


DECUS Munich Symposium 2004

Long-Distance Configurations for MSCS with IBM Enterprise Storage Server

Torsten Rothenwaldt
IBM Germany

Presentation 2E06

© 2004 IBM Corporation



DECUS Munich Symposium 2004

Agenda

- Cluster problems to resolve in long-distance configurations
- Stretched MSCS with remote copy
- IBM GDS solution for multi-site MSCS with ESS PPRC:
 - Design
 - Installation and configuration
 - Operation and fault scenarios
- Project considerations:
 - Certification and support
 - Typical issues

2 | 2E06 | Windows SIG

© 2004 IBM Corporation



DECUS Munich Symposium 2004 

Table of contents

- **Problems to solve**
 - Stretched MSCS with PPRC
 - ESS Geographically Dispersed Sites (GDS) for MSCS
 - Project considerations

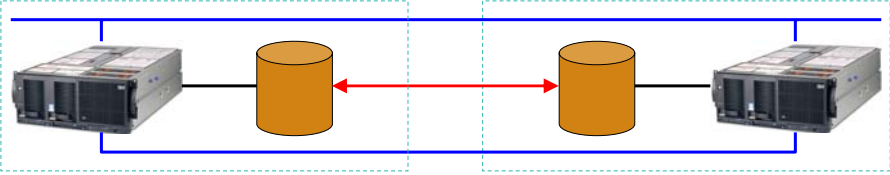
3 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 


Cluster problems to solve (1): Data transfer

- a) **Mirroring:**
 - One volume with two or more equal plexes, accessed symmetrically.
 - Data transfer is bidirectional.
- b) **Remote copy:**
 - Two volumes with primary/secondary roles, access only to primary.
 - Data transfer is unidirectional.
- c) **Replication:**
 - Two volumes, roles depending on implementation.
 - Changes intercepted and transferred by non-SCSI protocol, uni- or bidirectional.

How to integrate MSCS resource management and data operations?



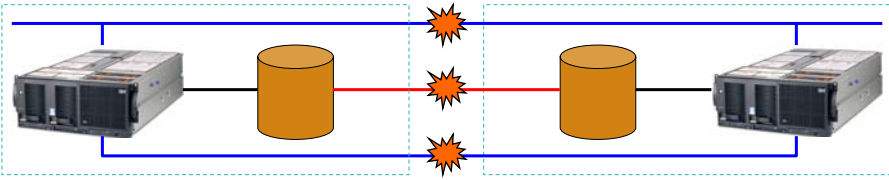
4 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 


Cluster problems to solve (2): Integrity

- **Transactional integrity:**
 - 1-safe versus 2-safe
 - Volume-level consistency: none vs. power-failure versus application stop
 - Single-volume consistency versus multi-volume consistency
 - What happens when data transfer becomes interrupted / is reestablished?
 - Safe mode versus unsafe mode
 - Full copy versus changelog versus bitmap
- **Cluster integrity:**
 - Partitioning in space and time

How to ensure MSCS quorum disk consistency and semantics?




5 | 2E06 | Windows SIG | © 2004 IBM Corporation


DECUS Munich Symposium 2004 

Peer-to-Peer Remote Copy with IBM ESS

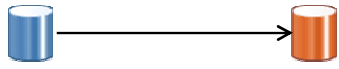
- **Dynamically balanced paths via mixed-traffic ports (FC or ESCON)**
- **Microcode feature with GUI, CLI and API support**
- **Characteristics:**
 - Synchronous or asynchronous, direct or cascading continuously
 - Can be combined with Point-in-Time copy from primary and secondary volume
 - Change recording in bitmaps ("Suspend" and "Copy changed tracks" operations)
 - Multi-volume consistency groups (with wait phase)
 - Rich spectrum of operation options (for example: "Allow read from secondary", "PPRC Failover", "PPRC Failback")
- **See for details:**
 - IBM Redbook SG24-5757 www.redbooks.ibm.com
 - IBM Systems Journal Vol. 42 Nr. 2 www.research.ibm.com/journal




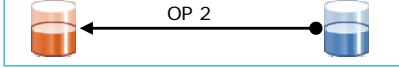
6 | 2E06 | Windows SIG | © 2004 IBM Corporation

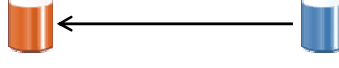
DECUS Munich Symposium 2004 

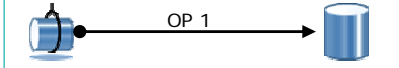
Details of failover/failback operations

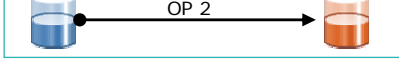
Normal operation (duplex) 

Failover phase 1 (reverse) 

Failover phase 2 (synchronize) 

Reverse operation (duplex) 

Failback phase 1 (reverse) 

Failback phase 2 (synchronize) 

7 | 2E06 | Windows SIG | © 2004 IBM Corporation



DECUS Munich Symposium 2004 

Table of contents

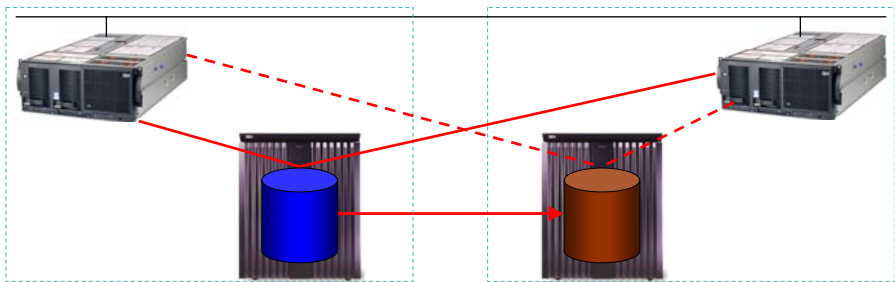
- Problems to solve
- **Stretched MSCS with PPRC**
- ESS Geographically Dispersed Sites (GDS) for MSCS
- Project considerations

8 | 2E06 | Windows SIG | © 2004 IBM Corporation


DECUS Munich Symposium 2004 

Stretched MSCS with PPRC: Principle

- Both cluster nodes access only the PPRC source volumes.
- Node outages and other non-storage failures handled by MSCS.
- Storage subsystem or PPRC failure may require manual procedure.
- Failure detection, consensus, partitioning like local MSCS.
- Transactional consistency according to PPRC details.

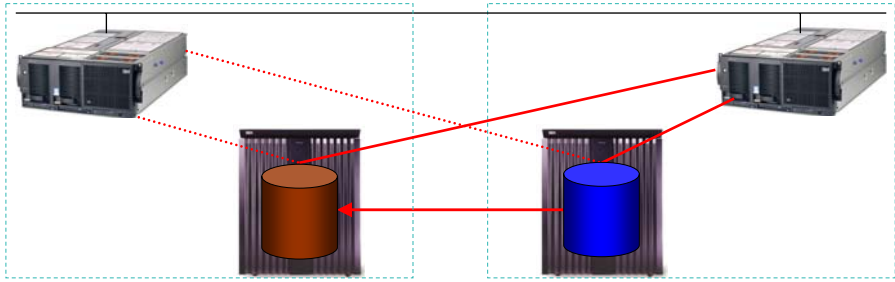


9 | 2E06 | Windows SIG | © 2004 IBM Corporation


DECUS Munich Symposium 2004 

Stretched MSCS with PPRC: Copy failover

- Shut down cluster node(s).
- Check PPRC and assignment status of secondary volumes.
- If possible, unassign primary volumes from cluster nodes.
- Reverse copy direction.
- Assign former secondary volumes to cluster nodes.
- Boot cluster node(s), check resources.
- If possible, synchronize reverse copy.



10 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Stretched MSCS with PPRC: Copy failback

1. Check situation. If unclear, unplug FC cables from former primary storage system.
2. Start former primary storage system.
3. Perform "cleanup operations" at former primaries.
4. Boot remaining cluster node(s), watch cluster join.
5. Establish and synchronize reverse copy.

Do you really want to perform a failback now?

6. Shut down cluster nodes.
7. Unassign current primary volumes from cluster nodes.
8. Reverse copy direction.
Terminate → Establish reversed → Suspend
9. Assign original volumes to cluster nodes.
10. Boot cluster nodes and check resources.
11. Synchronize new copies.

11 | 2E06 | Windows SIG | © 2004 IBM Corporation



DECUS Munich Symposium 2004 

Table of contents

- Problems to solve
- Stretched MSCS with PPRC
- ESS Geographically Dispersed Sites (GDS) for MSCS**
- Project considerations

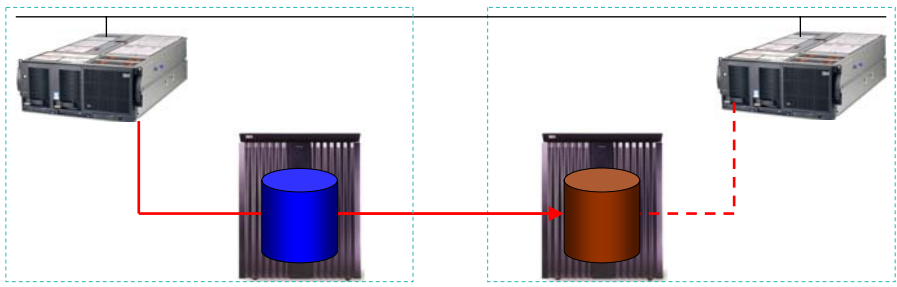
12 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 


ESS GDS for MSCS: Principle

- Additional PPRC cluster resource type
- The resource owner accesses the local volumes as PPRC source.
- All outages handled by combination of classical MSCS and additional software/firmware.

- Failure detection, consensus, partitioning depends on PPRC details.
- Transactional consistency according to PPRC details.

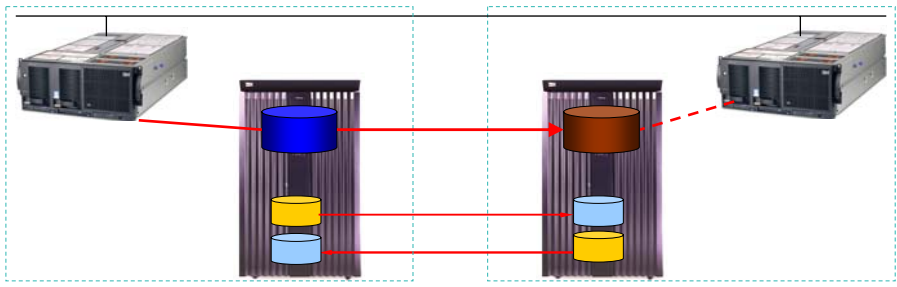


13 | 2E06 | Windows SIG | © 2004 IBM Corporation


DECUS Munich Symposium 2004 

ESS GDS for MSCS : Design

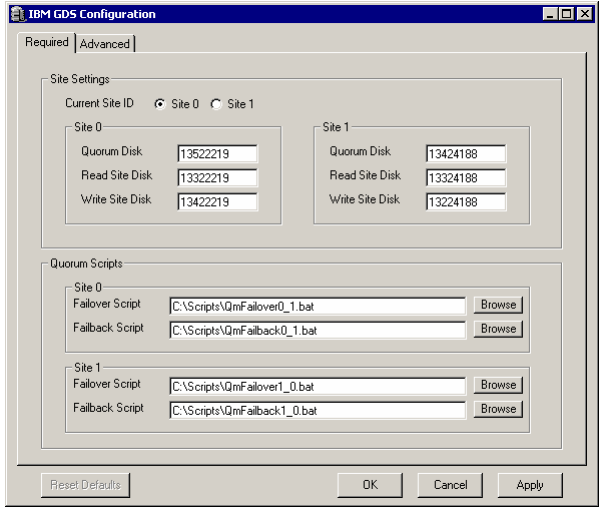
- Non-quorum "Physical Disk" depends on "IBM_PPRC" resource
- Communication between "IBM_PPRC" DLL and ESS via scripts
- Two pairs of additional site disks for PPRC heartbeat
- Software modules:
 - "IBM_PPRC" cluster resource DLL, extension DLL,
 - GDS Failover service, SDD Server service,
 - GDS Configuration utility



14 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

GDS configuration details (1)



IBM GDS Configuration

Required | **Advanced**

Site Settings

Current Site ID Site 0 Site 1


Site 0	Site 1
Quorum Disk: 13522219	Quorum Disk: 13424188
Read Site Disk: 13322219	Read Site Disk: 13324188
Write Site Disk: 13422219	Write Site Disk: 13224188

Quorum Scripts

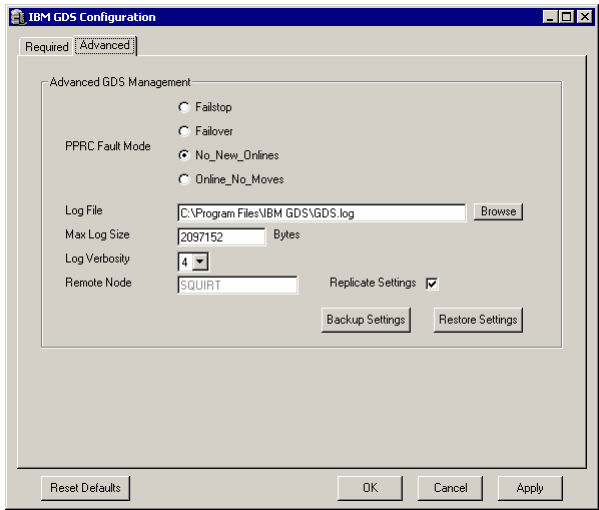
Site 0	Site 1
Failover Script: C:\Scripts\Qm\Falover0_1.bat	Failover Script: C:\Scripts\Qm\Falover1_0.bat
Failback Script: C:\Scripts\Qm\Failback0_1.bat	Failback Script: C:\Scripts\Qm\Failback1_0.bat

Buttons: Reset Defaults, OK, Cancel, Apply

15 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

GDS configuration details (2)



IBM GDS Configuration

Required | **Advanced**

Advanced GDS Management

PPRC Fault Mode

- Failstop
- Failover
- No_New_Onlines
- Online_No_Moves

Log File: C:\Program Files\IBM\GDS\GDS.log

Max Log Size: 2097152 Bytes


Log Verbosity: 4

Remote Node: SQUIRT

Replicate Settings:

Buttons: Backup Settings, Restore Settings, Reset Defaults, OK, Cancel, Apply

16 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

GDS configuration details (3)

Parameters

H_PPRC

Disk.SerialNumberSite0: 02224188

Disk.SerialNumberSite1: 426FCA28


EPRCFailoverScript0: C:\Scripts\HFFailover0_1.bat

EPRCFailbackScript0: C:\Scripts\HFfailback0_1.bat

EPRCFailoverScript1: C:\Scripts\HFFailover1_0.bat

EPRCFailbackScript1: C:\Scripts\HFfailback1_0.bat

17 | 2E06 | Windows SIG | © 2004 IBM Corporation

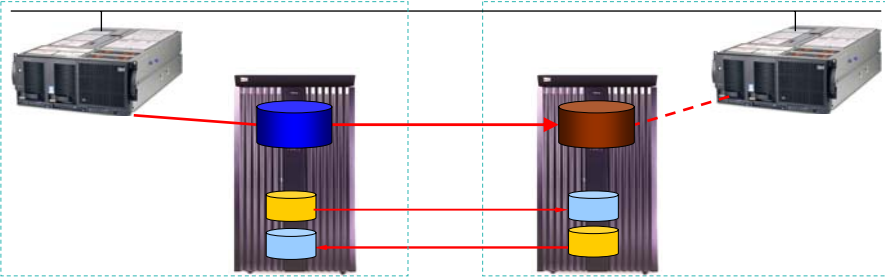
DECUS Munich Symposium 2004 

GDS operation

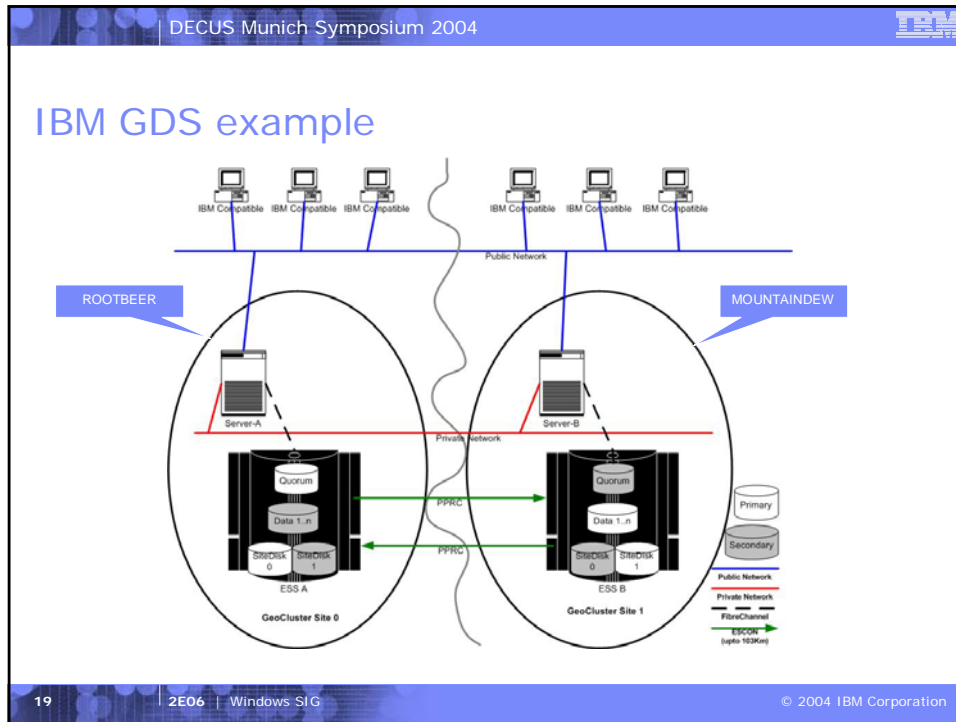
- 1) Server/application/network failure → MSCS reaction + reverse copy
- 2) ESS failure → MSCS reaction → PPRC failover
- 3) PPRC link failure → ? "PPRC Fault Mode" setting (data consistency!)
- 4) Site failure or partitioning → ? "PPRC Fault Mode" setting (split!)


"PPRC Fault Mode":

(a) Failstop	(c) Failover
(b) No New Onlines (default)	(d) No Moves

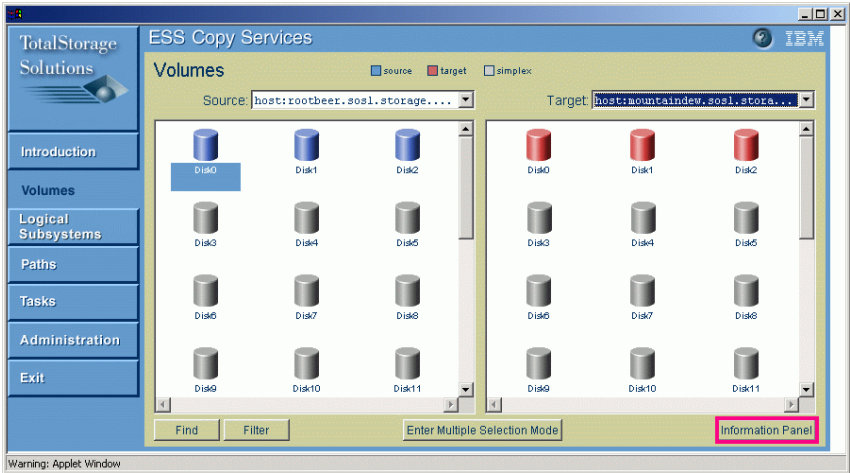


18 | 2E06 | Windows SIG | © 2004 IBM Corporation



DECUS Munich Symposium 2004 


PPRC primary and secondary disks



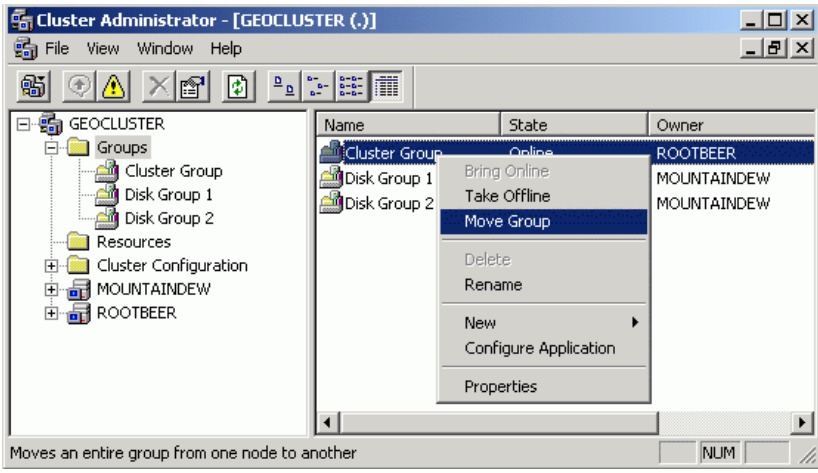
The screenshot shows the 'ESS Copy Services' interface. On the left is a navigation pane with options: Introduction, Volumes, Logical Subsystems, Paths, Tasks, Administration, and Exit. The main area is titled 'Volumes' and has radio buttons for 'source' (selected), 'target', and 'simplex'. Below are two columns of disk icons labeled Disk0 through Disk11. The 'Source' column has Disk0 selected. The 'Target' column has Disk0, Disk1, and Disk2 selected. At the bottom right, an 'Information Panel' is highlighted with a red box.

Warning: Applet Window

21 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Initiate move group on ROOTBEER



The screenshot shows the 'Cluster Administrator' window for 'GEOCLUSTER (.)'. The left pane shows a tree view with 'Groups' expanded to show 'Cluster Group', 'Disk Group 1', and 'Disk Group 2'. The right pane shows a table of cluster groups:

Name	State	Owner
Cluster Group	Online	ROOTBEER
Disk Group 1		MOUNTAINDEW
Disk Group 2		MOUNTAINDEW

A context menu is open over 'Disk Group 1', with 'Move Group' selected. Other menu items include 'Bring Online', 'Take Offline', 'Delete', 'Rename', 'New', 'Configure Application', and 'Properties'. A tooltip at the bottom reads: 'Moves an entire group from one node to another'.

22 | 2E06 | Windows SIG | © 2004 IBM Corporation


DECUS Munich Symposium 2004 

PPRC primary and secondary reversed

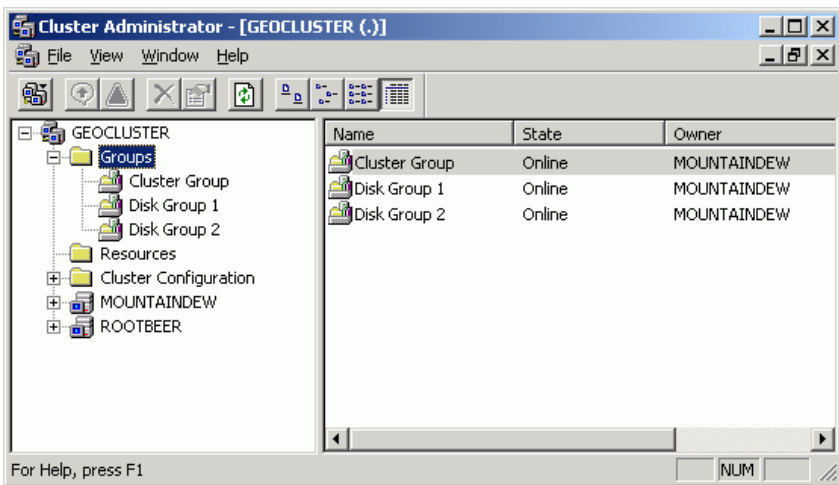


The screenshot shows the 'ESS Copy Services' window. On the left is a navigation pane with options: Introduction, Volumes, Logical Subsystems, Paths, Tasks, Administration, and Exit. The main area is titled 'Volumes' and has two columns: 'Source' and 'Target'. The Source column contains 12 red disk icons labeled Disk0 through Disk11. The Target column contains 12 blue disk icons labeled Disk0 through Disk11. Below the disks are buttons for 'Find', 'Filter', 'Enter Multiple Selection Mode', and 'Information Panel'. A warning message 'Warning: Applet Window' is visible at the bottom left of the window.

23 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

All groups now owned by MOUNTAINDEW




The screenshot shows the 'Cluster Administrator - [GEOCLUSTER (.)]' window. The left pane shows a tree view with 'Groups' expanded, containing 'Cluster Group', 'Disk Group 1', and 'Disk Group 2'. The right pane shows a table of these groups.

Name	State	Owner
Cluster Group	Online	MOUNTAINDEW
Disk Group 1	Online	MOUNTAINDEW
Disk Group 2	Online	MOUNTAINDEW

For Help, press F1 | NUM

24 | 2E06 | Windows SIG | © 2004 IBM Corporation

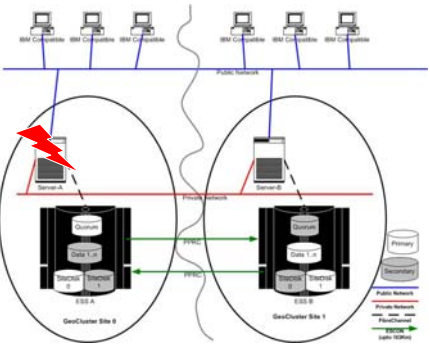
DECUS Munich Symposium 2004 

Fault scenario 1


Server/Application Fault or
Total Network Communication Fault

The remaining node arbitrates for the quorum resource and failover happens reversing PPRC.

When the failed server is repaired, resources may fail back to the preferred owner based on the user defined group failover policy.



25 | 2E06 | Windows SIG | © 2004 IBM Corporation

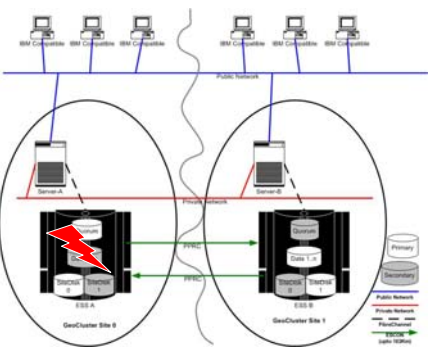
DECUS Munich Symposium 2004 

Fault scenario 2


ESS Fault

The node at the site with the functioning ESS will bring resources online, reversing PPRC if necessary.

When the ESS is repaired, PPRC is re-established and the resources will failback to their preferred owner.

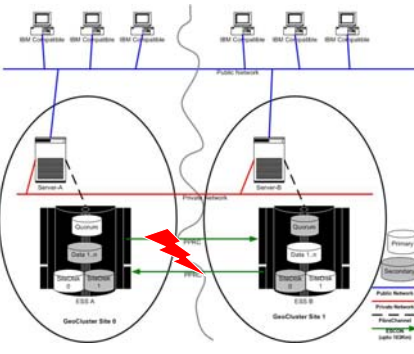


26 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 


Fault scenario 3

Storage Communication Fault



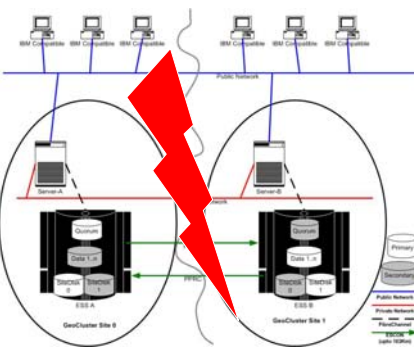
- ❖ PPRC Fault Mode = **Failstop**
All resources taken offline, MSCS on both nodes is shut down.
- ❖ PPRC Fault Mode = **No New Onlines**
Quorum owning node stays online. When PPRC link comes up changes are replicated.
- ❖ PPRC Fault Mode = **Failover**
Quorum owning node stays online and all resources are failed over to the quorum owning node.
- ❖ PPRC Fault Mode = **No Moves**
Both nodes stay online, no resources moved, no moves allowed.

27 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Fault scenario 4

Site Fault or Total Communication Fault



- ❖ PPRC Fault Mode = **Failstop**
All resources taken offline, MSCS on both nodes is shut down.
- ❖ PPRC Fault Mode = **No New Onlines**
Quorum owning node stays online. When PPRC link comes up changes are replicated.
- ❖ PPRC Fault Mode = **Failover**
Quorum owning node stays online and all resources are failed over to the quorum owning node.
- ❖ PPRC Fault Mode = **No Moves**
Both nodes stay online, no resources moved, no moves allowed.

28 | 2E06 | Windows SIG | © 2004 IBM Corporation



DECUS Munich Symposium 2004 

Table of contents

- Problems to solve
- Stretched MSCS with PPRC
- GeoCluster with PPRC
- Project considerations**


29 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Microsoft certification rules: Basic MSCS

- Up to Win2000: The combination "2 servers + HBAs + storage" must be MSCS-certified as whole configuration.
 - Some freedom with respect to CPU speed, memory size, cache size
 - IP components via default Network HCL
- Win2003: Configuration can be combined from building blocks which must be MSCS-certified separately.
 - Server block = single server + HBA
 - Storage block = storage solution + interconnect HBA
 - (in SAN: must also be qualified for Cluster/Multi-cluster device test)
 - A cluster server block is compatible with a cluster storage block if the HBAs, HBA firmware versions, number of HBAs, and HBA driver versions are all identical.
 - IP components via default Network HCL


30 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Microsoft definition of Geo-MSCS

- A Geographically Dispersed Cluster has following attributes:
 - Multiple storage arrays, at least one deployed at each site.
 - Nodes are connected to storage in such a way that in the event of a failure of a site or the communication links between sites, the nodes on a given site can access the storage on that site.
 - The storage fabric or host-based software provides a way to mirror or replicate data between the sites so that each site has a copy of the data.
- General requirements:
 - At least two networks (appear as single LANs on the same IP subnets)
 - Network round-trip delay at most 500 ms.
 - Shared disks provide/emulate SCSI-2 Reserve/Release semantics.
 - Quorum disk replicated in synchronous mode across all sites.


31 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Certification differences for GeoClusters

- No building block certification --- both the hardware and software configuration of a geographically dispersed server cluster must be validated.
- Configuration changes requiring refresh test:
 - Change VLAN technology
 - Update data mirroring software (if host-based software mirror)
- Configuration changes requiring full retest:
 - Change inter-site interconnect technology
 - Change quorum arbitration mechanism


32 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Keys to project success

- The customer has analyzed the business needs (with respect to the tiers of availability) and decided for a cluster-based solution.
- The executives are strong supporters of the project.
- There is a committed budget (including long-term maintenance costs) and an appropriate schedule.
- The customer has some basic experience with MSCS and SAN technology.
- The customer's technical staff is willing to undergo a deep technical training with regular drills.

33 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Notices and Disclaimers

Copyright © 2003 by International Business Machines Corporation.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product and price data have been reviewed for accuracy as of the date of initial publication. Product and price data are subject to change without notice. This information could include technical inaccuracies or typographical errors.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead. It is the user's responsibility to evaluate and verify the operation of any non-IBM product, program or service.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR INFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

IN NO EVENT SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY.


Performance data for IBM and non-IBM products contained in this document was derived under specific operating and environmental conditions. The actual results obtained by any party implementing and sub product will depend on a large number of factors specific to such party's operating environment and may vary significantly. IBM makes no representation that these results can be expected in any implementation of any such product. Accordingly, IBM does not provide any representations, assurances, guarantees or warranties regarding performance.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

34 | 2E06 | Windows SIG | © 2004 IBM Corporation

DECUS Munich Symposium 2004 

Trademarks

The following terms are trademarks or registered trademarks of the IBM Corporation in either the United States, other countries or both.

IBM, z/OS, S/390, AIX, FICON, ESCON, TotalStorage, Enterprise Storage Server, iSeries, pSeries, xSeries, zSeries

Windows NT is a registered trademark of Microsoft Corporation.

Other company, product, and service names mentioned may be trademarks or registered trademarks of their respective companies.

35 | 2E06 | Windows SIG © 2004 IBM Corporation