



Building Geo Clusters with HP Storage



Matthias Popp
Disaster Recovery Software Architect
Hewlett-Packard

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



Agenda

- What will you learn here
- Business Continuity Background
- Cluster Extension XP integration into MSCS
- What has changed
- Best practices
- What will be changed in the future

Cluster Extension XP

“The disaster recovery solution that extends a local cluster over metropolitan-wide distance.

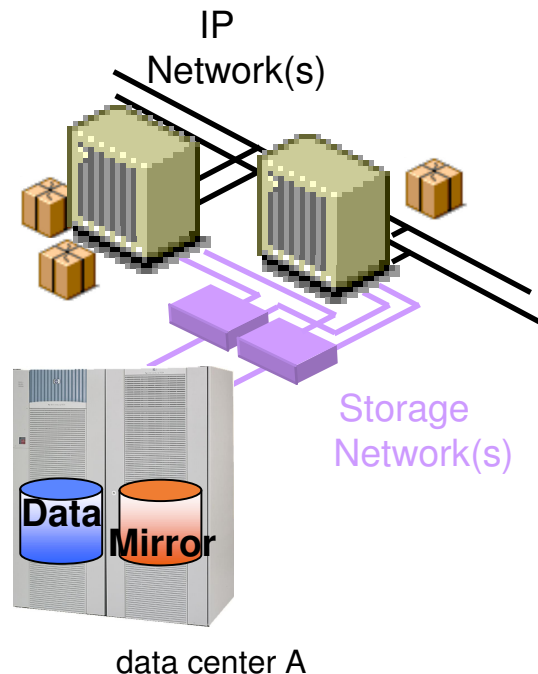
Hence, an extended cluster solution - for Solaris, AIX, Linux, Windows 2000 and now Windows Server 2003, 32-bit and 64-bit.”

Agenda

- What will you learn here
- **Business Continuity Background**
 - What and why
 - Product positioning and product structure
- Cluster Extension XP integration into MSCS
- What has changed
- Best practices
- What will be changed in the future

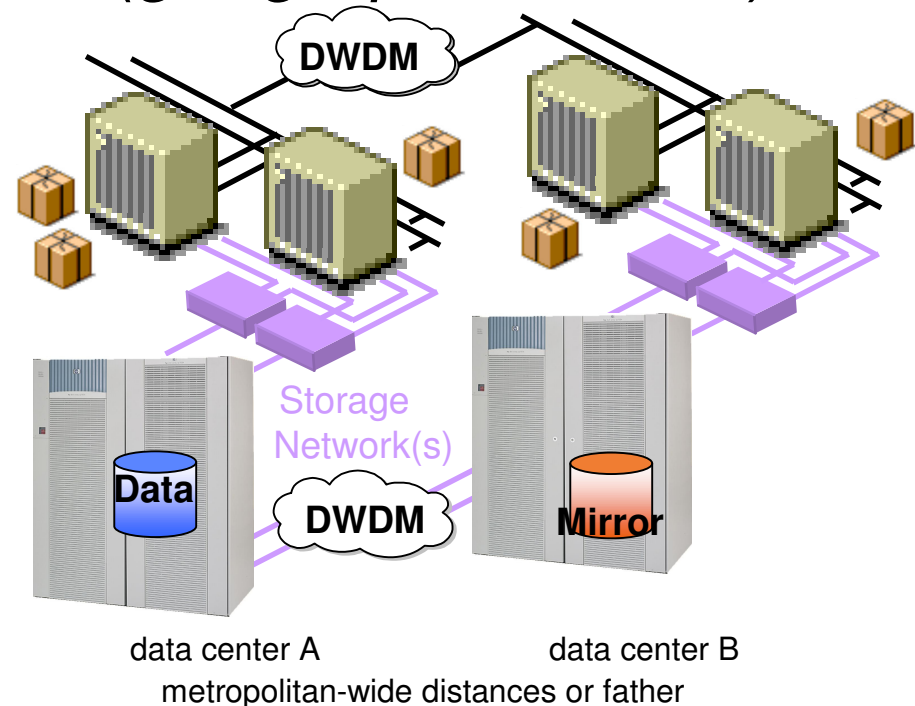
What is an extended cluster ?

local cluster



- single local cluster solution - distance limited to single data center
- limited disaster recovery solution capabilities

extended cluster (geographic cluster)



- extends single local cluster to multiple data centers – network latency is the limiting factor
- true disaster recovery solution

Why extended clusters ?

known causes of downtime

A/C Failure
 Acid Leak
 Asbestos
 Bomb Threat
 Bomb Blast
 Brown Out
 Burst Pipe
 Cable Cut
 Chemical Spill
 Condensation
 Construction
 Coolant Leak
 Cooling Tower
 Leak
 Corrupted Data
 Data Center
 Move
 Diesel Generator
 Earthquake
 Electrical Short

Erased Tapes
 Evacuation
 Explosion
 Fire
 Flood
 Fraud
 Frozen Pipes
 Hacker
 Hail Storm
 Halon Discharge
 Hardware Failure
 Human Error
 Humidity
 Hurricane
 HVAC Failure
 H/W Error
 Ice Storm
 Insects
 Lightning

Lost Data
 Low Voltage
 Microwave Fade
 Network Failure
 PCB
 Contamination
 Plane Crash
 Power Outage
 Power Spike
 Power Surge
 Programmer Error
 Raw Sewage
 Relocation Delay
 Rodents
 Roof Cave in
 Sabotage
 Shotgun Blast
 Shredded Data
 Smoke Damage
 Snow Storm

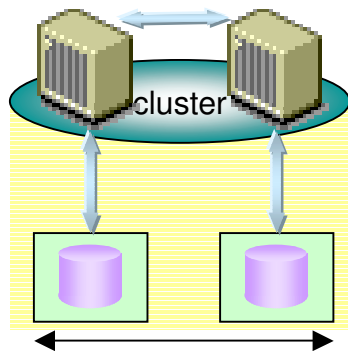

reasons for extended cluster solution

Source: Contingency Planning Research, Inc.

Sprinkler Discharge
 Static Electricity
 Strike Action
 S/W Error
 S/W Ransom
 Terrorism
 Theft
 Toilet Overflow
 Tornado
 Train Derailment
 Transformer Fire
 UPS Failure
 Vandalism
 Vehicle Crash
 Virus
 Water (Various)
 Wind Storm
 Volcano
 Epidemic

Why Cluster Extension XP? (1)

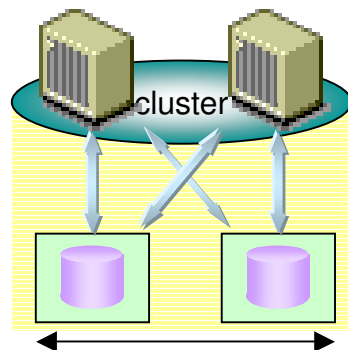
today: host-based data replication



host-based data replication

- host-based data replication utilizes additional CPU cycles
- “write-twice” host based mirroring means I/O performance impact and IP network congestion

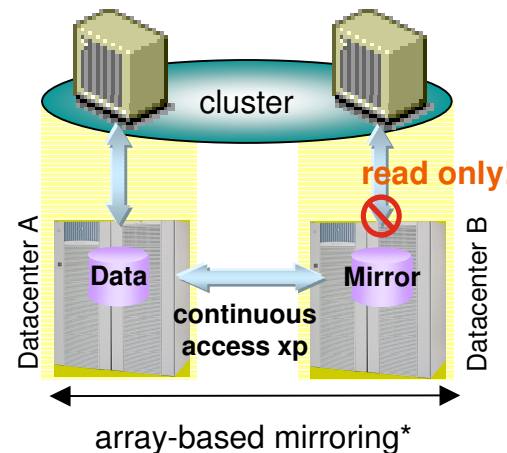
today: host-based mirroring



host-based mirroring

- multiple I/O paths to storage subsystems means solution limited by link distance
- “write-twice” host based mirroring means I/O performance impact

today: array-based mirroring*



array-based mirroring*

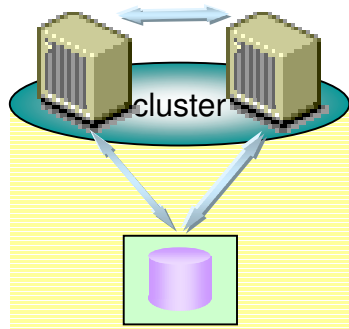
- systems not connected to Data and Mirror volumes
- system connected to mirror has read-only access (**expects** read/write in failover scenarios) or is not visible (which would require reboot to recognize devices once they become visible)

* without cluster extension xp

Why Cluster Extension XP? (2) (for MS Cluster Service only)



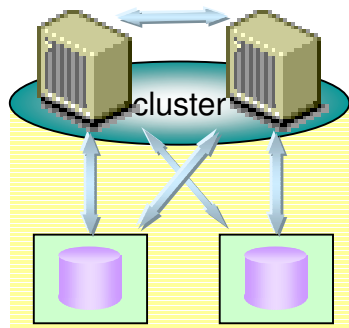
today: single shared quorum disk



single shared quorum disk

- if nodes lose all network connections they will compete for the quorum disk
- only quorum disk owner and node who can communicate with the quorum disk owner survive
- quorum disk is a single point of failure

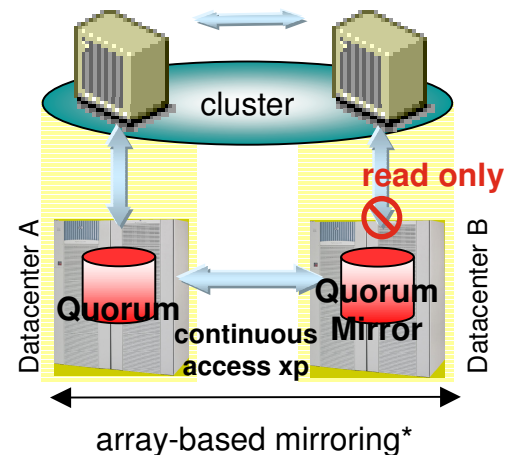
today: host-based mirrored quorum disk



host-based mirroring

- quorum disk host-mirrored in campus data centers (either LAN or SAN or both)
- introduces “split-brain” situations

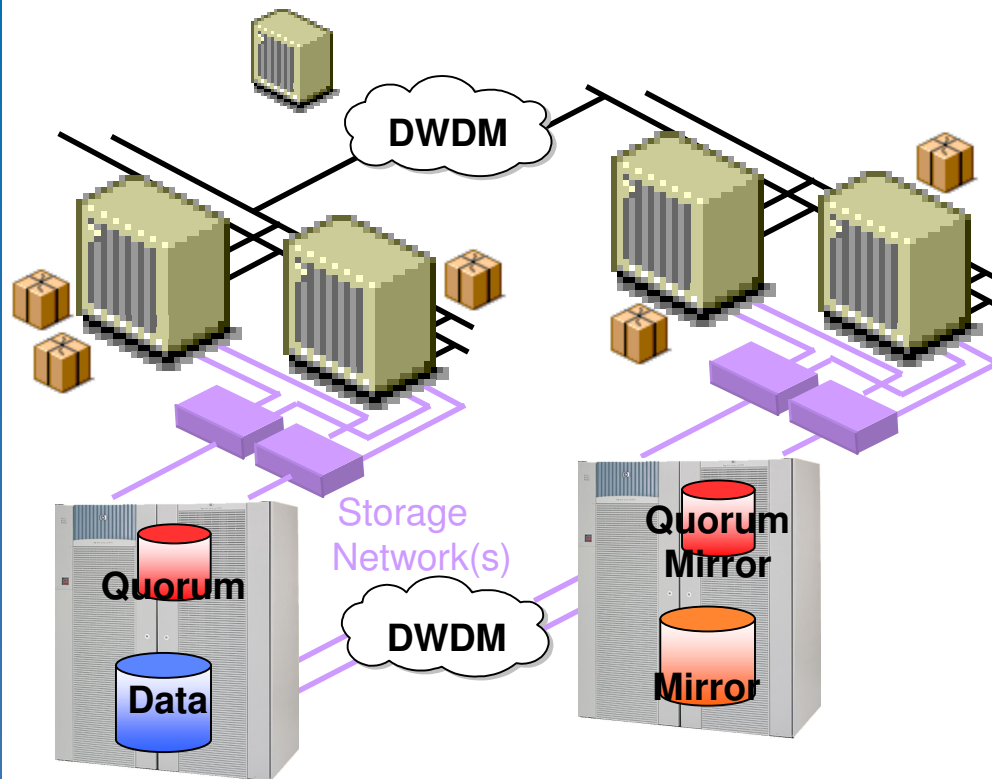
today: array-based mirrored quorum disk*



- systems not connected to primary quorum and its mirror
- system connected to mirror has read-only access (**expects** no access in normal mode and read/write in failover scenarios)

* without cluster extension xp

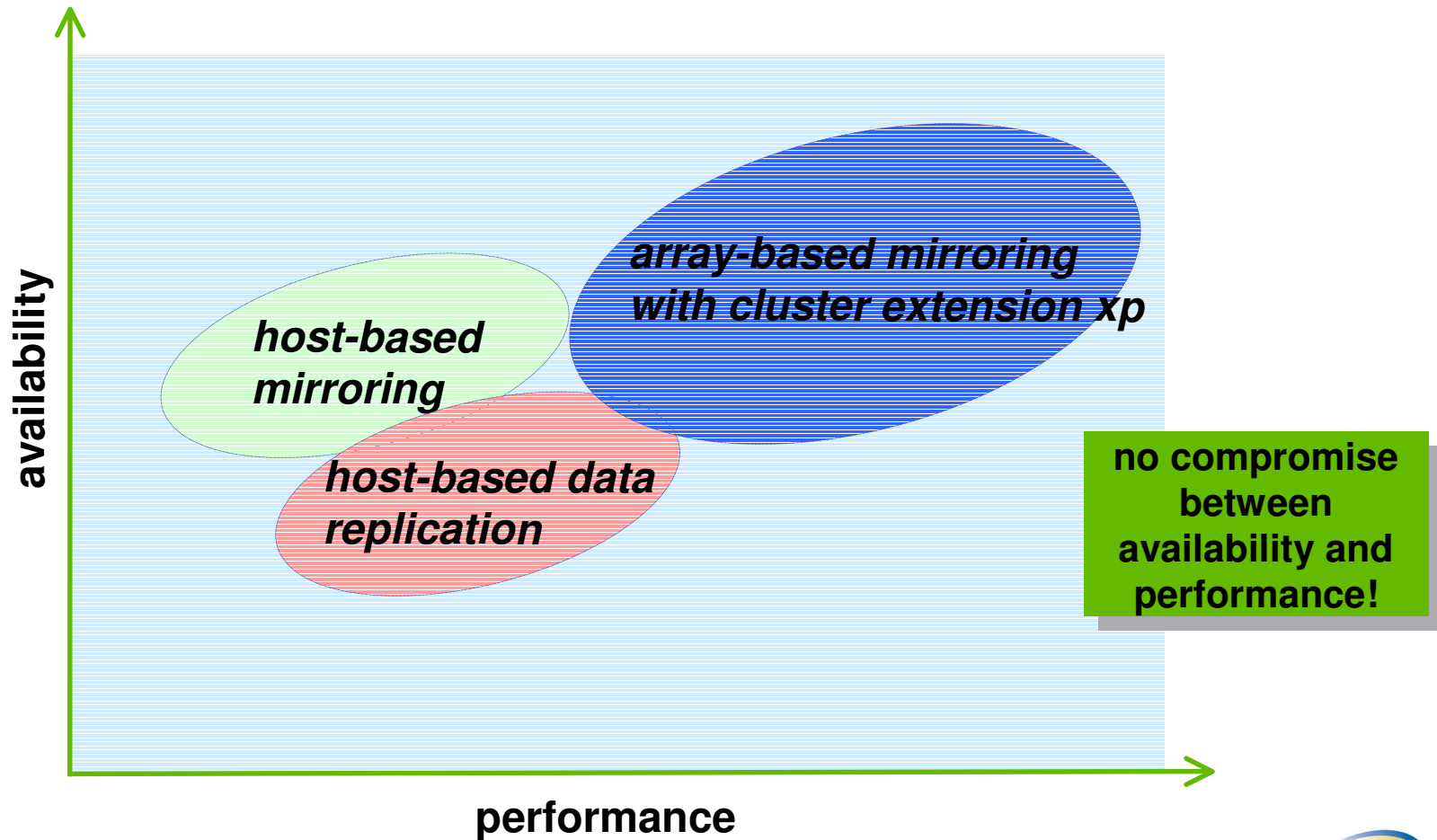
Cluster Extension XP - Benefits



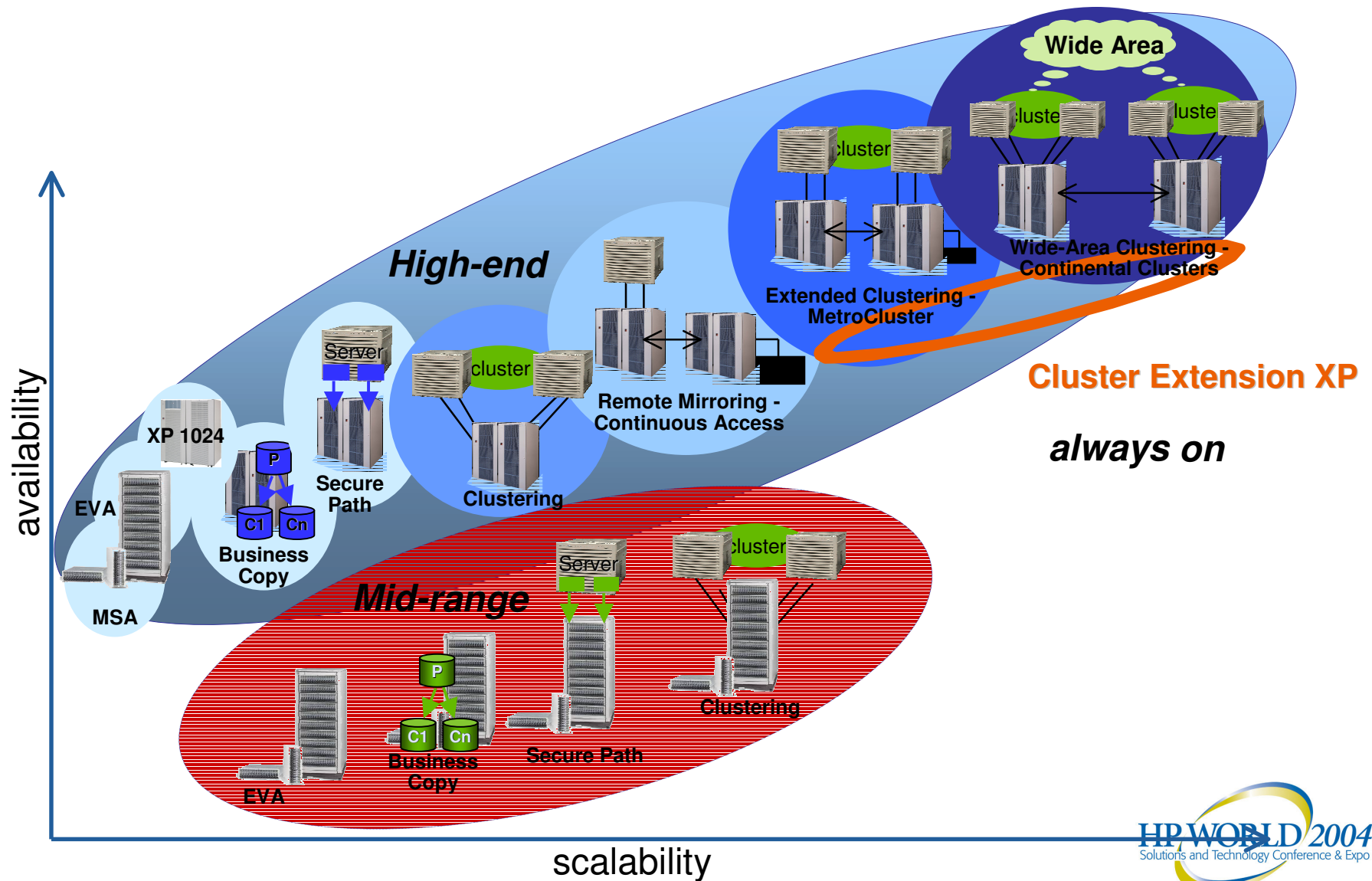
metropolitan-wide distances or
father

- Disaster recovery to protect against the risk of downtime, whether planned or unplanned
- Automatic failover/failback to reduce the complexity involved in a disaster recovery situation
- Ensures the highest standards in data integrity by leveraging XP disk array remote mirroring
- CLX quorum service provides market's only proven solution for quorum resource and full protection against 'split-brain' situations using arbitrator technology
- Fully Microsoft-certified: see Windows Catalog (HCL)
- World's first DR solution on Linux

Cluster Extension XP - Positioning



Scalable Business Continuity Solutions



CLX 2.04.00 - What is new ?

Microsoft Cluster service – CLX 2.04.00 version

- Supports Windows 2000 SP4 with Auto Path/Secure Path
- Supports Windows Server 2003 EE & DC with new filter driver (2 filter drivers necessary)
- Enhancements for CLX resource type (online performance improvements, added fixes for MS issues)
- Reliability enhancements for the Quorum Service/Filter Driver (e.g. faster, more automated recovery – less manual intervention)
- Fixes for the External Arbitrator, enhancements in Pair/Resync Monitor
- New install program
 - Allows separate installation of CLI
 - Allows separate installation of CLX resource type and quorum service
 - Allows choice of OS from start page
- Finished MS Geo Cluster certification in March '04

Cluster Extension XP - What is new ?



Veritas Cluster Server

- added support for Veritas Cluster Server 4.0

IBM HACMP

- added support for IBM HACMP 4.5 and 5.1 on AIX 5.1, 5.2 with Auto Path and native MPIO

HP Serviceguard on Linux

- added support for Serviceguard 11.14, 11.15
 - on Red Hat 7.2 AS 2.1 (kernel 2.4.2 or later) (CLX 2.02.00)
 - on SuSE Linux 8.0/United Linux 1.0 (CLX 2.02.01)
 - 64-bit kernel (SG 11.15 on SuSE/United Linux)

Cluster Extension XP - Summary

- Host-based software that seamlessly **integrates** continuous access xp remote mirroring and management with the leading cluster software:
 - Microsoft Cluster Service for Windows 2000 and Windows Server 2003
 - VERITAS Cluster Server for Solaris
 - HACMP for IBM AIX
 - Serviceguard on Linux (Red Hat and SuSE)
- **Extends** a local cluster beyond a single data center, enabling reliable metropolitan and WAN node-level failover/failback operation
- Automatic failover/failback and redirection of mirrored pairs **including** the quorum disk in failure scenario
- **No reboots** and LUN presentation/mapping changes necessary during failover
- Monitors and recovers disk pair synchronization on an app level
- **Offloads** the host from data replication tasks
- Extensive volume states rules database

Cluster Extension XP - Product Requirements



- **synch** replication: Continuous Access XP & Raid Manager XP (bundled w/ CA XP)
- **async** replication: Continuous Access XP + Continuous Access XP Extension
- Cluster software:
 - VERITAS Cluster Server for Sun Solaris
 - HACMP for IBM AIX
 - Microsoft Cluster Service for
 - Windows 2000 Advanced Server or Datacenter Server
 - Windows Server 2003 Enterprise Edition or Datacenter Edition (32/64-bit)
 - Serviceguard on Linux
- Command View XP on Management PC
- XP 48, XP 512, XP 128, XP 1024

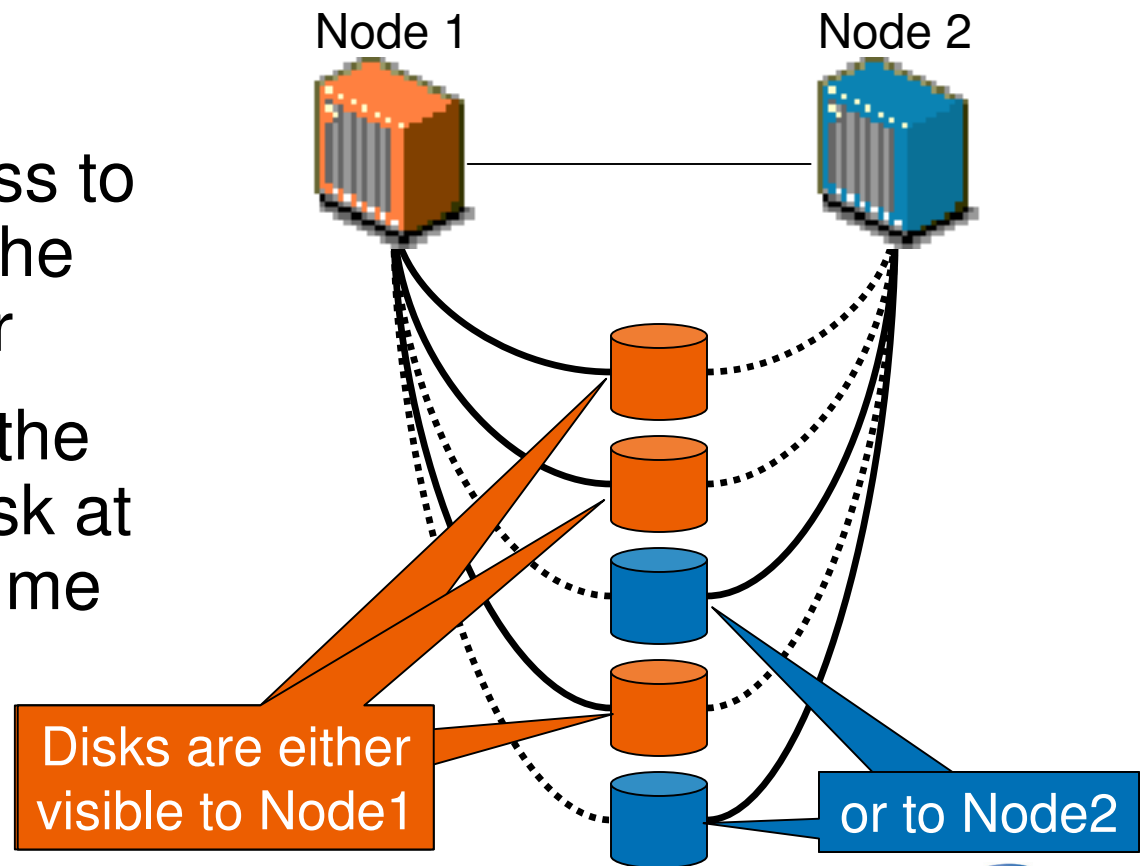
Agenda

- What will you learn here
- Business Continuity Background
- Cluster Extension XP integration into MSCS
 - The issue
 - Distributed quorum (filter) service and external arbitration
 - Resource type and cluster administrator extension
 - Resource dependencies
- What has changed
- Best practices
- What will be changed in the future

MS cluster service components (1)

Cluster disk driver

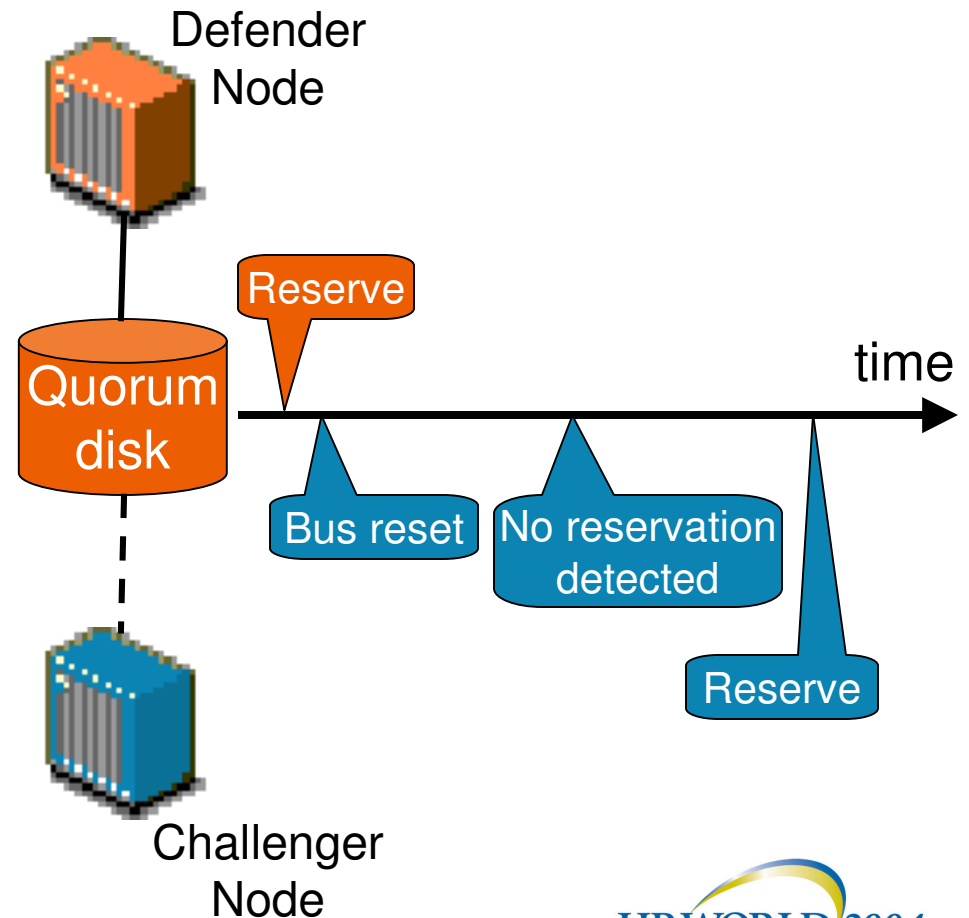
- Is a filter driver
- Coordinates access to shared disks for the nodes in a cluster
- Only one node in the cluster owns a disk at a single point in time



MS cluster service components (2)

Cluster quorum disk

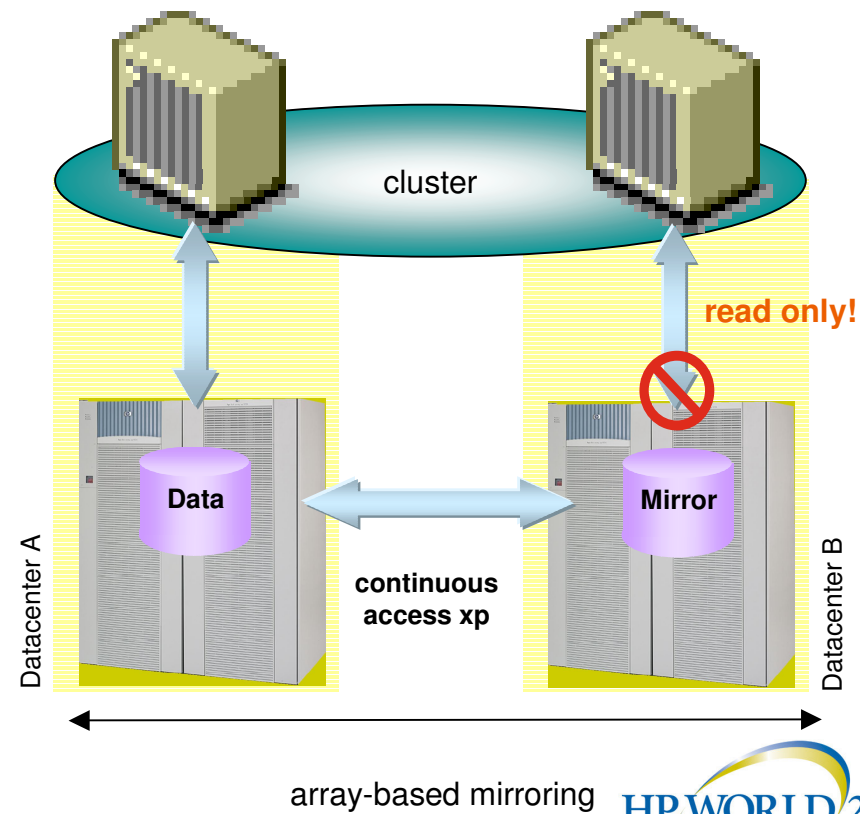
- To prevent split-brain-phenomena when all IP network links fail
- Implemented as a challenge/defense protocol, that uses SCSI Reserve and SCSI Bus Reset commands
- Complex timing and disk access protocol



MS Cluster Service Components (3)

Physical Disk resources – data disks only

- Do not “know” about read-only replicated disks
- CLX resource type necessary to check CA XP disk pairs before they can be read/write enabled (either by swapping or suspending)
- Physical Disk resource needs to depend on CLX resource



Challenges with dispersed MSCS cluster systems



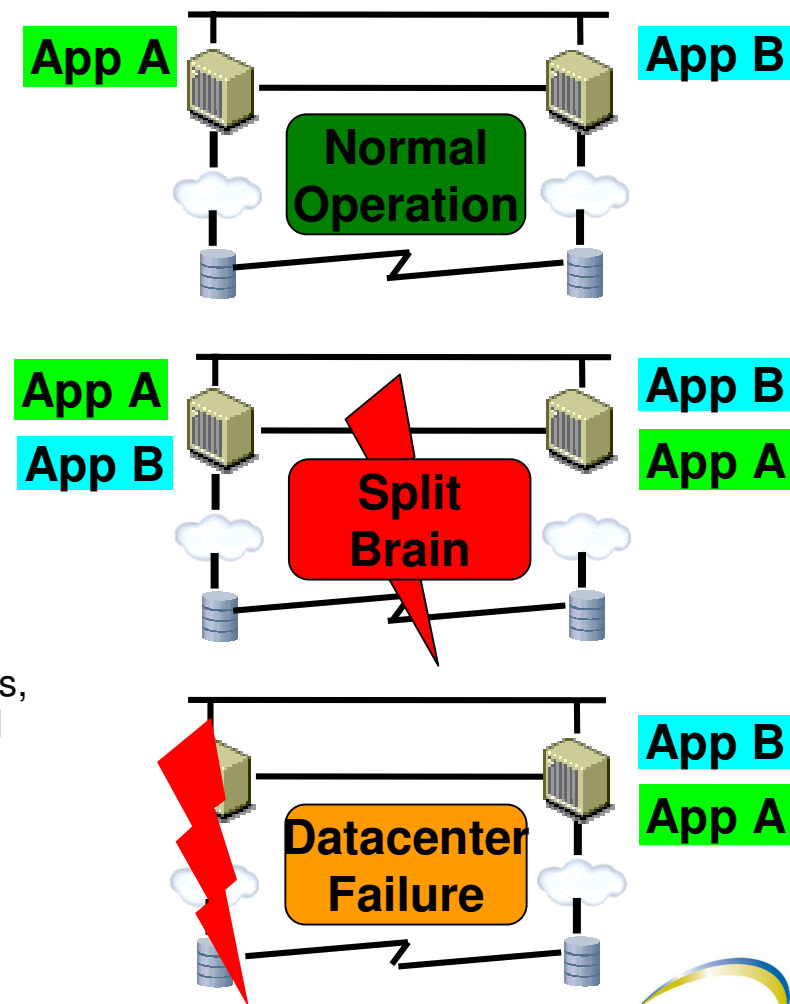
- Mirror **quorum** disk to remote disk array
- Implement quorum disk pair **and keep** challenge/defense protocol working as if it is a single shared resource
- **Filter** SCSI Reserve/Release/Reset and any necessary IO commands without performance impacts
- **Prevent** split-brain phenomena
- **Check data** disk pairs on failover
- Allow data disk failover only if current and consistent

CLX Distributed Quorum Filter Service

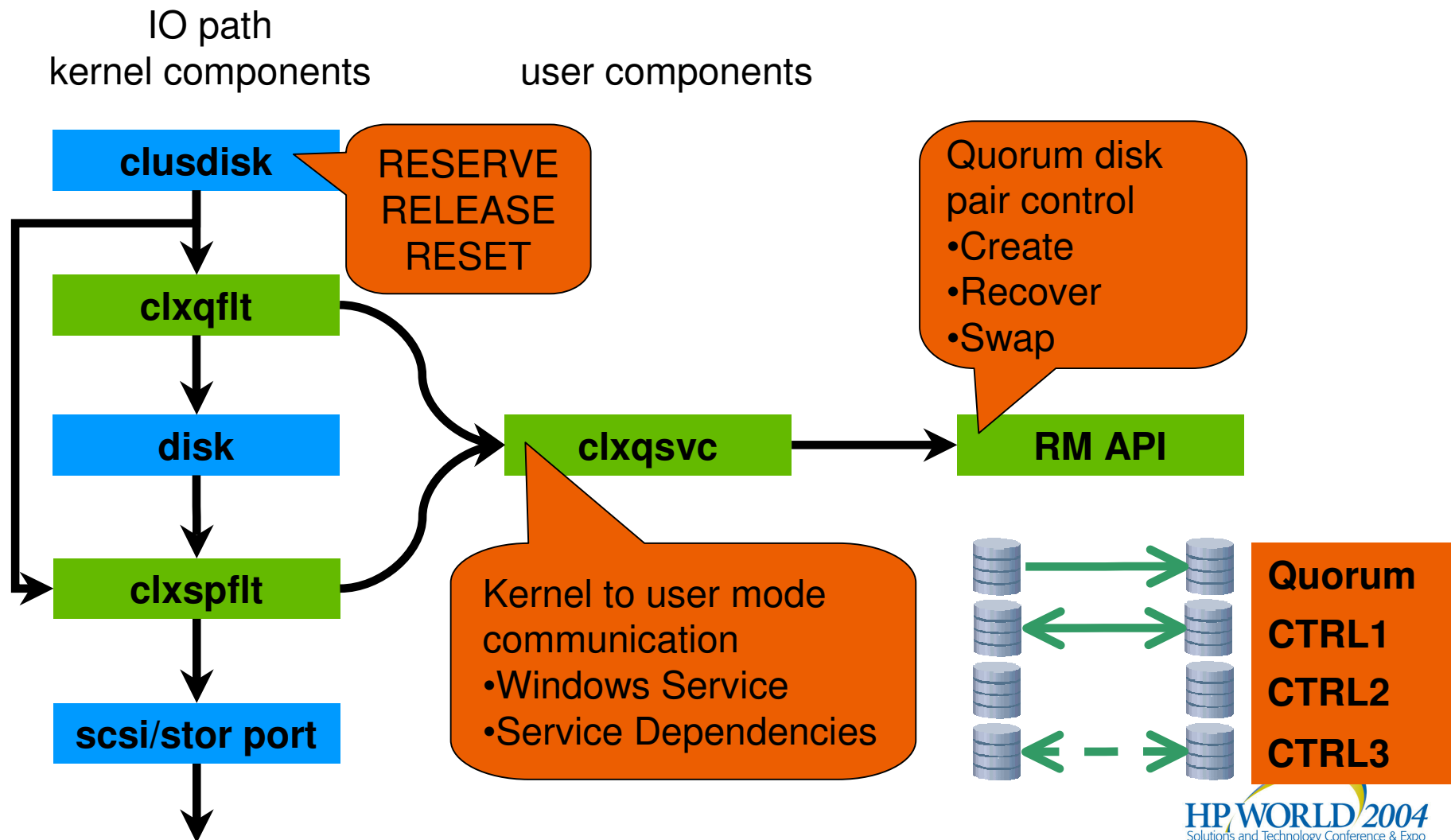


Why is it needed?

- **Prevents** split-brain syndrome
- Handles more failure scenarios in a safe and automatic way (i.e. Datacenter Failure)
- Replicates standard quorum disk transparently - no need for Majority Node Set (MNS)
- Microsoft on MNS: „A majority node set cluster is best used in controlled, targeted scenarios, as part of a cluster solution offered by your Original Equipment Manufacturer (OEM), Independent Software Vendor (ISV), or Independent Hardware Vendor (IHV). The single quorum device server cluster, **the standard cluster model available** in Windows NT and Windows 2000, will still cover the vast majority of your cluster deployment needs.“



CLX Distributed Quorum Filter Service Components



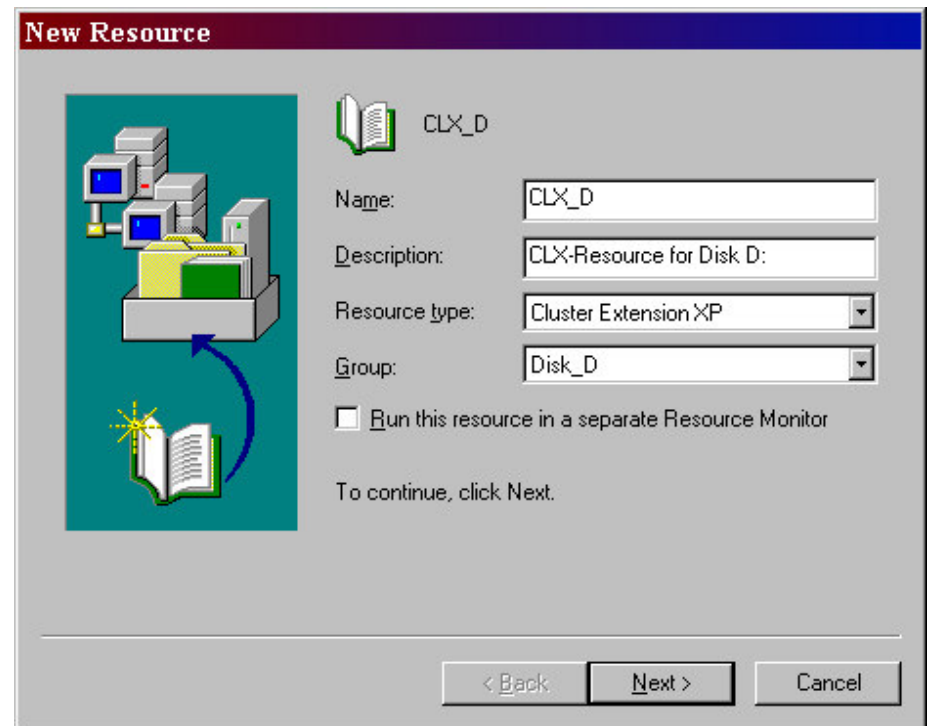
Challenges using MNS

- MNS solely relies on the IP network
 - The MNS quorum is achieved by accessing file shares between the systems.
 - A network/node failure causes the arbitration between nodes to start. MNS relies on the part that might have failed.
 - Quorum disk as alternative path removed.
- You need a 3rd node as part of the cluster.
 - It can't just be any external client. It has to be a cluster member on the same subnet.
- The majority calculation is not dynamic.
 - If you have 6 nodes you can lose 2 before the cluster fails.
 - If you lose one, 5 remain and theoretically you would have to recalculate the majority set. This means with the 5 remaining you could lose another 2 and the cluster would still run. But MS doesn't do that in their MNS implementation. They never recalculate and therefore, in a 6 node cluster you lose 1 node and then you lose 1 more node and the cluster is down.
- With CLX quorum service the cluster would be running until the last node failed!
- CLX quorum disk uses the CA link as alternative medium to IP.
 - One could argue that long distance DR solution would use the same link for CA and for IP. But 90% of our Metrocluster and CLX installations are metro/campus installations where the IP and CA links are separate.

Resource Type „Cluster Extension XP

What does it do?

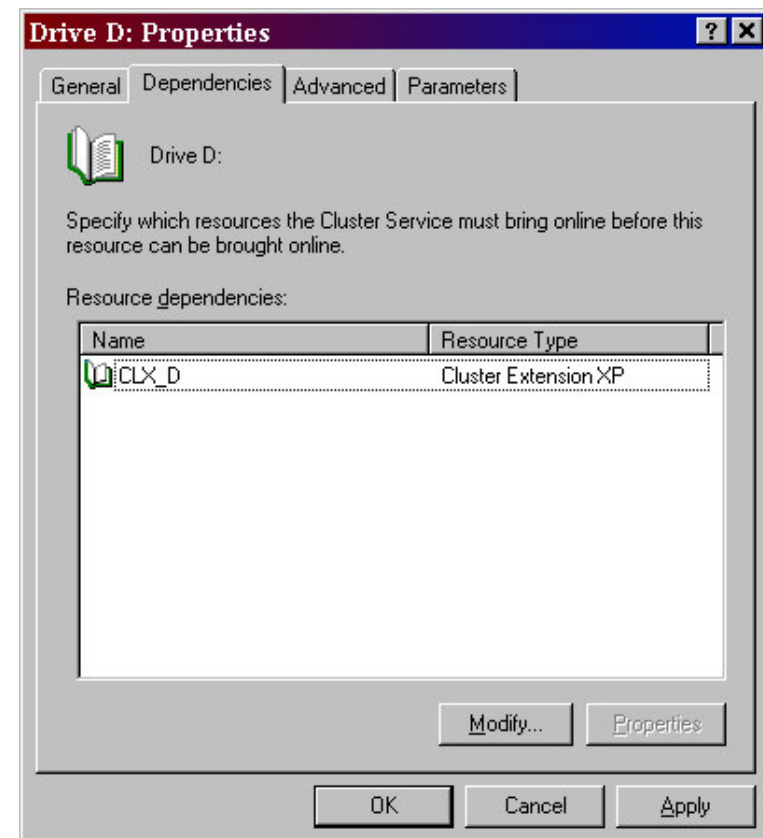
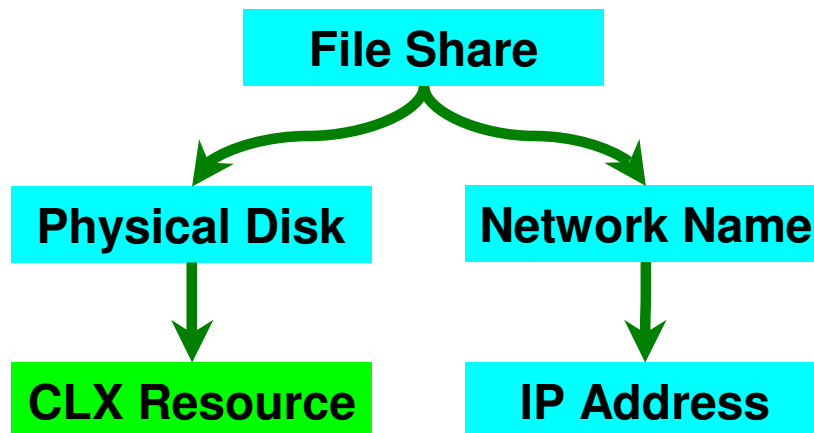
- Responsible for takeover actions of the data/application disk pairs
- Uses Raid Manager XP instance(s)
- Triggers the Pair-/Resync Monitor, if configured
- Double-checks fence level, XP serial number



Resource Dependencies

Integration of CLX Resource Type:

- All „Physical Disk“ resources of a resource group depend on a CLX resource
- Very smooth integration



Cluster Administrator Extension DLL

Parameters

CLX_FILESHARE

XP Serial Numbers: 30060
30080 [Add] [Delete]

RM XP Instances: 101
102 [Add] [Delete]

RM XP Device Group: clxfileshare

CA XP Fence Level: NEVER [Advanced...]

Data Center Node Name List: W2K2 in DC A
W2K3 in DC B
W2K4 in DC C [Add] [Delete]

Application Directory: C:\Program Files\Hewlett-F

[Failover/Failback...] [Pair/ResyncMon...] [Pre/Post Exec...]

[< Back] [Finish] [Cancel]

Advanced Fence Level Failover Behaviour

☐ DATA Lose Mirror

☒ DATA Lose Data Center

ASYNC Takeover Timeout: 3600 seconds

[OK] [Cancel]

Add Node to Data Center List

Available hosts: W2K2
W2K3
W2K4

Data Center Name: A
A
B

[OK] [Cancel]

Failover/Failback Behaviour for CLX_FILESHARE

Application Startup Behaviour:

☒ RESYNCWAIT
RESYNCWAIT Timeout: 300 seconds

☐ FASTFAILBACK

☐ AutoRecover

[OK] [Cancel]

Pair/Resync Monitor Properties for CLX_...

☐ Use Pair/Resync Monitor

☐ Pair/Resync Monitor AutoRecover

Monitor Interval: 60 seconds

[OK] [Cancel]

Pre/Post Executable Properties for CLX_F...

Pre/Post Takeover Executable Paths

PreExecScript: []

PostExecScript: []

☐ Check Remote Disk States after Takeover

[OK] [Cancel]

Agenda

- What will you learn here
- Business Continuity Background
- Cluster Extension XP integration into MSCS
- **What has changed**
 - Quorum service recovery enhancements
 - Parallel resource online performance enhancements
- Best practices
- What will be changed in the future

Changes in 2.04 (I)

- 2.04.00 is the 64-bit Windows release plus CLX resource type enhancement
 - 2.04.00 is latest release for W2k, W2k3 (32-bit and 64-bit), Solaris
 - Includes parallel resource online enhancements.
- Installation program change to use xpinfo-like inquiry as response to a disk discovery issue
 - CLX install program fails to identify Command Device when disks are visible to Windows but not accessible as Windows disks. In this case, disk registry entries are enumerated but not equal to their respective physical disk drive numbers. This incident happened on multi-OS Itanium servers but could also happen if LUN0 (XP512) is mapped to all servers but disabled in Disk Manager. Workaround is to map the Command Devices with Physical Drive numbers that are higher than the number of “invisible” disks.

Changes in 2.04 (2)

- Installation program uses HP-written programs internally because 64-bit InstallShield doesn't provide all 32-bit functions
 - Functions like "GetRemoteRegistryKey" from the InstallShield program are not working/supported. In order to exchange information for the quorum service between nodes we had to write our own 64-bit programs.
- Install program allows independent install of CLX resource type and quorum service
 - All previous versions had independent checkboxes but required quorum disk, command devices and CLX control disks
- Installation program creates control and quorum disks pairs from the last node in the cluster
 - All previous versions required to have all quorum service related disks either being paired correctly or not paired at all.

Secure Path 4.0C-SP1 required

- If you want to use persistent reservations (= IO load balancing in the cluster) in conjunction with CA XP, then you must have SP4.0C-SP1.
- SP4.0C-SP1 fixes issues with reserved/not reserved LUNs on CA XP SVOL devices.
- Use this or later version with CLX.

Agenda

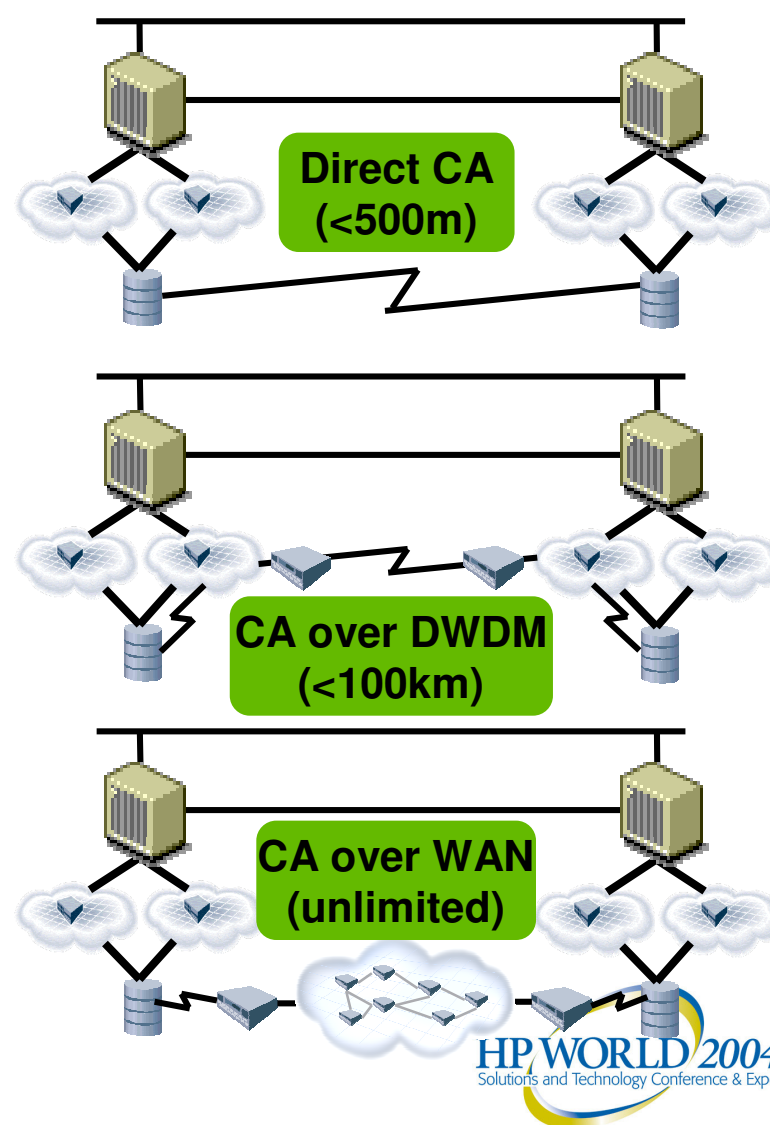
- What will you learn here
- Business Continuity Background
- Cluster Extension XP integration into MSCS
- What has changed
- **Best practices**
 - SAN and LUN Mapping
 - Raid Manager XP Configuration
 - Parameter Settings
- What will be changed in the future

Old (but common) issues and don'ts

- Pair/resync monitor doesn't work with RM instance 0 (fixed in CLX 2.02.00)
- Pre-exec needs to be removed when creating new CLX resource (fixed in CLX 2.03.00)
- Common user mistakes:
 - Cluster Group (quorum disk) has CLX resource configured
 - Quorum service shares command device with all other HP apps (CLX resource, Data Protector ZDB, Performance Advisor)
 - CLX resource name changed without delete/create process

SAN

- Use two independent fabrics and Secure Path software
- All devices including Raid Manager Command Devices should have redundant paths to be protected by Secure Path
- Prefer synchronous CA XP, with async CA XP APM is useful
- If the CA XP connection is done over a less reliable network (i.e. switched Wide Area Network), use a dedicated fabric for CA to avoid fabric reconfigurations



LUN Mapping

host groups(XP128/1024)

- Map all LUNs to CL1-A and CL2-A:
 - LUN0 CMD-Dev QFS (36MB)
 - LUN1 CMD-Dev RMgr (36MB)
 - LUN2 Quorum (100-500MB)
 - LUN3-5 CTRL1-3 (36MB)
 - LUN6-n Shared Data Drives

no host groups(XP48/512)

- Map all LUNs to CL1-A and CL2-A:
 - LUN0 all Svr (Emulex & QLogic)
 - LUN1 CMD-Dev QFS Svr1
 - LUN2 CMD-Dev RMgr Svr1
 - LUN2 Quorum Svr1
 - LUN3-5 CTRL1-3 Svr1
 - LUN6-...Shared Data Drives Svr1
 - LUN20 CMD-Dev QFS Svr2
 - LUN21 CMD-Dev RMgr Svr2
 - LUN22 Quorum Svr2
 - LUN23-25 CTRL1-3 Svr2
 - LUN26-...Shared Data Drives Svr2
 - LUN30 CMD-Dev QFS Svr3
 - LUN31 CMD-Dev RMgr Svr3
 - LUN32 Quorum Svr3
 - LUN33-35 CTRL1-3 Svr3
 - LUN36-...Shared Data Drives Svr3

Configure Control and Raid Manager CMD Devices



- Create a Primary Partition, but do not format Control and RM CMD devices
- Use GUID and \\.\PhysicalDriveX in config file
- Use separate RM CMD Devices for Quorum Filter Service and Raid Manager XP Instances

```
horcm10.conf - Notepad
File Edit Format View Help

HORCM_MON
#ip_address      service      poll(10ms)      timeout(10ms)
NONE            horcm10        1000            1000

HORCM_CMD
#dev_name        dev_name        dev_name
\\.\Volume{0a30881b-b023-11d6-b995-00306e12ea34} \\.\PHYSICALDRIVE2

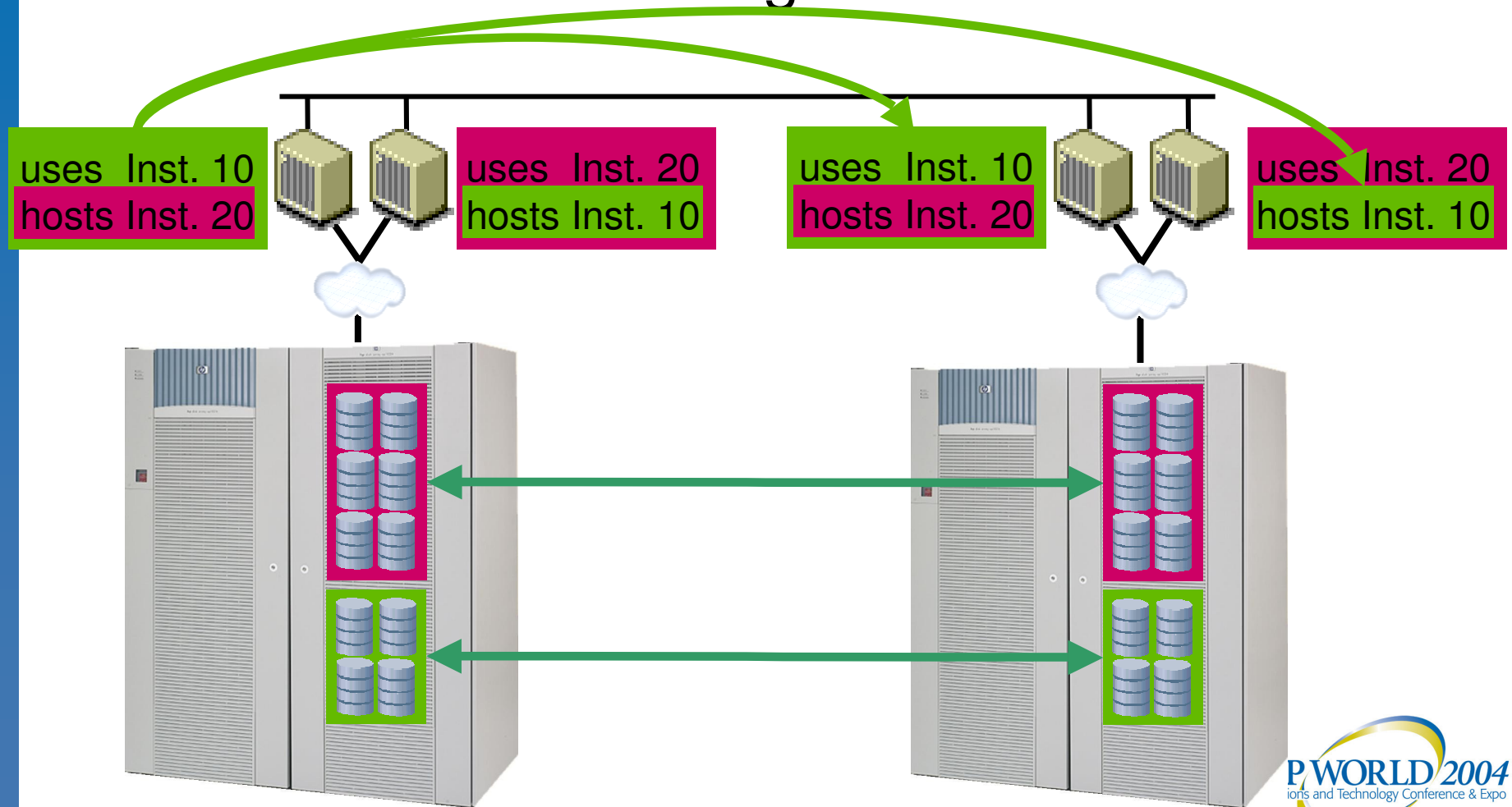
HORCM_DEV
#dev_group       dev_name        port#           TargetID
EXCL3_qfs        100_200         CL1-A          0
EXCL3_qfs        101_201         CL1-A          0
EXCL3_qfs        102_202         CL1-A          0
EXCL3_qfs        103_203         CL1-A          0
EXCL3V1          104_204_s       CL1-A          0
EXCL3V1          105_205_t       CL1-A          0
EXCL3V2          106_206_u       CL1-A          0
EXCL3V2          107_207_v       CL1-A          1
```

```
C:\WINDOWS\System32\cmd.exe
C:\> raidscan -x findcmddev 0,20
\\.\PhysicalDrive1
\\.\PhysicalDrive2
\\.\Volume{35533e66-f2b7-11d6-b9c7-00306e12ea34}
\\.\Volume{0a30881b-b023-11d6-b995-00306e12ea34}

C:\>
```

Raid Manager XP Configuration

- Redundant Raid Manager XP Instances



Naming Raid Manager XP Device Groups and Devices



- Name the **device groups** after the virtual server, i.e.:
-> **EXCL3V1**
- Name the **device** after the P-Vol and S-Vol plus drive letter, i.e.: P-Vol 1:04, S-Vol 1:04, S:\
-> device name: **104_104_s**

horcm10.conf - Notepad

```
File Edit Format View Help
HORCM_MON
#ip_address      service
NONE             horcm10

HORCM_CMD
#dev_name        dev_name
\\.\Volume{0a30881b-b023-11...
```

HORCM_DEV

#dev_group	dev_name	port#	TargetID	LU#	MU#
EXCL3_qfs	100_200	CL1-A	0	1	
EXCL3_qfs	101_201	CL1-A	0	2	
EXCL3_qfs	102_202	CL1-A	0	3	
EXCL3_qfs	103_203	CL1-A	0	4	
EXCL3V1	104_204_s	CL1-A	0	5	
EXCL3V1	105_205_t	CL1-A	0	6	
EXCL3V2	106_206_u	CL1-A	0	7	
EXCL3V2	107_207_v	CL1-A	1	0	

C:\WINDOWS\System32\cmd.exe

```
C:\> pairdisplay -g EXCL3V1_ST -fx -CLI
```

Group	PairVol	L/R	Port#	TID	LU	Seq#	LDEV#	P/S	Status	Fence	Seq#	P-LDEV#	M
EXCL3V1	104_104_s	L	CL1-A	0	5	30064	104	P-VOL	PAIR	NEVER	60024	104	-
EXCL3V1	104_104_s	R	CL1-A	0	5	60024	104	S-VOL	PAIR	NEVER	-	104	-
EXCL3V2	105_105_t	L	CL1-A	0	6	30064	105	P-VOL	PAIR	NEVER	60024	105	-
EXCL3V2	105_105_t	R	CL1-A	0	7	60024	105	S-VOL	PAIR	NEVER	-	105	-

```
C:\>
```

Raid Manager XP Configuration File

```

horcm10.conf - Notepad
File Edit Format View Help

HORCM_MON
#ip_address      service      poll(10ms)      timeout(10ms)
NONE            horcm10          1000            1000

HORCM_CMD
#dev_name        dev_name        dev_name
\\.\Volume{0a30881b-b023-11d6-b995-00306e12ea34} \\.\PHYSICALDRIVE2

HORCM_DEV
#dev_group      dev_name        port#      TargetID      LU#      MU#
EXCL3_qfs       100_200         CL1-A      0              1
EXCL3_qfs       101_201         CL1-A      0              2
EXCL3_qfs       102_202         CL1-A      0              3
EXCL3_qfs       103_203         CL1-A      0              4
EXCL3V1         104_204_s       CL1-A      0              5
EXCL3V1         105_205_t       CL1-A      0              6
EXCL3V2         106_206_u       CL1-A      0              7
EXCL3V2         107_207_v       CL1-A      1              0

HORCM_INST
#dev_group      ip_address      service
#Client LAN to EXCL3N2
EXCL3_qfs       172.16.4.10     horcm10
EXCL3V1         172.16.4.10     horcm10
EXCL3V2         172.16.4.10     horcm10
#Private LAN to EXCL3N2
EXCL3_qfs       10.16.4.10      horcm10
EXCL3V1         10.16.4.10      horcm10
EXCL3V2         10.16.4.10      horcm10
#Client LAN to EXCL4N2
EXCL3_qfs       172.16.4.20     horcm10
EXCL3V1         172.16.4.20     horcm10
EXCL3V2         172.16.4.20     horcm10
#Private LAN to EXCL4N2
EXCL3_qfs       10.16.4.20      horcm10
EXCL3V1         10.16.4.20      horcm10
EXCL3V2         10.16.4.20      horcm10
#last line
  
```

CLX Resource Parameter Settings (1)



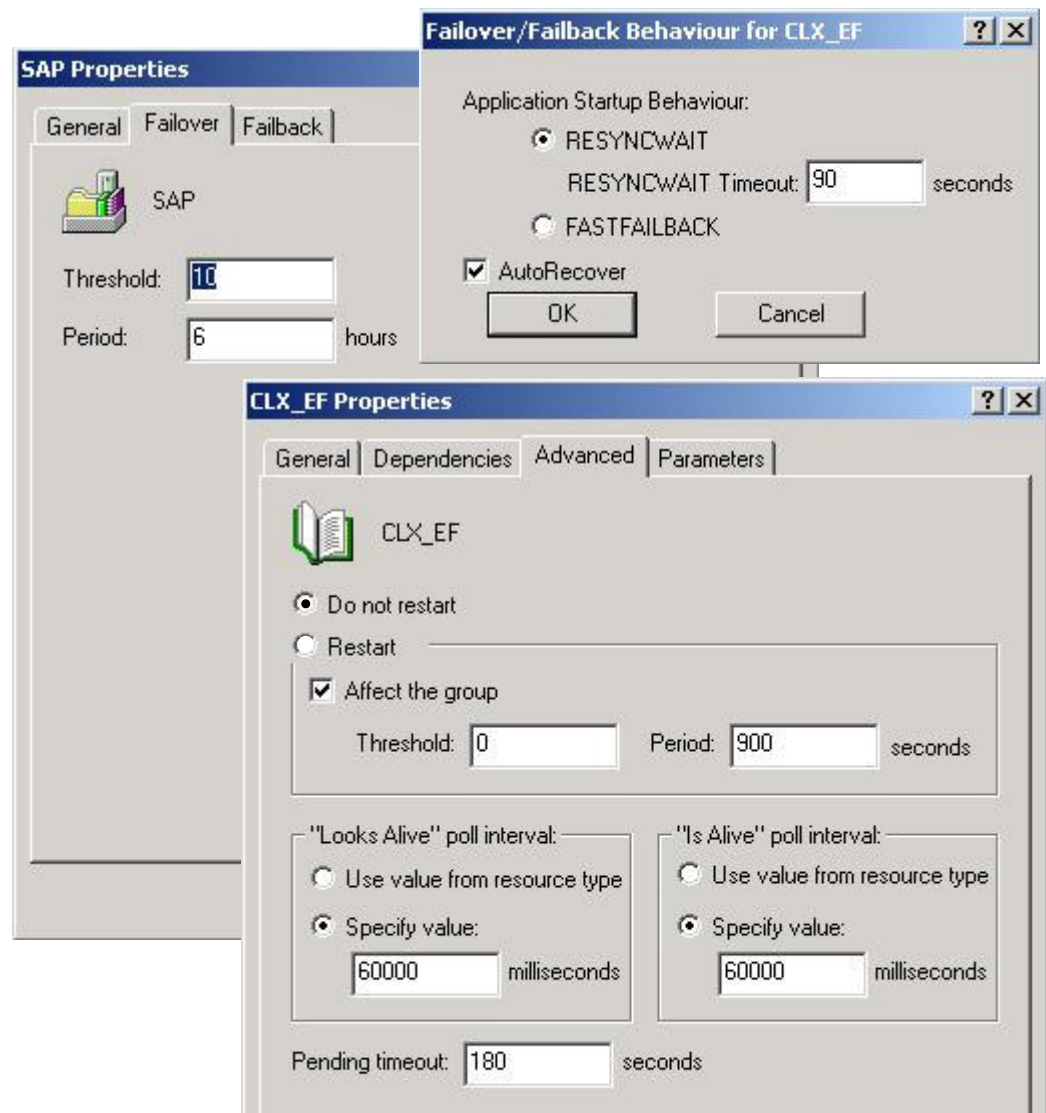
- For higher availability use fence level NEVER
- For guaranteed data consistency and currency over the two sites use fence level DATA
- Only to avoid latency issues, use fence level ASYNC
- In two-node clusters use „Resync Wait“
- Use the „Pair/Resync Monitor“



CLX Resource Parameter Settings (2)



- Use „AutoRecover“ with care
- „Failover Period“ of group and „Pending Timeout“ of CLX resource must be greater than the „ResyncWait Timeout“



Agenda

- What will you learn here
- Business Continuity Background
- Cluster Extension XP integration into MSCS
- What has changed
- Best practices
- What will be changed in the future

Futures

- Plans to enhance the install program
 - Get rid of too many windows.
- License keys required!
 - We will include AutoPass.

HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:



RECOMMENDED TRAINING VENUE FOR THE
HP Certified Professional

