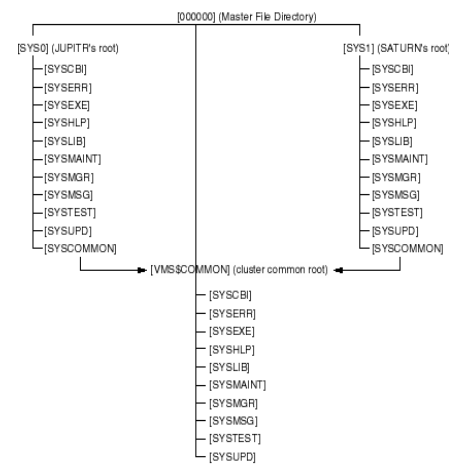
▪ TruCluster



Cluster Interconnect

OVMS

▪ VMScluster:



VM-0001A-AI

Ausführung von I/O im Client/Server Mode

hp
invent

- ◆ Client Systeme: Zugriff auf Serversysteme
 - Erforderlich sind 3 I/Os für jeden Disk Zugriff
 - Beispiel umfasst: NTFS auf Windows, MSCP auf OpenVMS, NFS auf sonstigen Systemen

Ausführung von I/O mit Direct Access I/O

hp
invent

- ◆ Direct access I/O bedeutet, dass alle Systeme im Cluster direkt auf alle Disks im Cluster zugreifen können
 - Mit voller Transparenz und Cache Koherenz
 - beseitigt 2/3 der I/Os bei jedem Zugriff auf die Disk
 - nur Tokens und ein paar Locks über den Interconnect



Cluster File Systems (1 of 2)

	Network File Systems I/O	Direct Access I/O	Distributed Lock Mgr
HACMP AIX, Linux	Yes	Raw devices and GPFS	Yes (API only)
LifeKeeper Linux, Windows	NFS	Supplied by Oracle	Supplied by Oracle
NonStop Kernel	Data Access Manager	Effectively Yes	Not applicable
OpenVMS Cluster Software	Mass Storage Control Protocol	Files-11 on ODS-2 or -5	Yes
Oracle 9i RAC Many O/S's	No (supplied by native O/S)	Raw devices and Oracle FS	Yes
PolyServe Matrix Server Linux, Windows	No (supplied by native O/S)	Yes	Yes (for file system only)



Cluster File Systems (2 of 2)

	Network File Systems I/O	Direct Access I/O	Distributed Lock Mgr
Serviceguard HP-UX, Linux	Yes	Supplied by Oracle	OPS/RAC Edition
SunCluster Solaris	Yes	Supplied by Oracle	Supplied by Oracle
Veritas Cluster Server AIX, HP-UX, Linux, Solaris, Windows	Yes	Yes (with DBE/AC for 9i RAC)	Yes (with DBE/AC for 9i RAC)
TruCluster Tru64 UNIX	Device Request Dispatcher NFS	Cluster File System	Yes
Windows 2000/2003 DataCenter		Supplied by Oracle	Supplied by Oracle



Cluster – File Systems

Verschiedene Cluster File Systeme:

	HP		VERITAS	IBM	Sun
	Service-guard	TruCluster Server	VCS	HACMP	SunCluster
File systems supported	Online JFS, JFS	AdvFS, UFS	VxFS	JFS, JFS2	UFS
Cluster-wide file system	No	CFS	CFS	GPFS	GFS ⁽³⁾
Cluster-wide I/O device access	No	Yes	No	No	Yes
Cluster-wide workload management	Yes (WLM, PRM)	Yes	Yes (VCS)	No	No



Cluster File System Anforderungen:

- für Betriebssystem xx?:
 - unterstützt SCSI und FibreChannel
 - Alle Devices (tapes, disks, CD, etc) können geshared werden
 - Non-shared Devices über Client/Server I/O
 - Cluster weite Cache Koherenz und single system file semantics
 - Cluster wide device naming
 - names are consistent across boots
 - Disk, group und user quotas sind unterstützt



Cluster Configurations (1 of 2)

	Max Servers in a Cluster	Cluster Interconnect	Quorum Scheme
HACMP AIX, Linux	32	Network, Serial, Disk Bus (SCSI, SSA)	No
LifeKeeper Linux, Windows	16	Network, Serial	Yes (Optional)
NonStop Kernel	255	ServerNet, TorusNet	No
OpenVMS Cluster Software	96	CI, Network, MC, Shared Memory	Yes
Oracle 9i RAC Many O/S's	Dependent on the O/S	Dependent on the O/S	n/a
PolyServe Matrix Server Linux, Windows	16	Dependent on the O/S	n/a



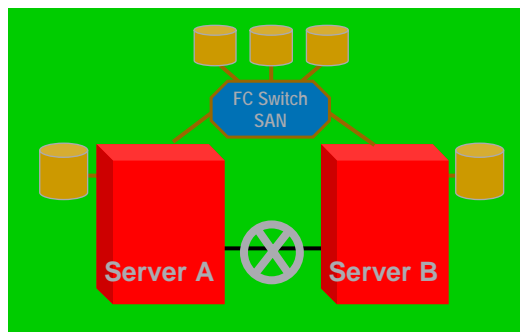
Cluster Configurations (2 of 2)

	Max Servers In A Cluster	Cluster Interconnect	Quorum Scheme
Serviceguard HP-UX, Linux	16	Network, HyperFabric	Yes = 2, optional >2
SunCluster Solaris	8	Scalable Coherent Interface (SCI), 10/100/1000 E	Yes (Optional)
TruCluster Tru64 UNIX	8 generally, 512 w/Alpha SC	100/1000 E, Memory Channel, QSW	Yes
Veritas Cluster Server AIX, HP-UX, Linux Solaris, Windows	32	Dependent on the O/S	Yes (using Volume Mgr)
Windows 2000/2003 DataCenter	4/8	Network	Yes

Ein 2-Knoten Cluster ohne eine Quorum-Disk



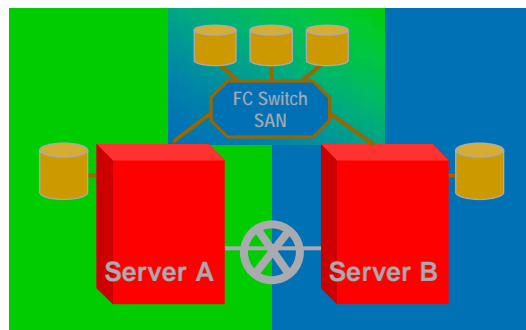
- Alle Disks werden cluster wide gemounted
 - Quorum = $(\text{expected_votes} + 2) / 2 = (2+2)/2 = 2$
 - Quorum = $(\text{actual_votes} + 2) / 2 = (2+2)/2 = 2$



Ein 2-Knoten Cluster ohne eine Quorum-Disk



- Server A und B versuchen jeder für sich ein Cluster zu bilden
 - Quorum = $(\text{actual_votes} + 2) / 2 = (1+2)/2 = 1$
 - Ergebnis ist $<$ expected_votes, es wird kein Cluster gebildet!

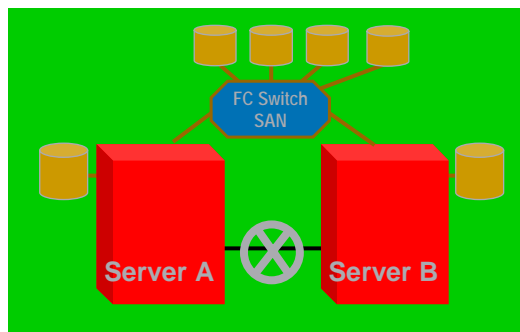


Ein 2-Knoten Cluster mit einer Quorum-Disk



Alle Disks werden cluster wide gemounted

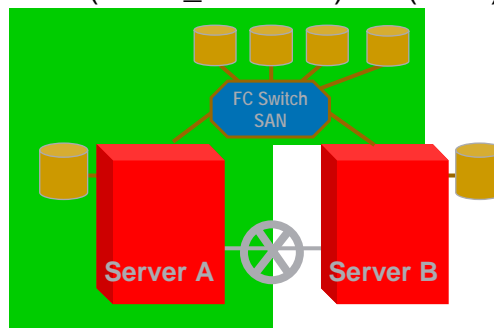
- Quorum = $(\text{expected_votes} + 2)/2 = (3+2)/2 = 2$
- Quorum = $(\text{actual_votes} + 2)/2 = (3+2)/2 = 2$



Ein 2-Knoten Cluster mit einer Quorum-Disk



- Server A bildet ein Cluster
 - Quorum = $(\text{actual_votes} + 2)/2 = (2 + 2)/2 = 2$
- Server B bildet kein Cluster
 - Quorum = $(\text{actual_votes} + 2)/2 = (1 + 2)/2 = 1$






Applikationen: single Instance

- Features
 - Kompatibel mit nicht geclusterten Systemen
 - jede Applikation, die auf einem einzelnen System läuft, läuft ebenfalls auf einem System in einem Cluster
 - Applikationen müssen nur einmal installiert sein sind aber konfiguriert und lizenziert per System
- Besonderheiten:
 - TruCluster Applikationen nutzen CAA, um zu registrieren und um Action scripts for failover zu erzeugen (in_single class),
 - VMSccluster nutzt /RESTART bei batch queues




Applikationen: Multi Instance

- Features
 - nutzt den DLM um multi-instance Applikationen zu erzeugen
 - nutzt das CFS mit standard file locking APIs um Files über das gesamte Cluster zu sharen
 - nutzt Cluster-weite Namen für Ressourcen
 - nutzt Cluster Alias um hereinkommende Requests zu verteilen




Application Support

	Single-instance (failover mode)	Multi-instance (cluster-wide)	Recovery Methods
Linux	Yes	No	Scripts
NonStop Kernel	Yes	Effectively Yes	Paired Processing
TruCluster	Yes	Yes	Cluster Application Availability
Serviceguard HP-UX, Linux	Yes	No	Packages and Scripts
VMScluster	Yes	Yes	Batch /RESTART
Windows 2000 DataCenter	Yes	No	Scripts




Resilience

	Data High Availability	Dynamic Partitions	Disaster Tolerance
Linux	DRBD	No	No
NonStop Kernel	Remote DataCenter Facility	No	Remote DataCenter Facility
TruCluster - MC S-Guard	LSM RAID-1 yes	No yes	Continous access (CA)
VMScluster	HBVS RAID-1	Yes	DTCS, CA
Windows 2003 DataCenter	NTFS RAID-1	No	CA



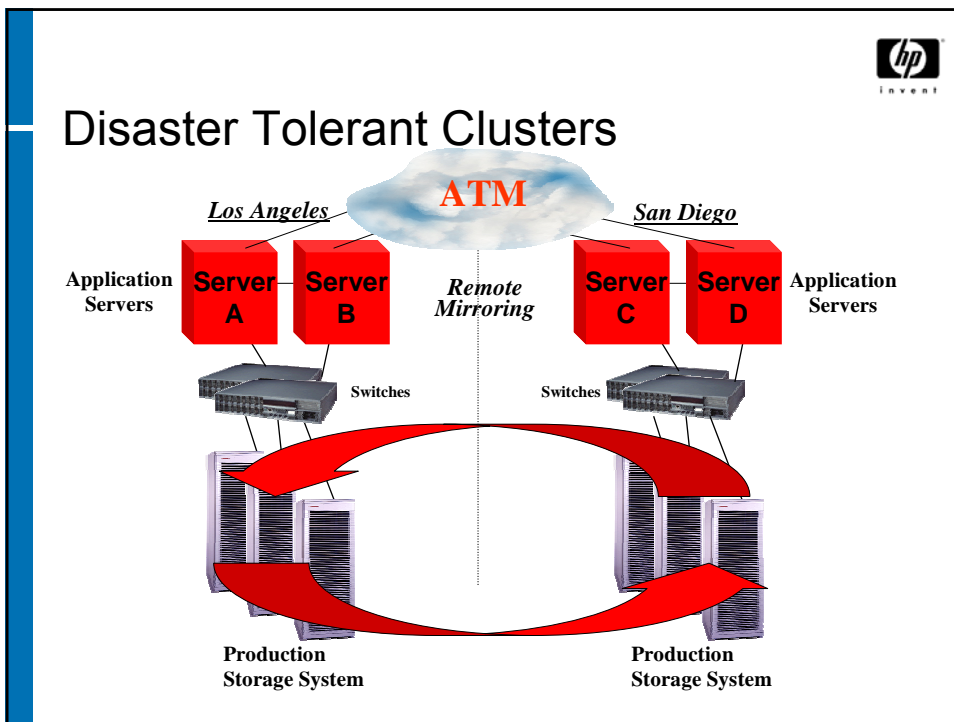
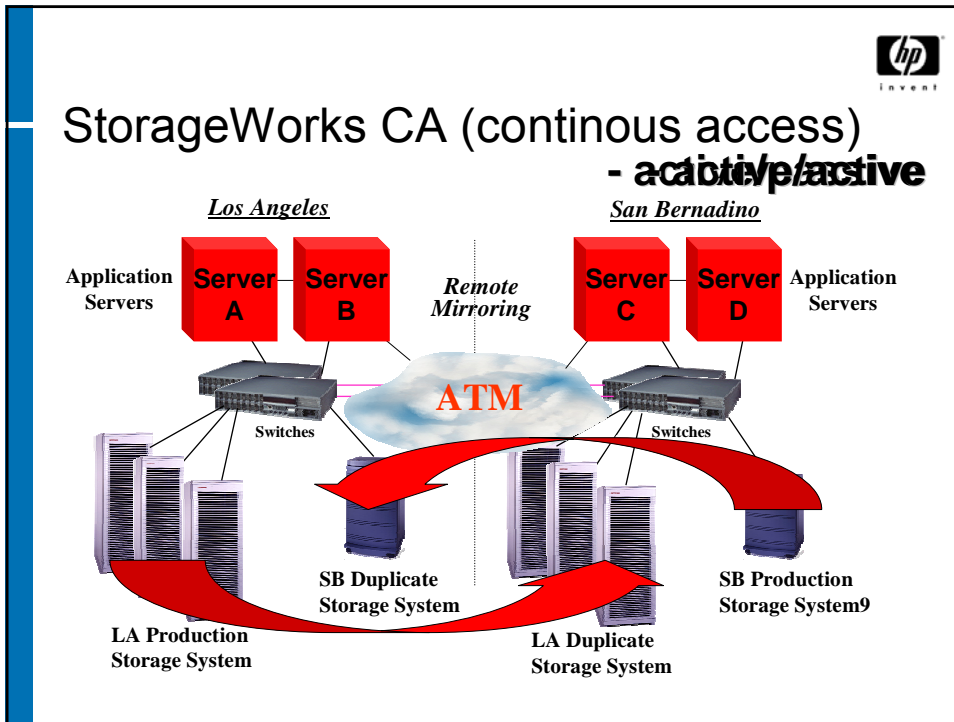
Resilience (1 of 2)

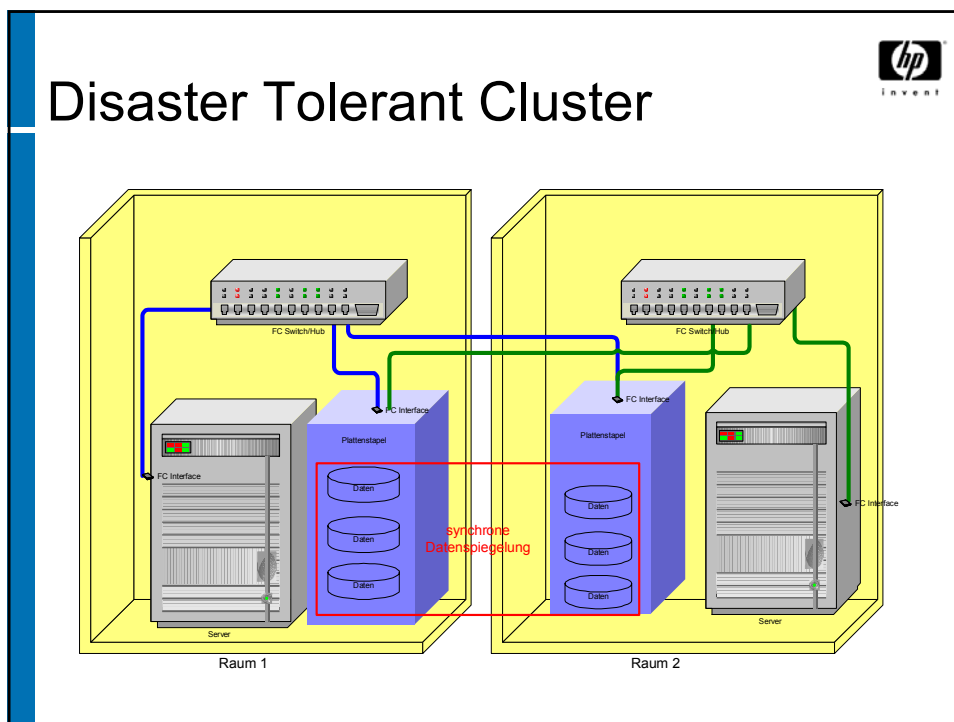
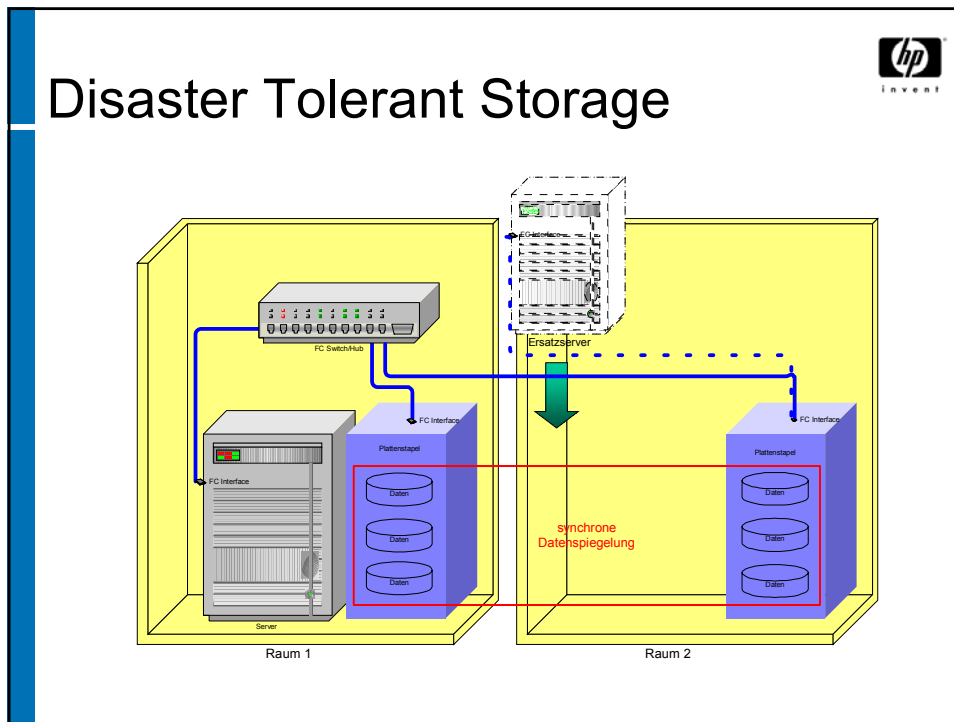
	Dynamic Partitions	Disk High Availability	Path High Availability	Alias
HACMP AIX, Linux	DLPars (AIX 5.2)	RAID-1 (Logical Volume Manager)	Multi-path I/O (a/p)	Not shared
LifeKeeper Linux, Windows	No	RAID-1 Distributed Replicated Block Device (DRBD)	Multi-path I/O (a/p)	Not shared
NonStop Kernel	No	RAID-1, Process Pairs	Multi-path I/O (p)	Shared
OpenVMS Cluster Software	Galaxy	RAID-1 (Host Based Volume Shadowing)	Multi-path I/O (p)	Shared
Oracle 9i RAC Many O/S's	No	Dependent on the O/S	Dependent on the O/S	No
PolyServe Matrix HA Linux, Windows	No	Dependent on the O/S	Dependent on the O/S	No



Resilience (2 of 2)

	Dynamic Partitions	Disk High Availability	Path High Availability	Alias
Serviceguard HP-UX, Linux	vPars	RAID-1 (MirrorDisk/UX)	Multi-path I/O LVM (a)	Not shared
SunCluster Solaris	No	RAID-1 (Solaris Volume Manager)	Multi-path I/O (p)	Not shared
TruCluster Tru64 UNIX	No	RAID-1 (Logical Storage Manager)	Multi-path I/O (a)	Shared
Veritas Foundation Suite AIX, HP-UX, Linux, Solaris, Windows	No	RAID-1 (Veritas Volume Manager)	Multi-path I/O (p)	No (simulate by Traffic Director)
Windows 2000/2003 DataCenter	No	RAID-1 (NTFS)	Secure Path (a/p)	Not shared







Zusammenfassung

- Haupteigenschaften moderner Cluster sind:
 - ✓ Single root,
 - ✓ single system image,
 - ✓ cluster file system,
 - ✓ quorum scheme,
 - ✓ interconnects,
 - ✓ host based RAID
 - ✓ Parallelized I/O und multiple reader/writer direct access I/O
 - ✓ Active/active HBAs und active/active Cluster interconnects
 - ✓ 32-Knoten QSW Super-computing Cluster und 96 Knoten mixed Architektur/interconnect Cluster
- In Kürze alles auch für HPUX verfügbar



Zusammenfassung

- jedes System bietet eine "high availability" Option
 - Aber die Recovery Zeiten variieren von "sehr lange" bis zu transparent kurz
- jedes System skaliert "outside the box"
 - aber in unterschiedlichen Größen von 2-Knoten bis zu 255-Knoten Clustern
- jedes System hat die Option von Disaster Toleranz
 - aber die Technologien variieren von one-way Datenreplikation zwischen separaten Clustern bis zu voll active/active Kooperation von einem Cluster, das sich über geographisch verteilte Rechenzentren erstreckt
 - →→→
- **Man verstehe die Optionen und wähle die richtigen Technologien!**
(die do's und don'ts jeder Technologie)



Resources

- Linux
 - <http://linux-ha.org/>
- NSK
 - <http://h71033.www7.hp.com/page/techdoc.html>
 - TruCluster
 - <http://h30097.www3.hp.com/docs/>
 - VMScluster
 - <http://h71000.www7.hp.com/doc/>
 - Win2003
 - <http://www.microsoft.com/windowsserver2003/techinfo/overview/articleindex.msp#Books>
 - "In Search of Clusters", Gregory F. Pfister
 - ISBN 0-13-899709-8



Resources

- SunCluster
 - <http://www.sun.com/software/cluster/index.html>
- TruCluster
 - http://h30097.www3.hp.com/docs/pub_page/cluster_list.html
- Veritas Cluster Server
 - <http://www.veritas.com/van/articles/3245.html>
- Windows 2000/3
 - <http://www.microsoft.com/windows2000/en/datacenter/help>
- Books
 - "Clusters for High Availability", Peter Weygant, ISBN 0-13-089355-2
 - "In Search of Clusters", Gregory F. Pfister, ISBN 0-13-899709-8

**Herzliche
n Dank!**

Engagement für
den Kunden



© 2004 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice



Ein Baustein für
“Adaptive Enterprise”
Business
continuity:

Cluster im Vergleich
wie gehts weiter?



© 2004 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice

CPU-Leistungen



CPU Leistung 2003

*Noch immer:
Verdoppelung der
Performance alle 18
Monate*

Itanium2 bereits ganz
vorne mit dabei

Sun fällt immer weiter
zurück
(Stand 8-2003)

Chart 1: SPECint2000

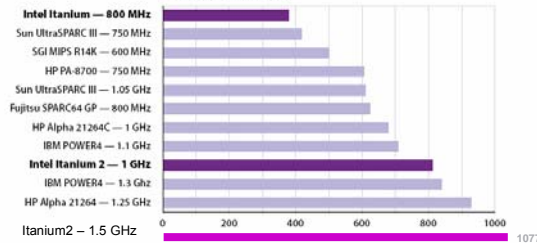
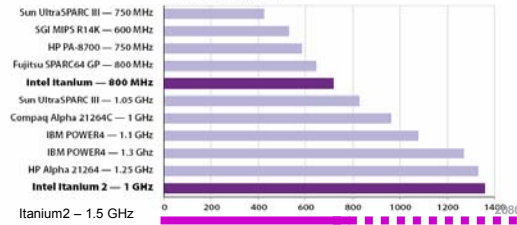


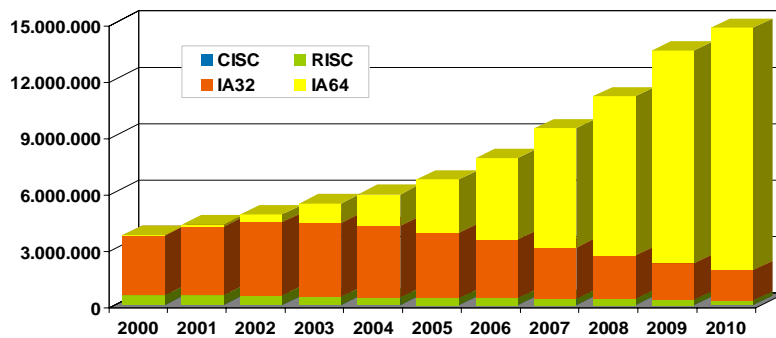
Chart 2: SPECfp2000



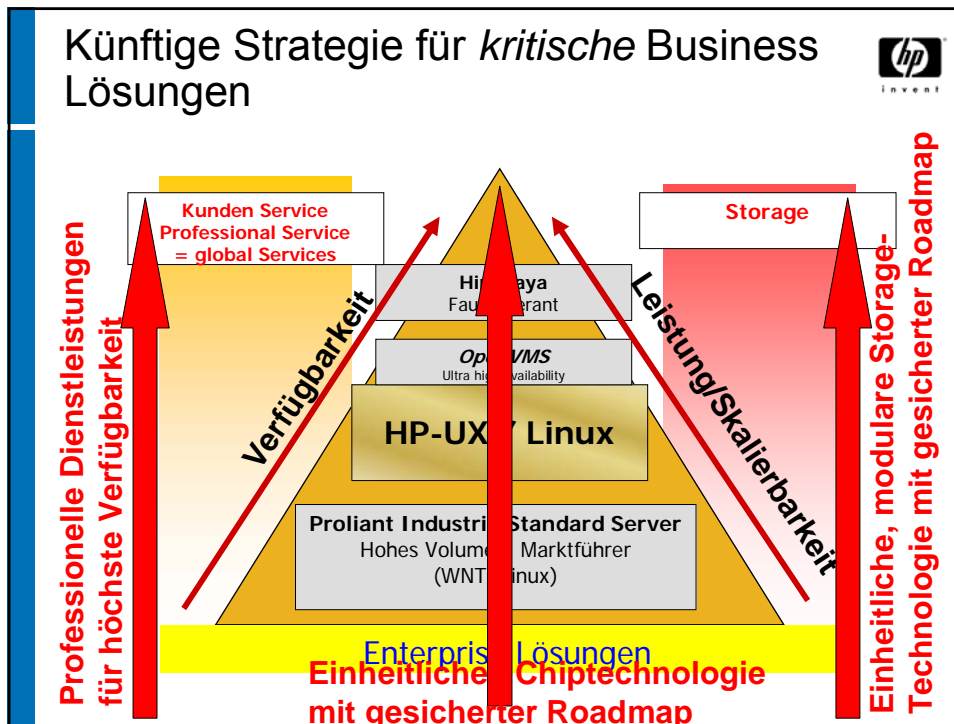
Server-Architektur Marktübersicht , 2000-2010



Unit shipments



Source: IDC



Hewlett Packard wandelt sich....

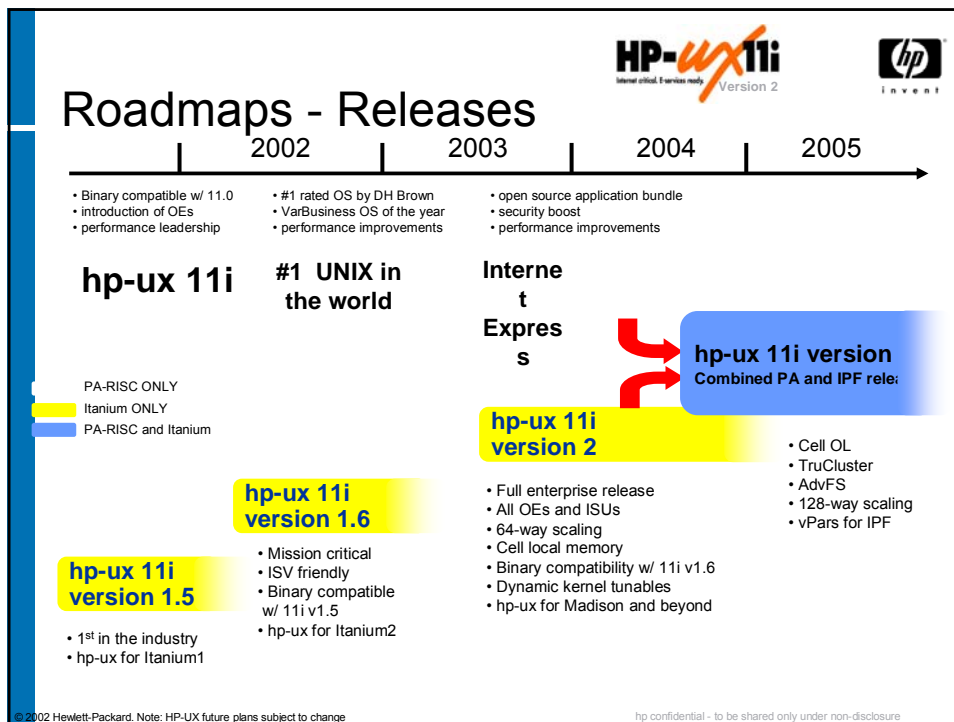
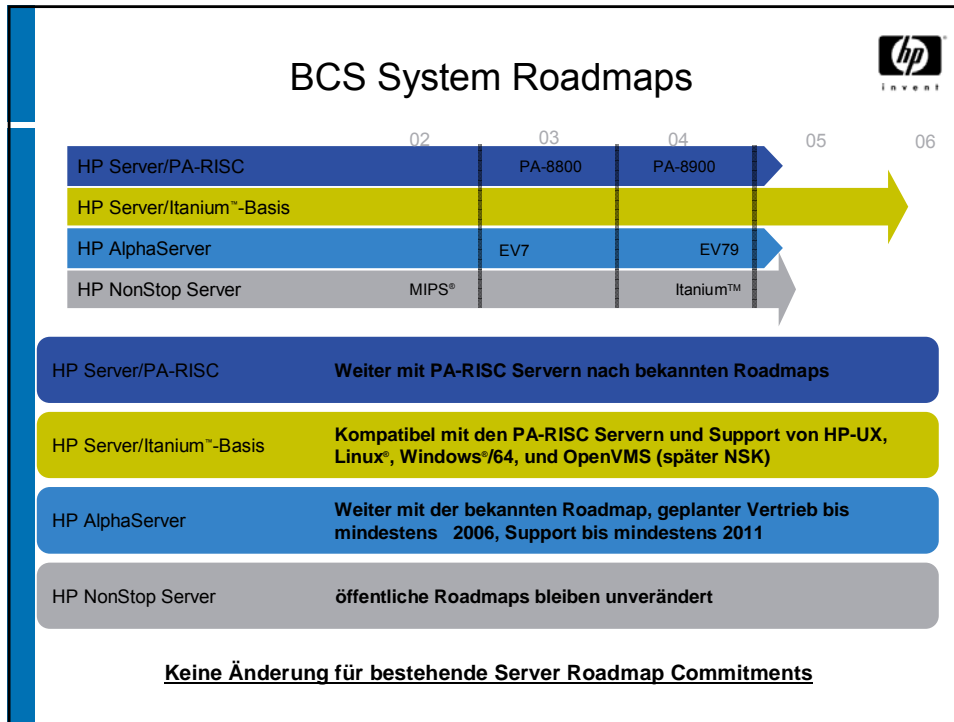
Unsere Basis:

- Wir besitzen exzellentes Ingenieurwissen und Innovationskraft (fundierte Kapitalausstattung) → **technisches Know-How**
- Wir führen bei offenen Systemen und Architekturen und definieren Standards → **innovative Systemarchitekturen**
- Wir haben tiefes Verständnis komplexer Kundenanforderungen (→ Hochleistungs- und fehlertolerante Systeme) → **Integration**

Wir bieten Ihnen:

- beste Produkte – Kosten effektive Infrastruktur (geringe TCO)
- erweiterte Global Services Kompetenz und weltweiten Support
- mehr als nur ein Produkt...


optimale IT-Infrastruktur



hp-ux 11i:

das beste, high-end UNIX Operating System ist jetzt das #1-zertifizierte Operating System weltweit

Basierend auf die Auslieferungen der ersten 7 Monate zeigt hp-ux 11i nur 1/3 SW-Fehler und braucht nur die Hälfte der benötigten Patches im Vergleich zu Solaris 8

System	Patches	Defects
hp-ux 11i	~120	~230
Solaris 8	~280	~750

>3x bessere Qualität

source: www.sun.com

2002 UNIX Function Review

#1 in all categories:

- Scalability
- Reliability, availability, and serviceability
- Systems management
- Internet and Web application services
- Directory and security services

5 4 3 2 1

hp-ux 11i

Solaris 8

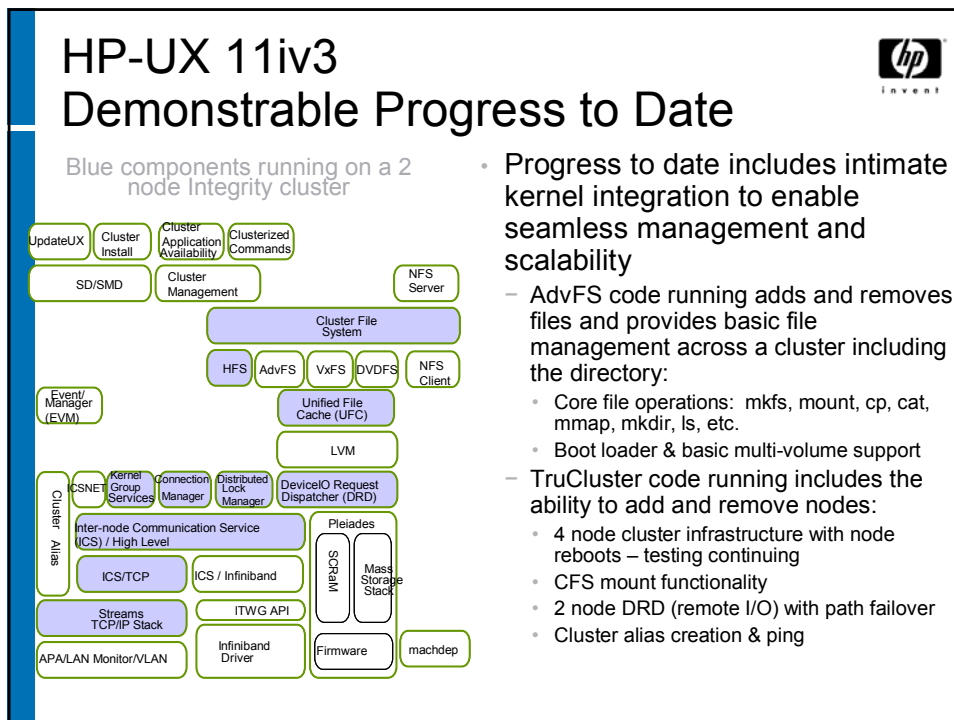
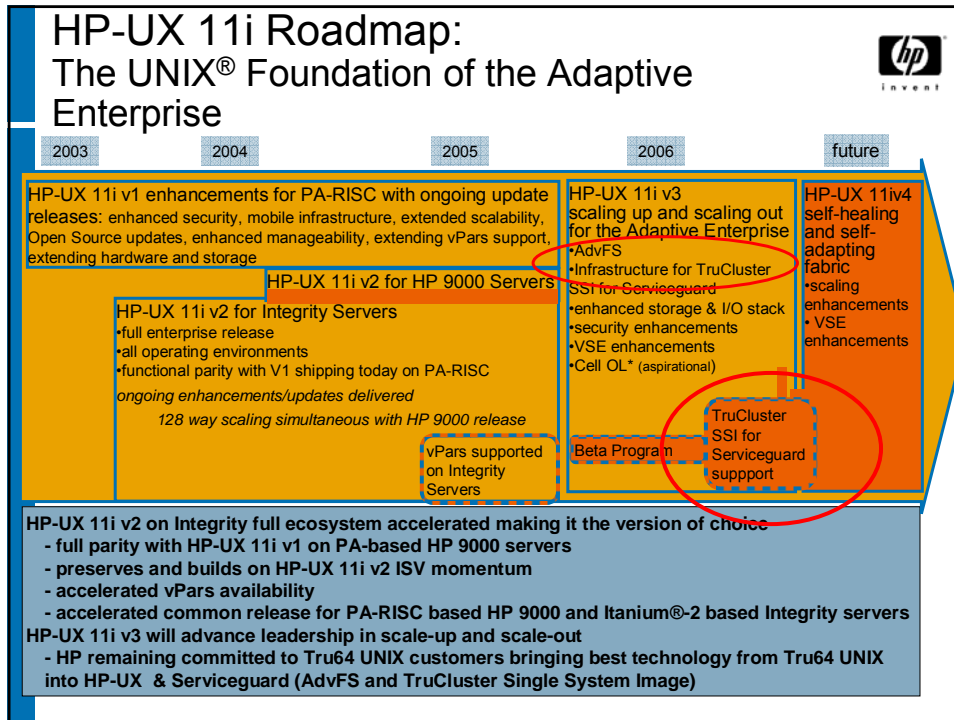
IBM AIX 5L

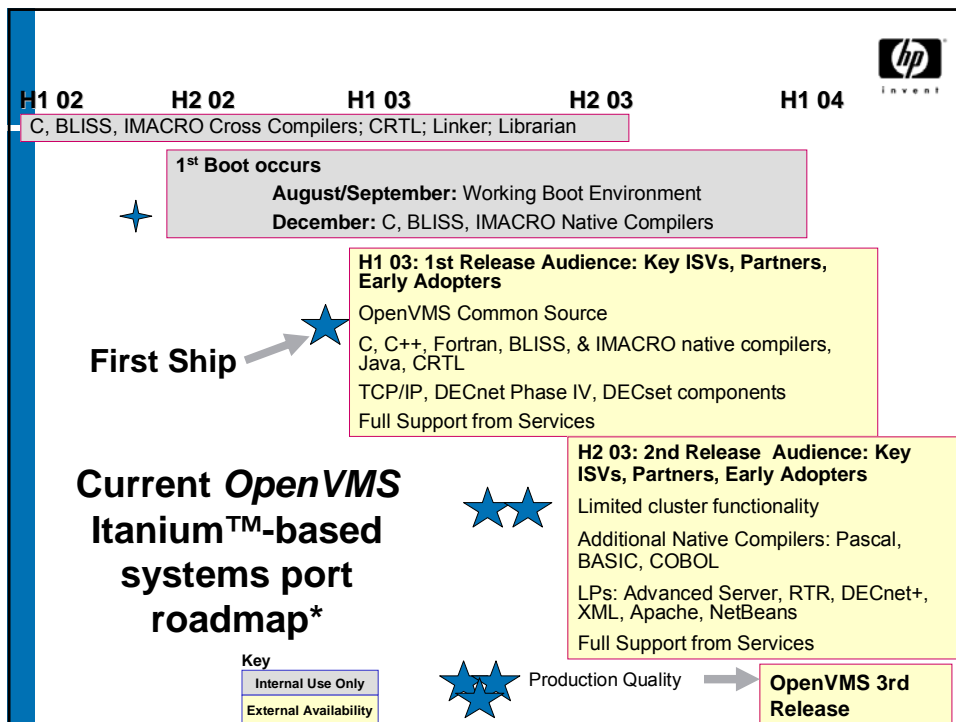
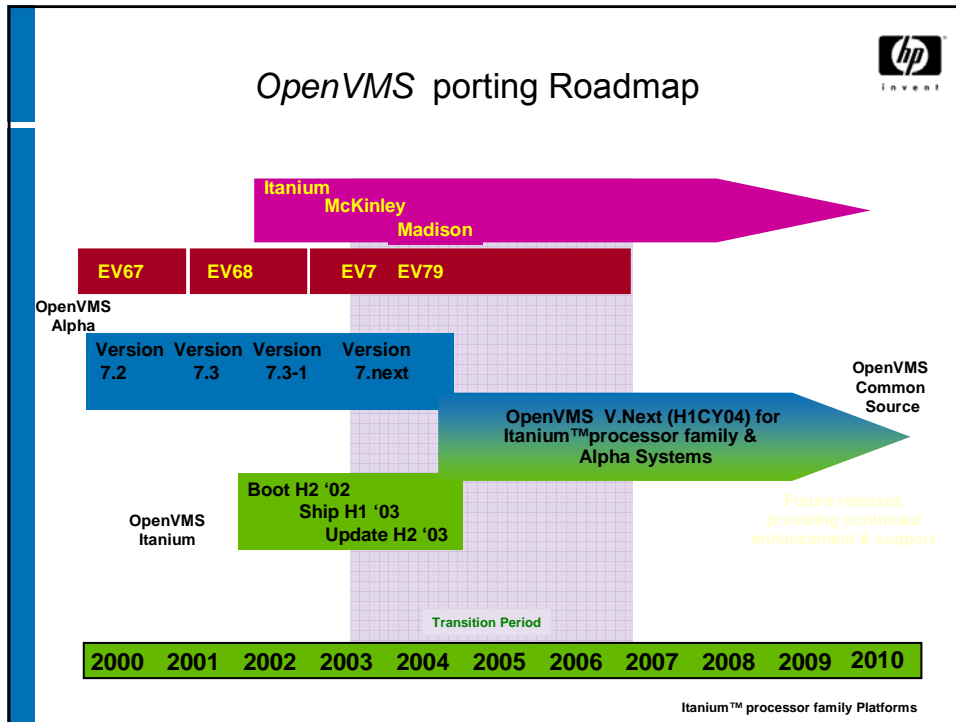
HP Server Roadmap Details

	02	03	04	05	06
hp server PA-RISC	HP Superdome PA-8700 speed-up	PA-8800	PA-8900		
	HP Server rp8400 PA-8700 speed-up	PA-8800	PA-8900		
	HP Server rp7410 PA-8700 speed-up	PA-8800	PA-8900		
	HP Server rp5400 PA-8700	HP Server rp5610 PA-8700	PA-8900		
	HP9000 A-class PA-8700 speed-up				
Itanium™ hp Server	HP Server rx9610 Itanium™ processor	HP Superdome Madison 32p future Itanium™ 64p Madison Itanium™ 16p	future Itanium™ 32-128p future Itanium™ 16p	future Itanium™ 32-128p future Itanium™ 16p	
	McKinley 4p	Madison 4p	future Itanium™ 8p	future Itanium™ 8p	
	McKinley 2p	Madison 2p	future Itanium™ 4p	future Itanium™ 4p	
			future Itanium™ 2p	future Itanium™	
hp AlphaServer	HP AlphaServer GS EV68 (1-32p)	EV7 (8-64p) GS1280	EV79		
	HP AlphaServer ES EV68 (1-4p)	EV7 (2-8p) ES 80	EV79		
	HP AlphaServer DS EV68 (1-2p)				

in-box HP Upgrades und binäre Kompatibilität

nahtlose Migration auf Kundenwunsch





Herzlichen Dank!

Engagement für
den Kunden

