Efficient Data Path Object Migration for a Highly Available, Scalable IP SAN

Robert E. Gilligan, Kadir Ozdemir,

Ismail Dalgic, Peter Wang

Intransa, Inc.

http://www.intransa.com

Robert.gilligan@ieee.org

Introduction

- Overview of the architecture of the Intransa IP SAN system
- Applies object migration technology
- Leverages the redirection feature of iSCSI
- Delivers an IP SAN that is both scalable and highly available

Goals and Objectives

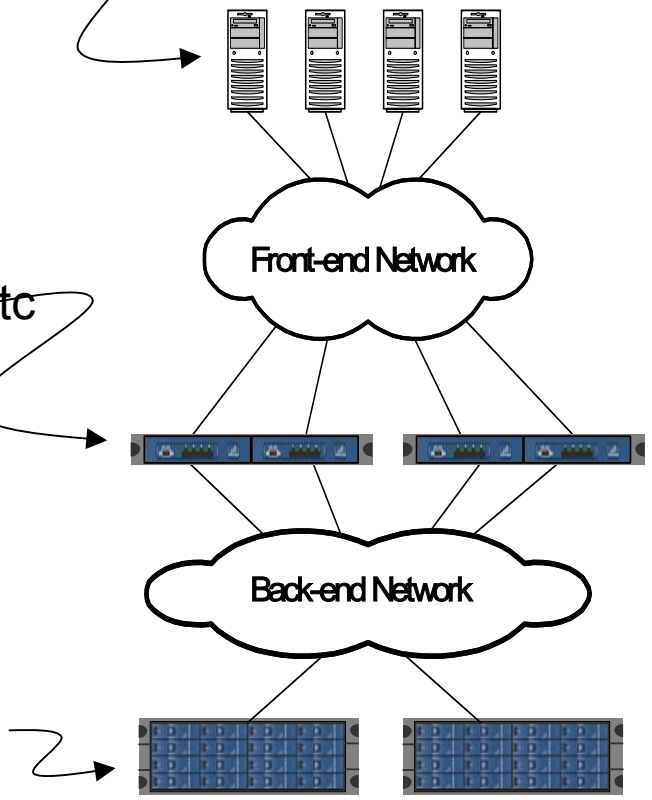
- IP SAN – iSCSI storage service
  - Virtualization and storage
  - iSCSI only – no NAS of FC at this time
- Highly scalable in two dimensions
  - Scalable performance: Able to add controller units to increase throughput and IOPS
  - Scalable storage capacity: Able to add disk arrays to increase storage capacity
- Highly available
  - Robustness in the face of single-unit faults (e.g. controller failure, link failures, etc.)
  - Fast recovery time after faults

System Architecture - Elements

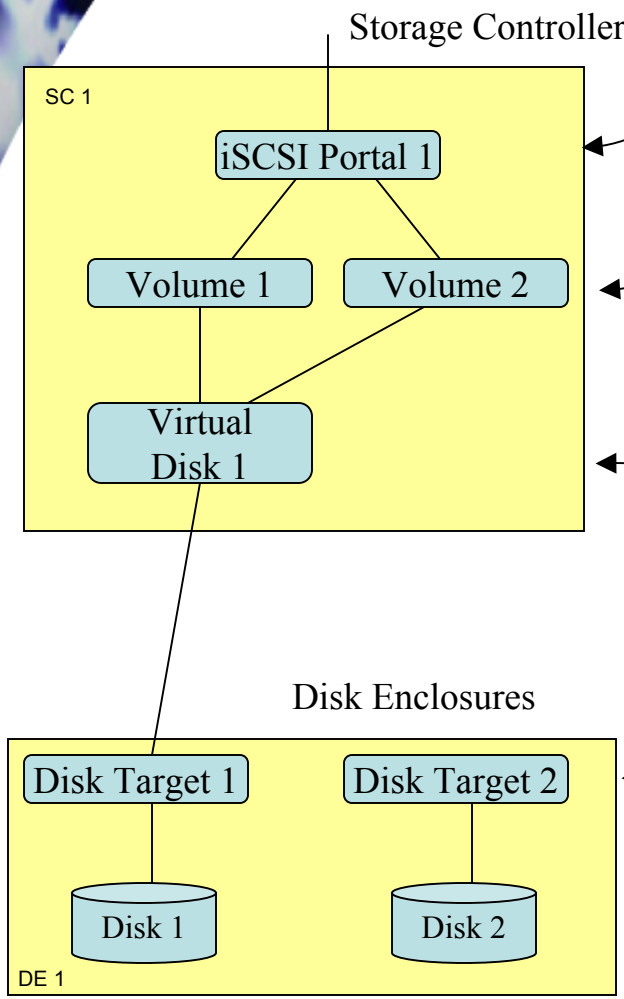
- *Application Hosts* - Speak iSCSI to storage controllers over the front-end network

- *Storage Controllers* - Provide virtualization services - volumes, cache, RAID, snapshot, replication, etc
- Speak iSCSI or tBlock to other storage controllers over back-end network
- Speak iSCSI or xBlock to Disk Enclosures over back-end network

- *Disk Enclosures* - Hold IP-Addressable disks
- Each disk accessible by all storage controllers over back-end network

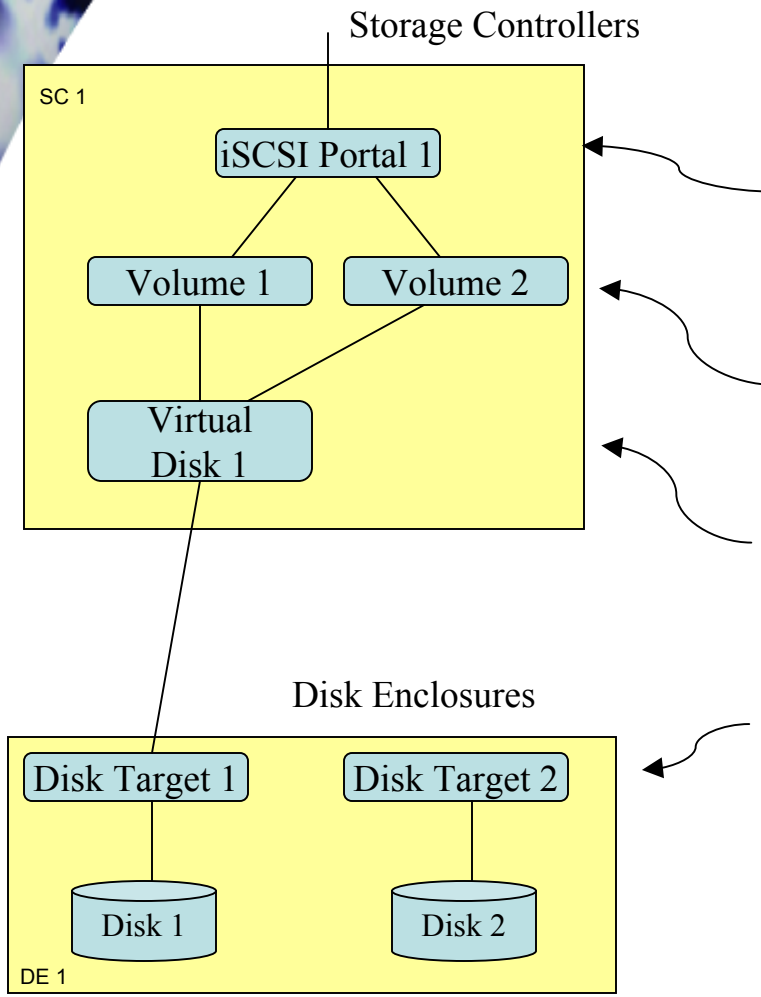


Storage Stack Layer Functions



- iSCSI Portal Layer:
  - Endpoint of iSCSI connection
  - 1 Per controller
- Volume Layer:
  - Virtualizes user volume (LUN) out of extents from one or more Virtual Disk
  - Provides Cache and RAID 0, 1, 10
  - Many per controller
- Virtual Disk Layer:
  - Virtualizes whole physical disks
  - Provides RAID 0, 1, 10
  - Many per controller
- Disk Target Layer
  - Endpoint of communication for one physical disk
  - 1 Per disk

Storage Stack Object Migration

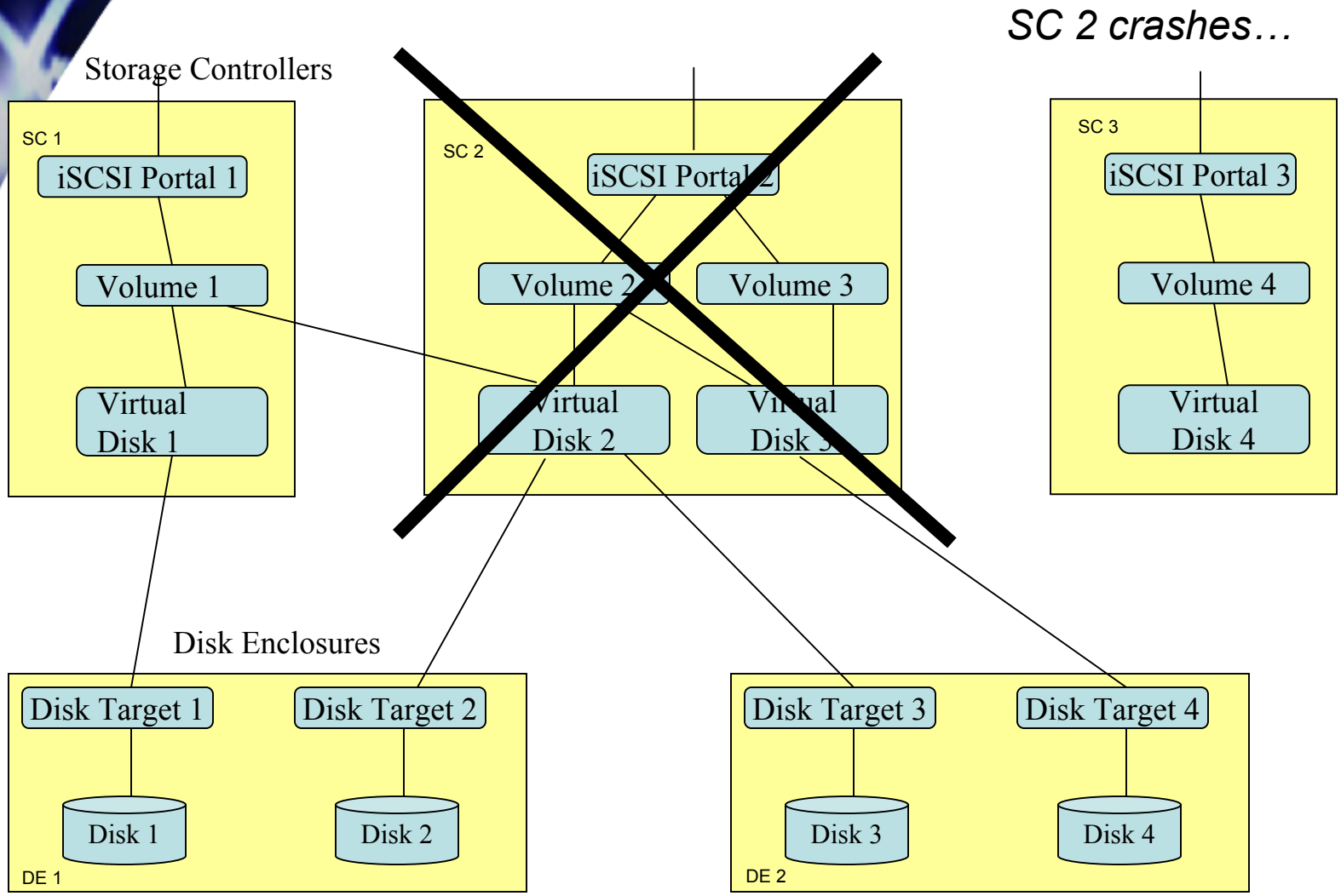


- Only two of the four layers are capable of migrating:
- iSCSI Portal Layer:
  - *Never Migrates*
- Volume Layer:
  - *Migrates*
- Virtual Disk Layer:
  - *Migrates*
- Disk Target Layer:
  - *Never Migrates*

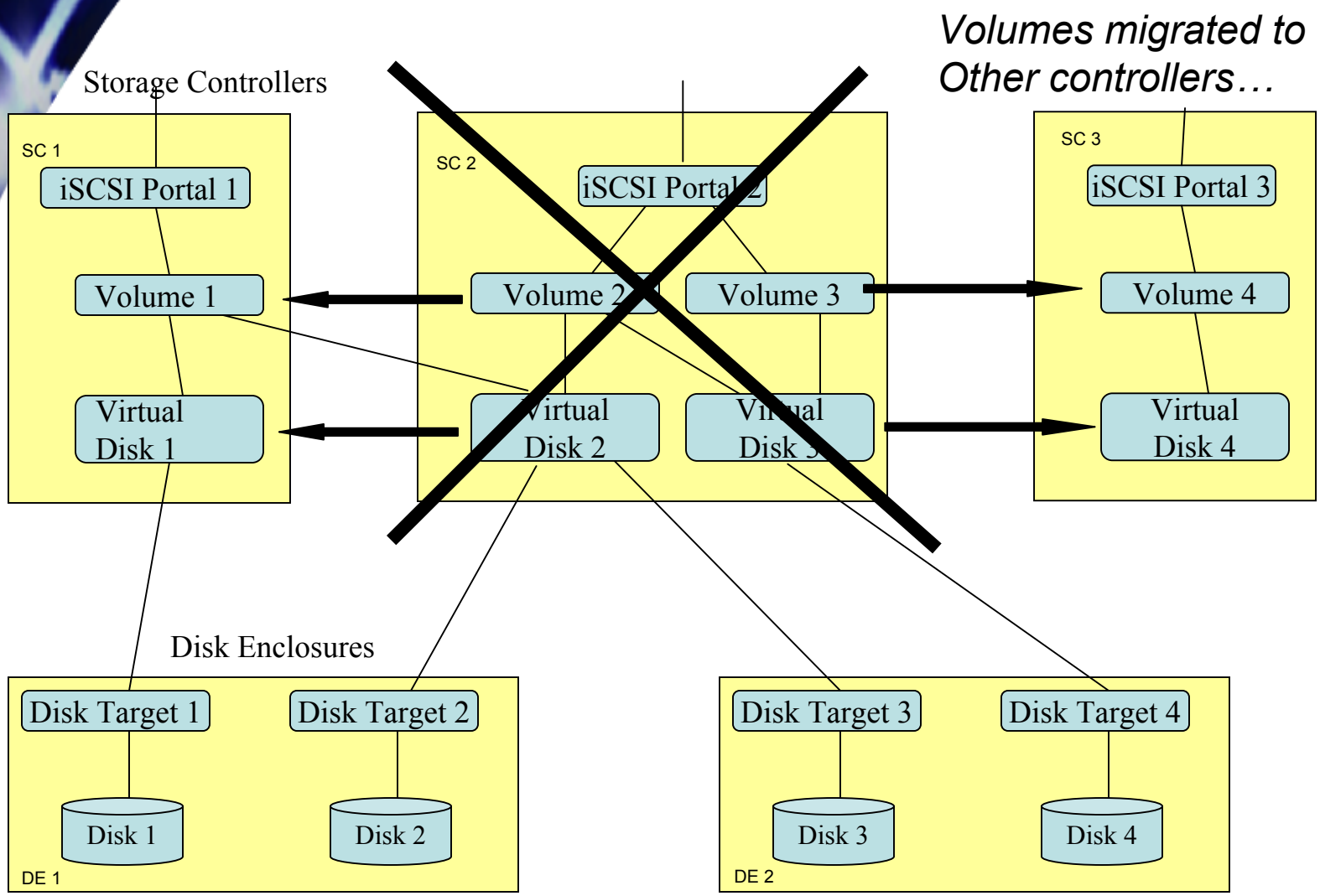
Migration Conditions

- Storage Controller failure
  - All volumes and virtual disks located on controller are migrated to other controllers
  - Migrated objects are distributed among all remaining controllers per a load balancing algorithm
- Loss of link to front-end switch
  - All volumes located on controller are migrated to other controllers
  - Virtual disks located on controller are NOT migrated
- Loss of link to back-end switch
  - All virtual disks located on controller are migrated
  - Volumes located on controller MAY be migrated for performance advantage (e.g. to avoid extra network hop via front-end network)
- Storage controller overload
  - Some volumes and virtual disks are migrated to reduce load
  - Other volumes and virtual disks remain

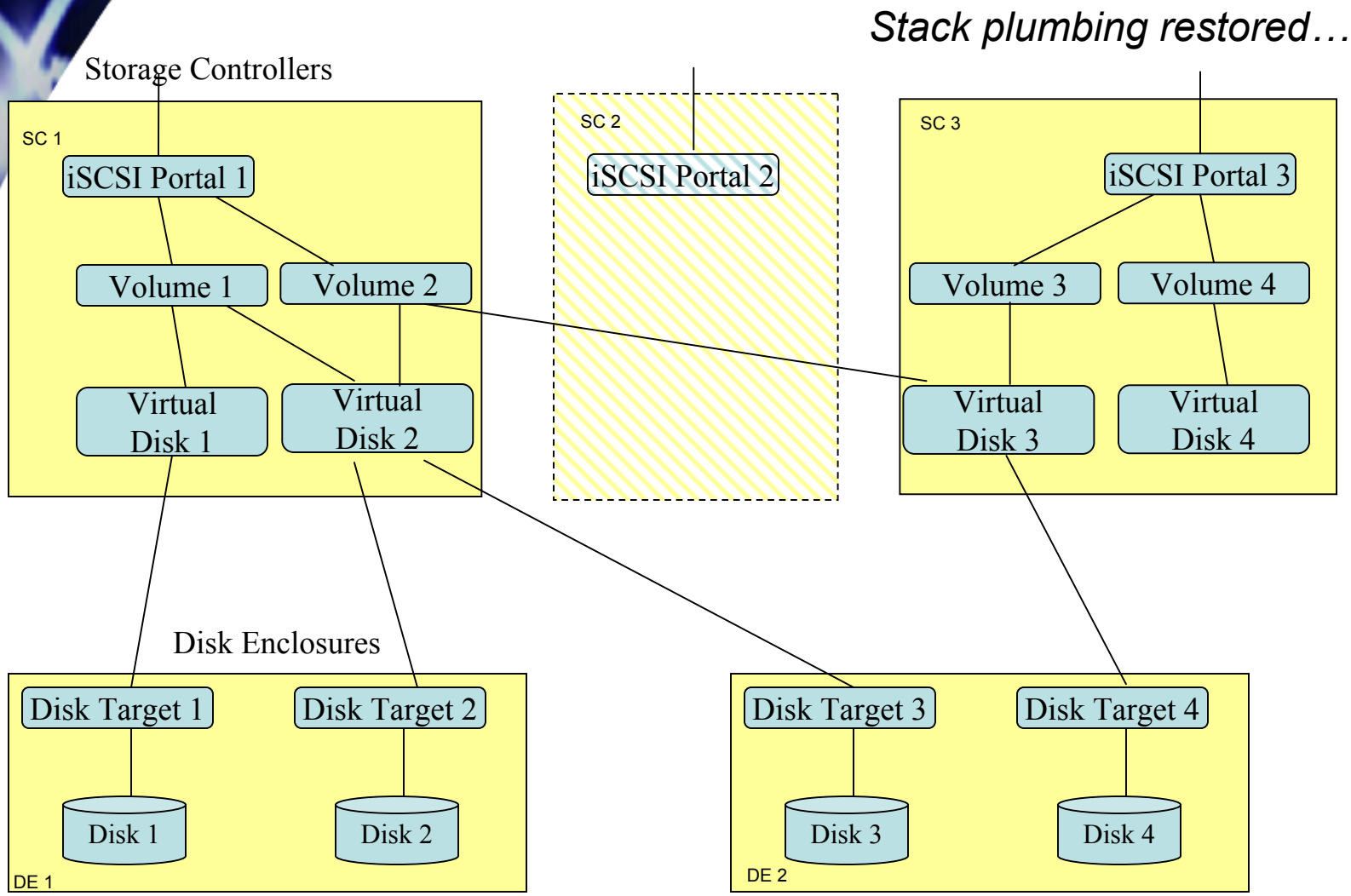
Migration Example 1: Storage Controller Failure



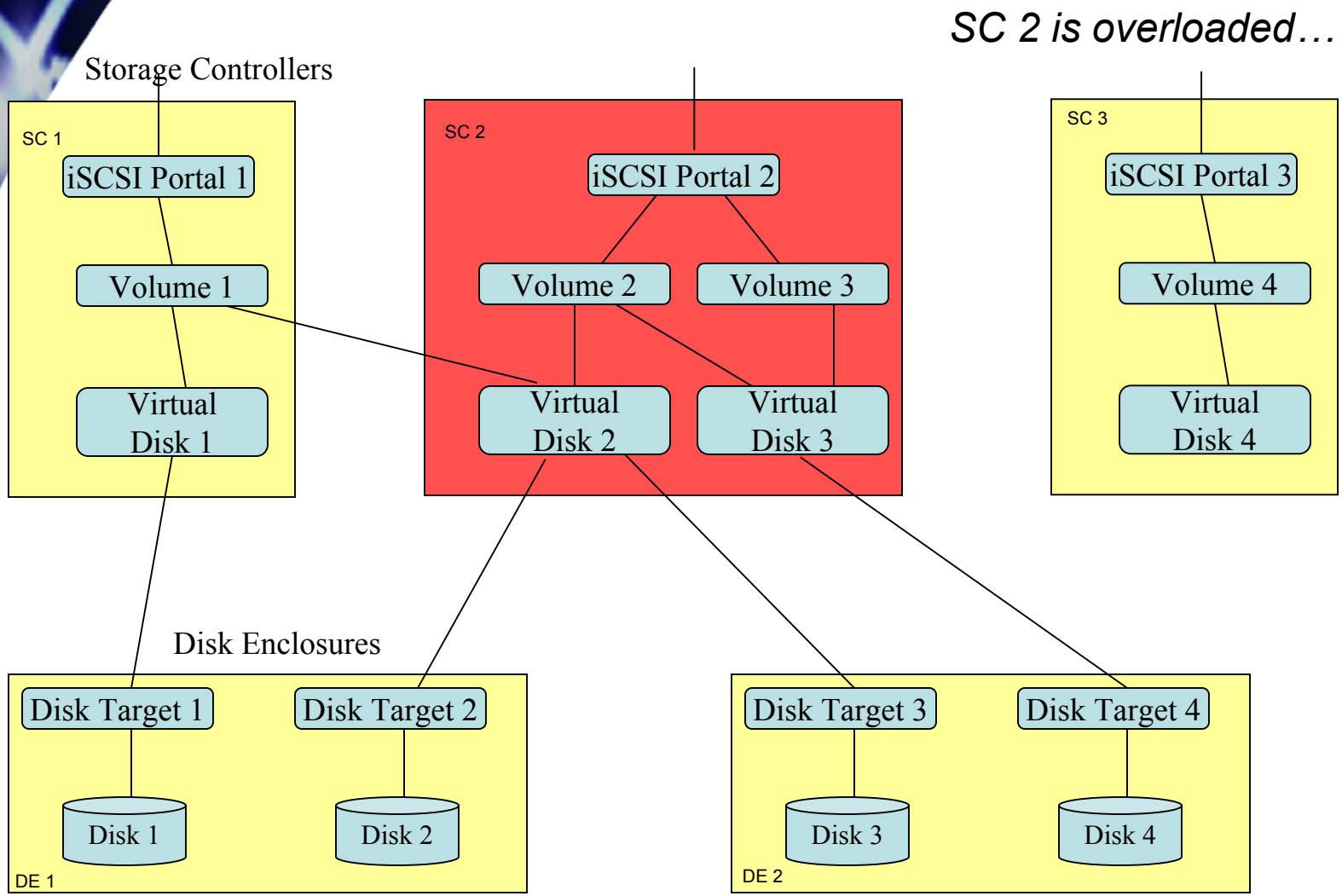
Migration Example 1: Storage Controller Failure



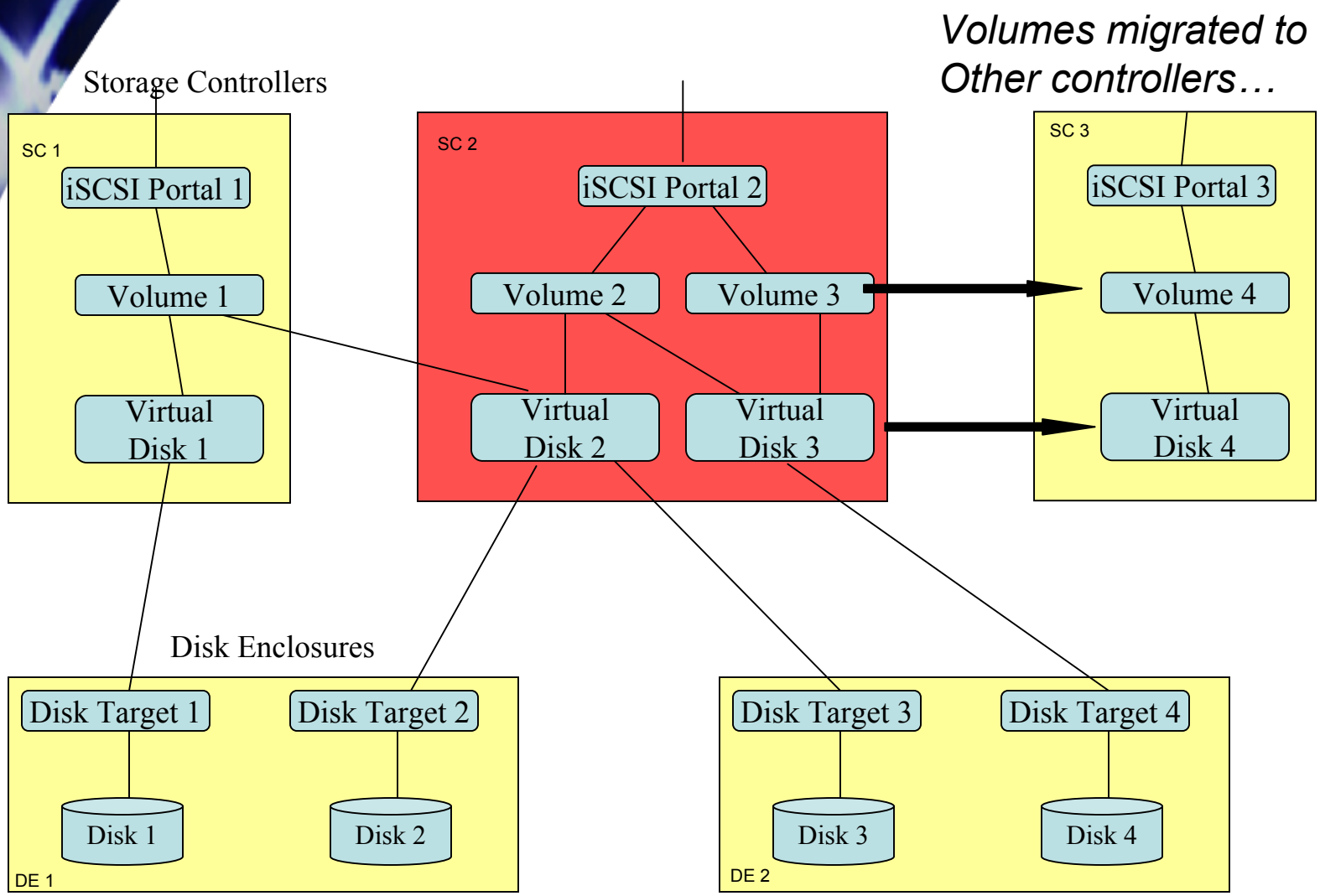
Migration Example 1: Storage Controller Failure



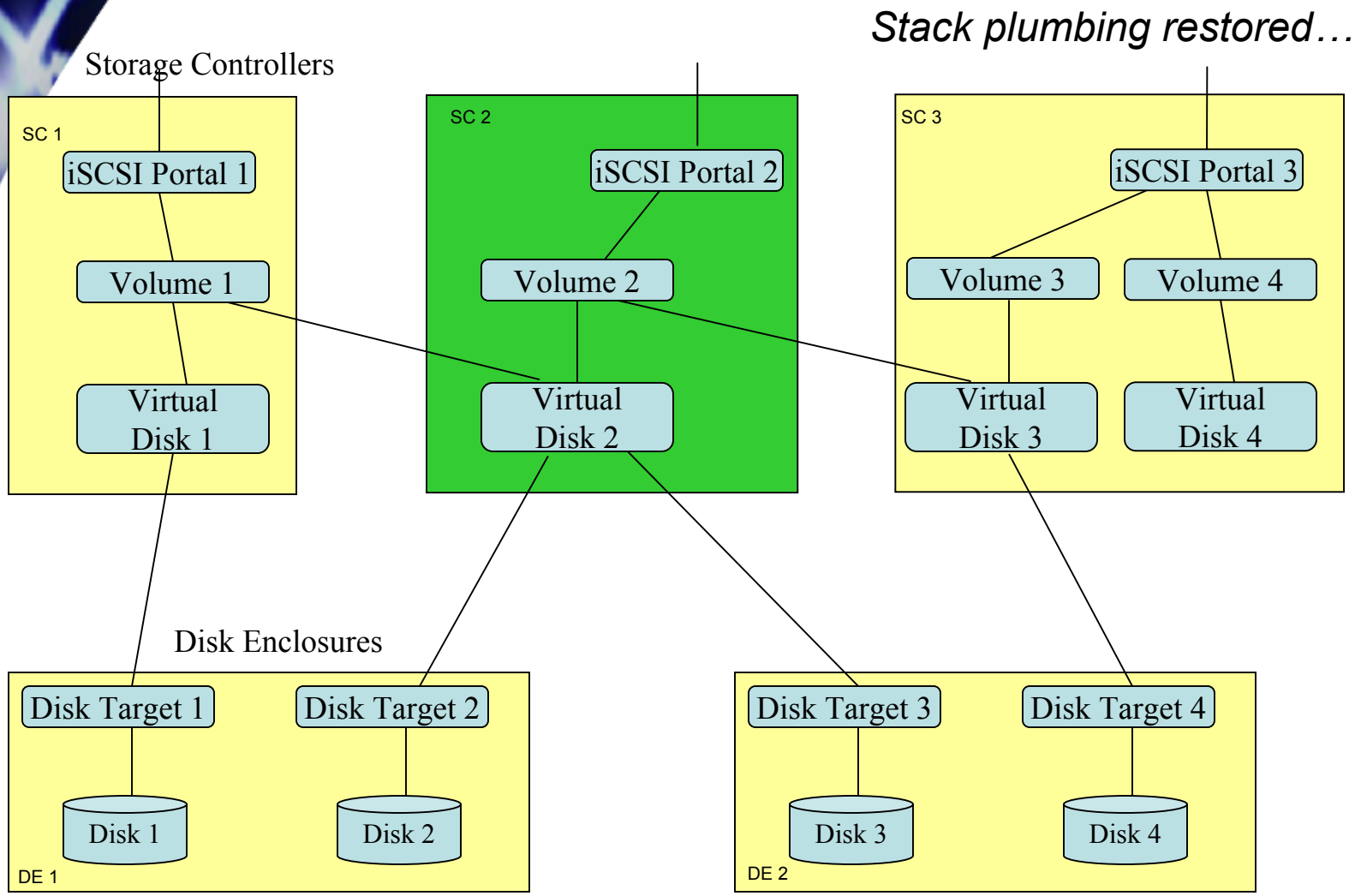
Migration Example 2: Storage Controller Overloaded



Migration Example 2: Storage Controller Overloaded



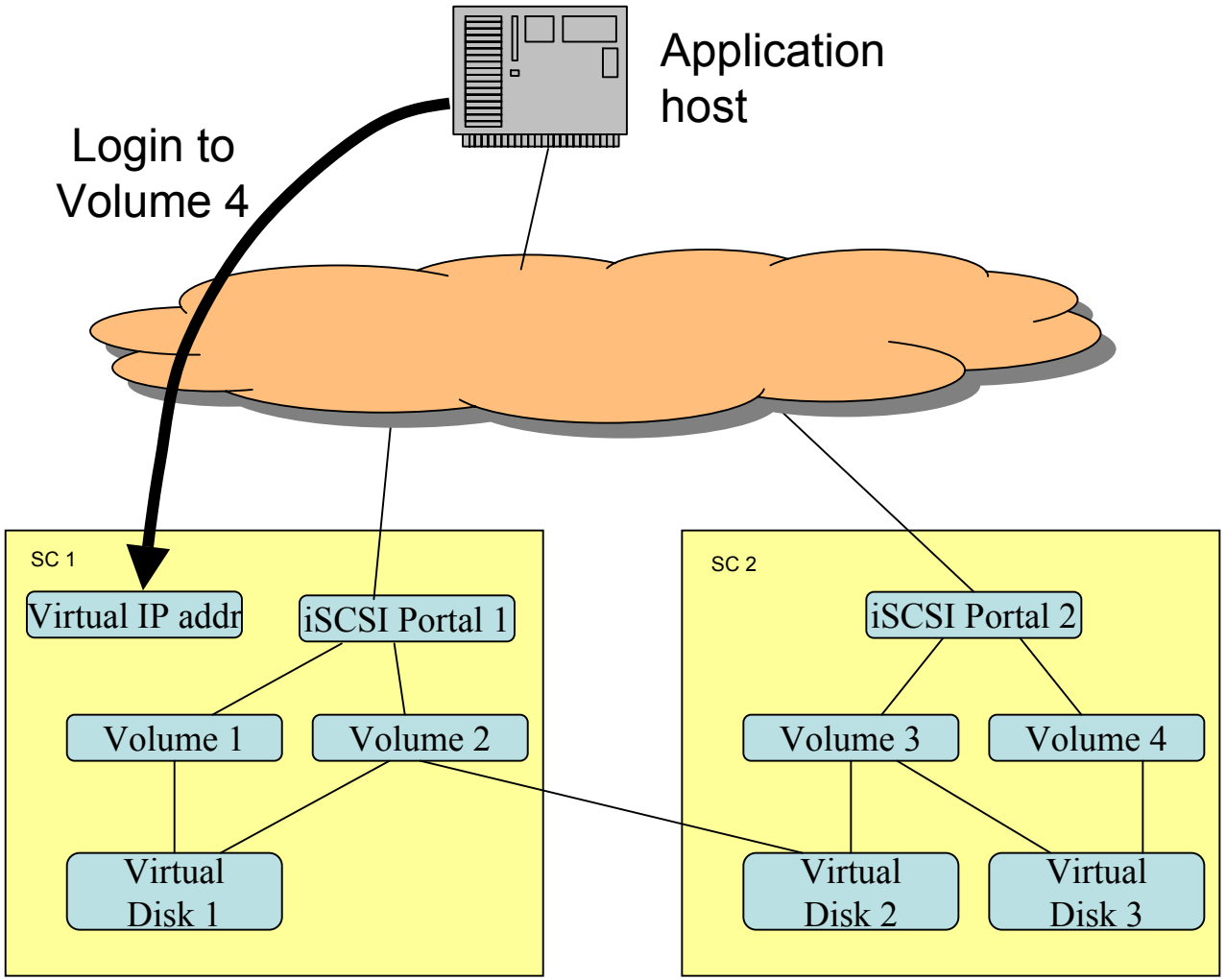
Migration Example 2: Storage Controller Overloaded



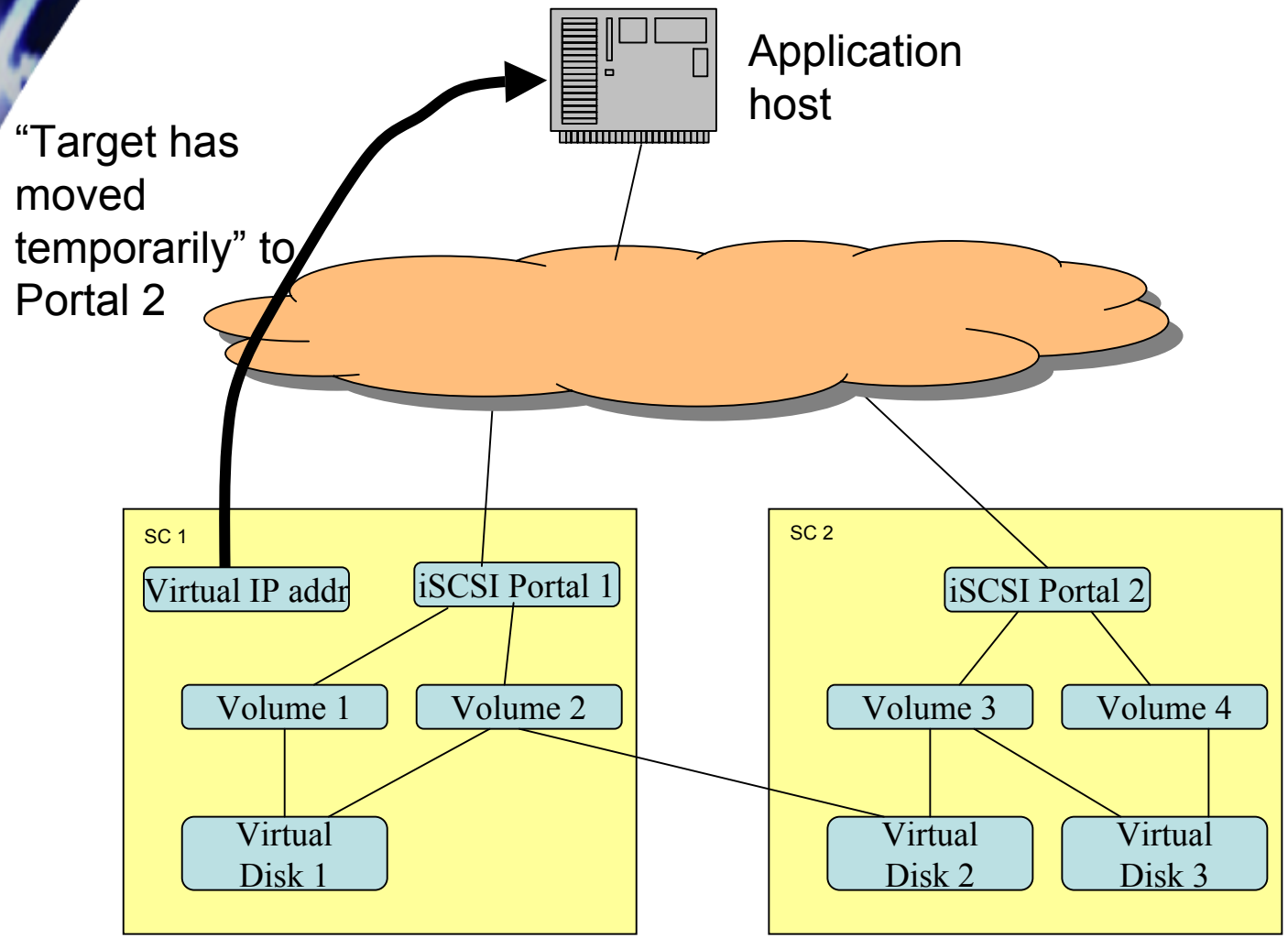
iSCSI Redirection

- Problem: How to route initiator to the controller holding volume of interest?
- Solution: iSCSI redirection feature
  - Single virtual IP address assigned to the cluster
  - Virtual IP address is served by one controller is used for initial connection from initiator
  - Controller issues “Target Moved Temporarily” login response to direct initiator to correct controller
  - Virtual IP address migrates when controller it is located on fails

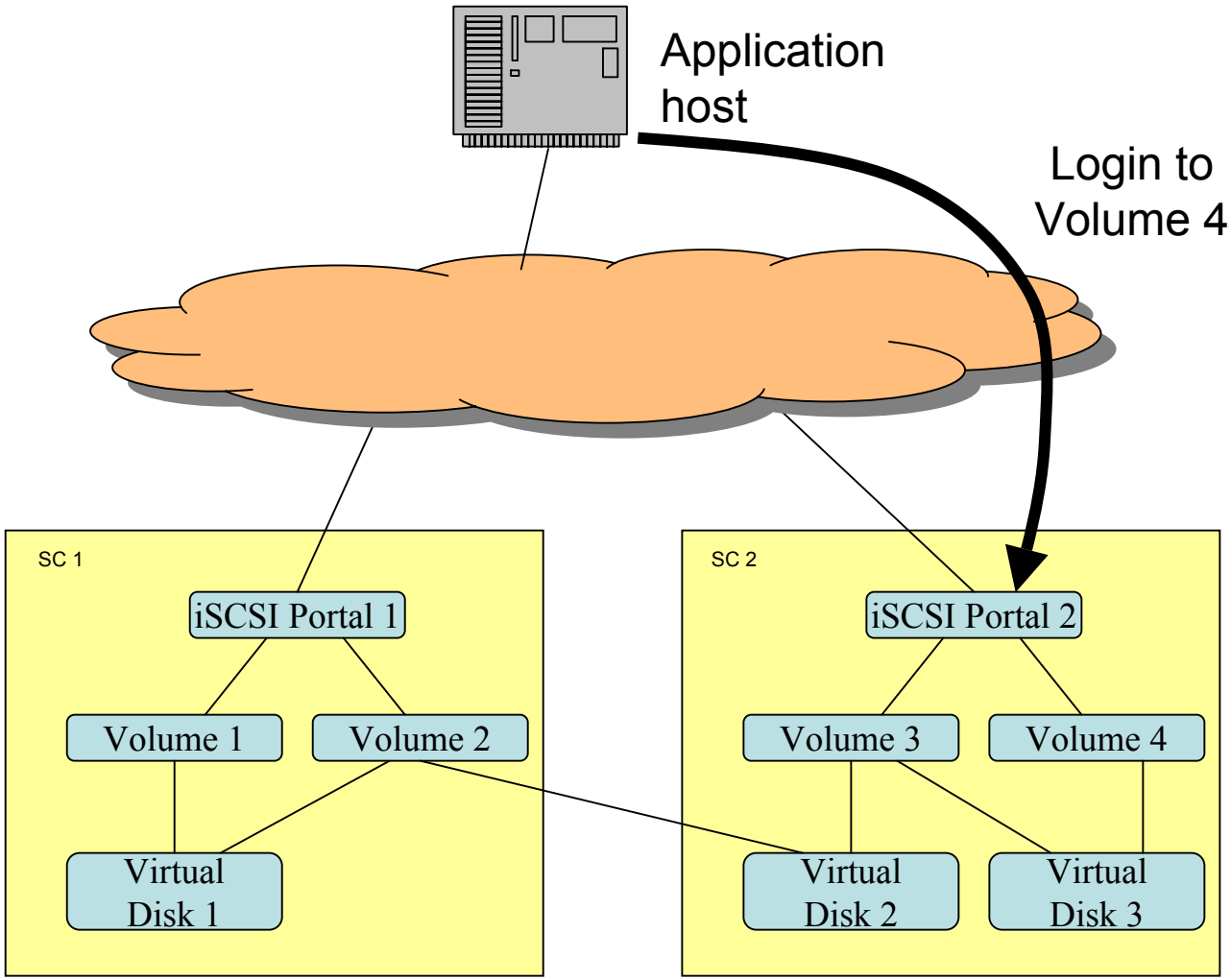
iSCSI Redirection Example Step 1



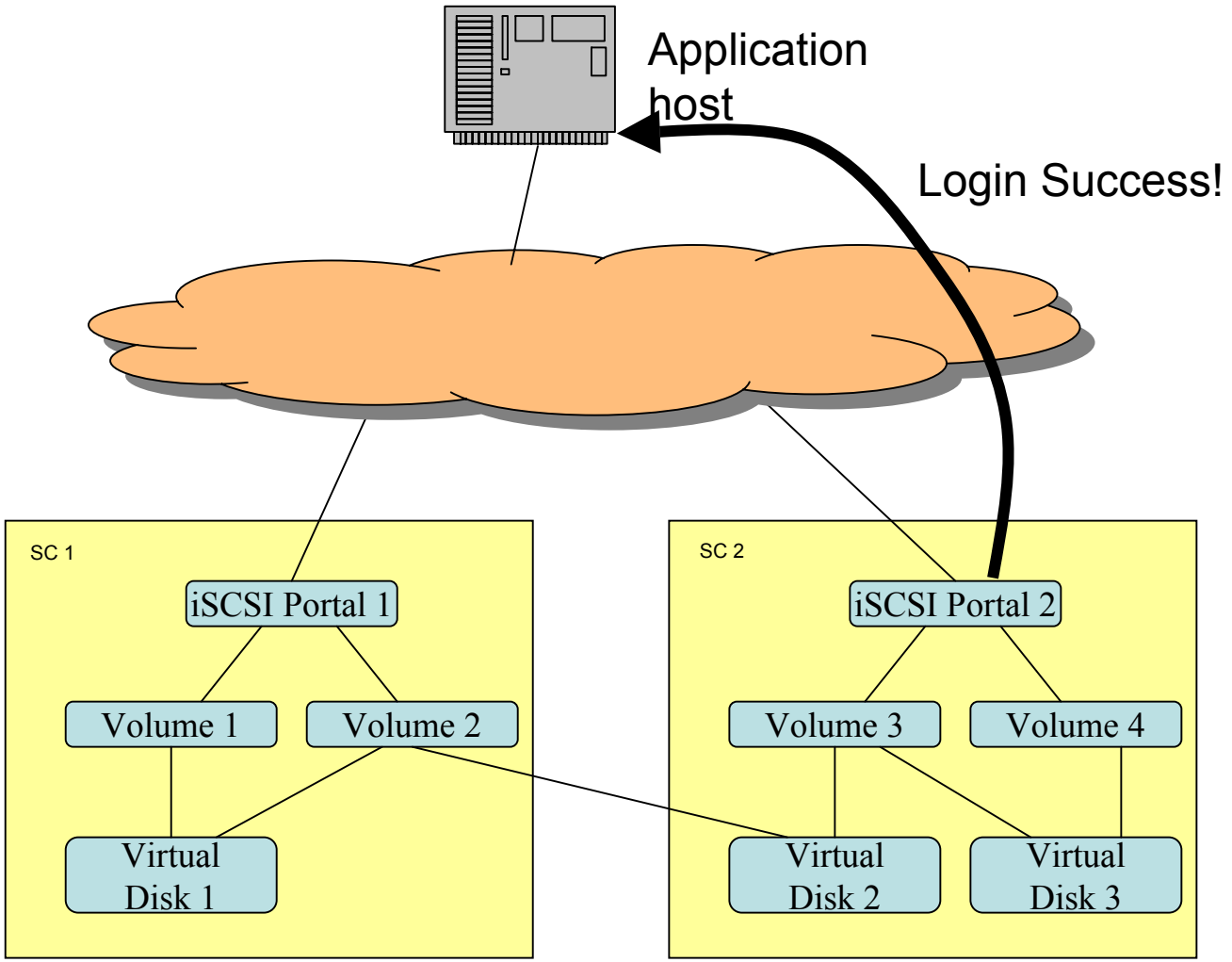
iSCSI Redirection Example Step 2



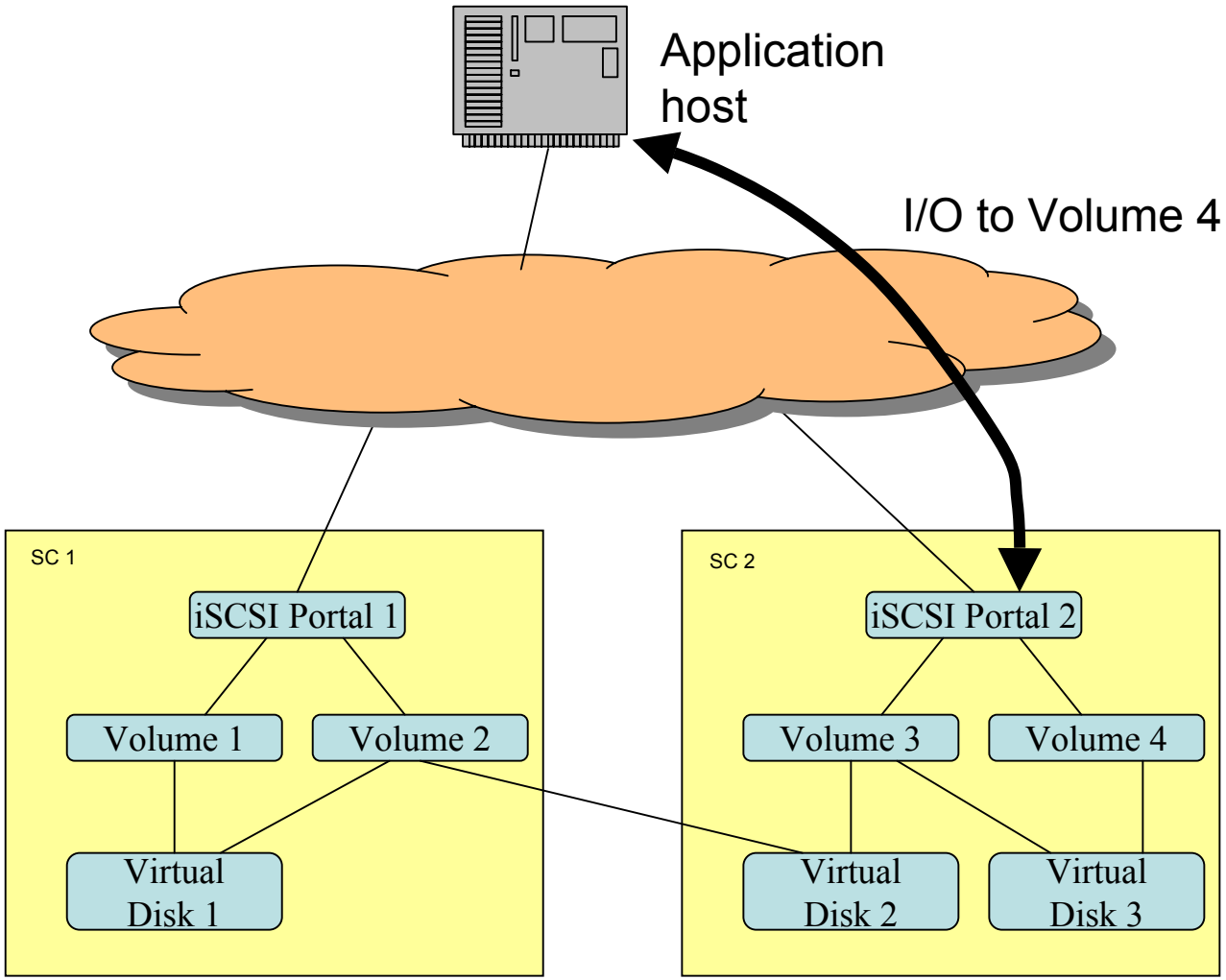
iSCSI Redirection Example Step 3



iSCSI Redirection Example Step 4



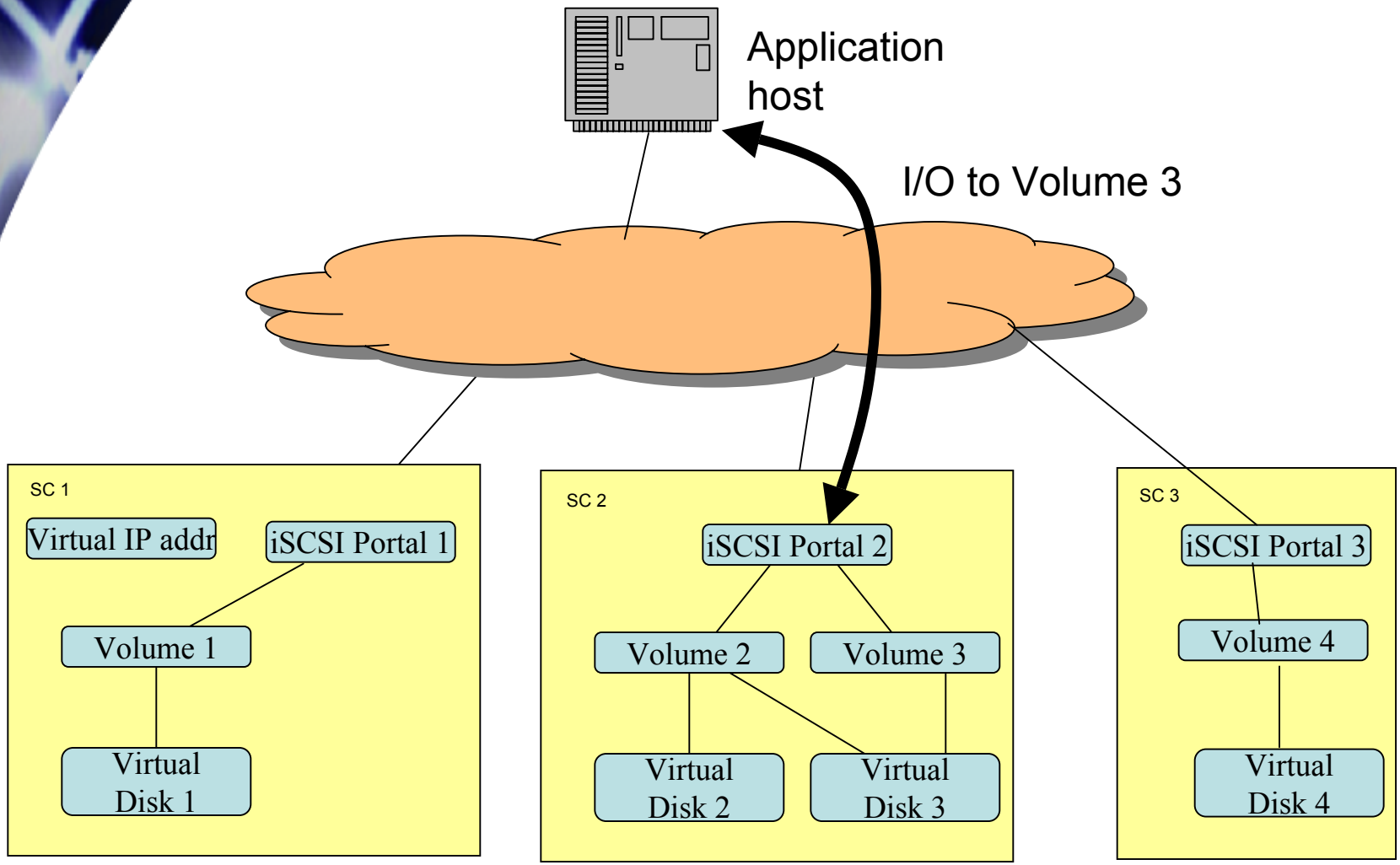
iSCSI Redirection Example Step 5



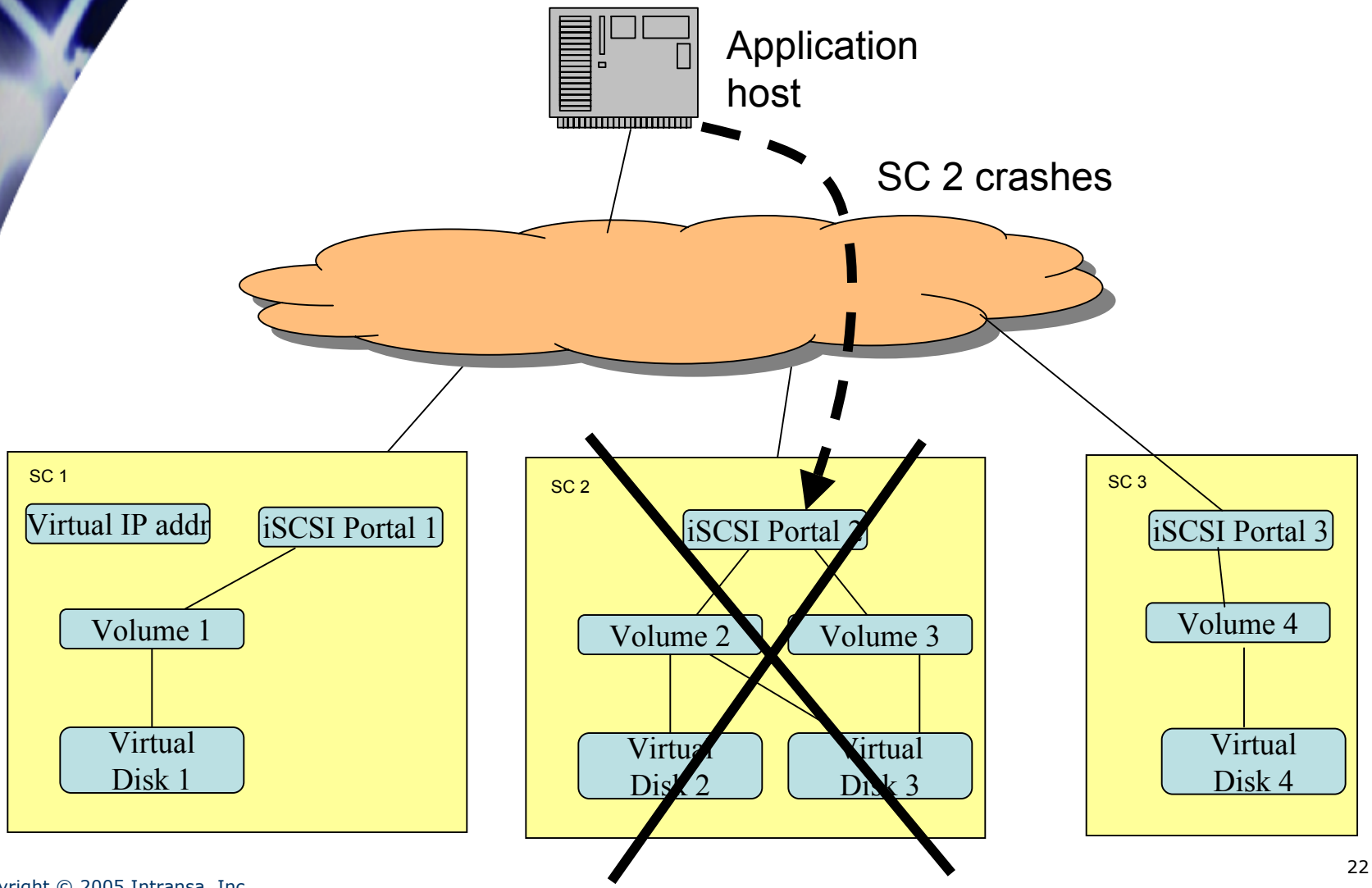
iSCSI Session Failover Three Cases

- Controller Failure:
- Front-end Link Failure:
  - All volumes and virtual disks are migrated to different controllers
  - iSCSI connections in progress are aborted or timed out
  - Initiators re-connect and log into virtual IP address
  - Initiators are re-directed to new controller
- Overloaded Controller:
  - Some volumes and virtual disks are migrated to other controllers
  - Controller issues iSCSI “asynchronous logout” on connections to volumes being moved
  - Initiators re-connect and log into virtual IP address
  - Controller issues “target moved temporarily” response with IP address of controller now holding volume
  - Initiators closes connection and re-log into new address

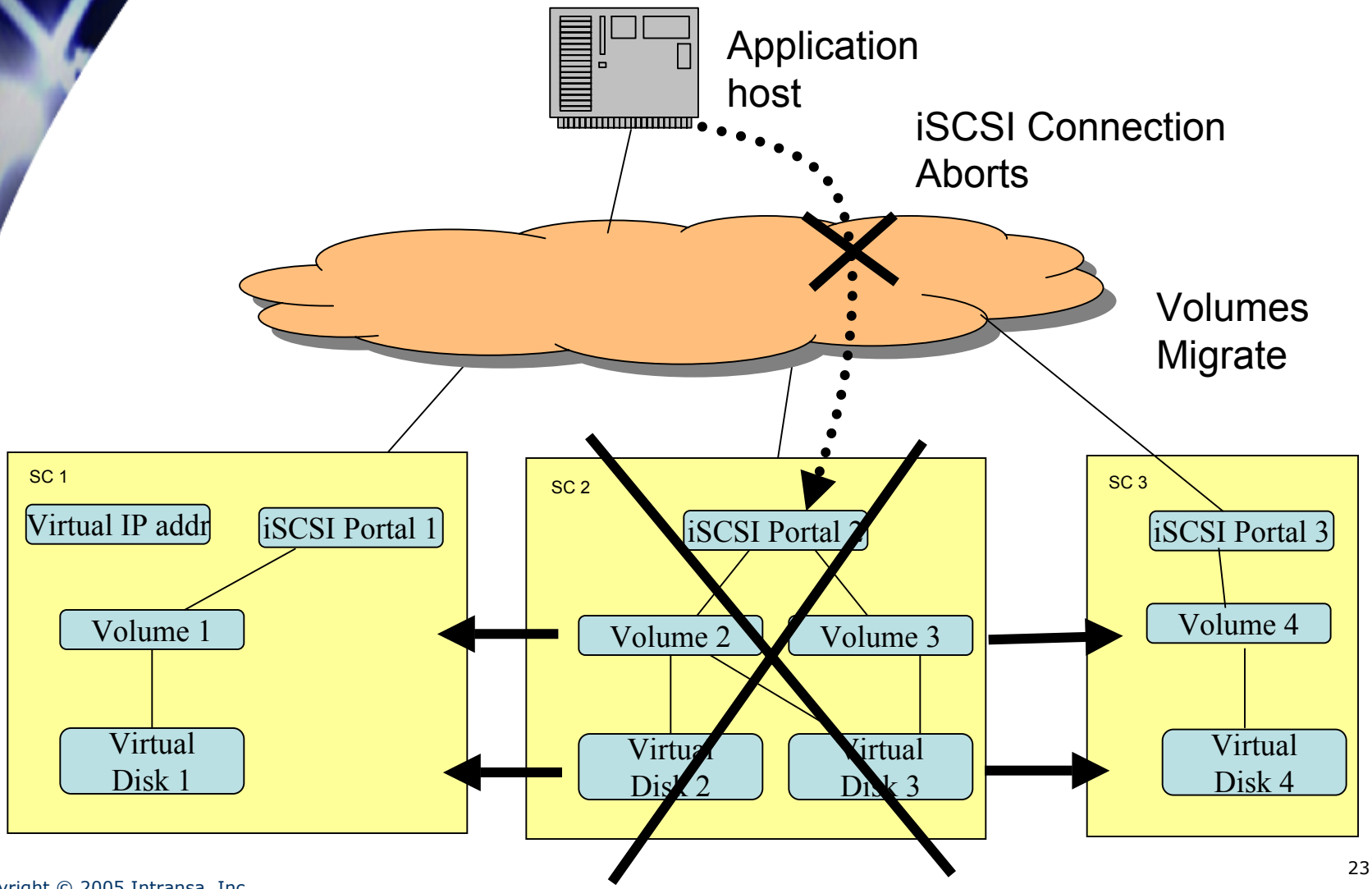
iSCSI Session Failover Example Step 1



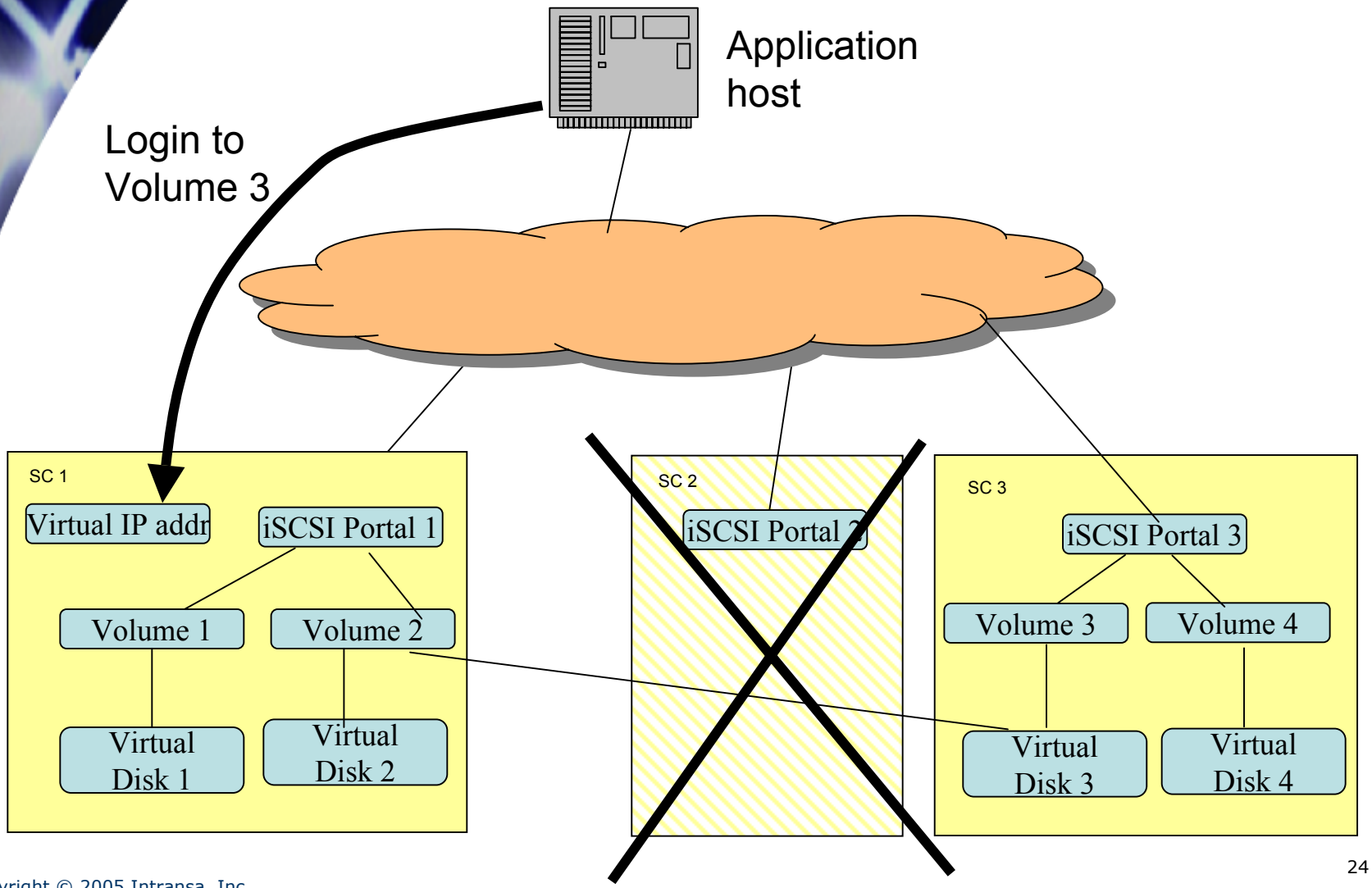
iSCSI Session Failover Example Step 2



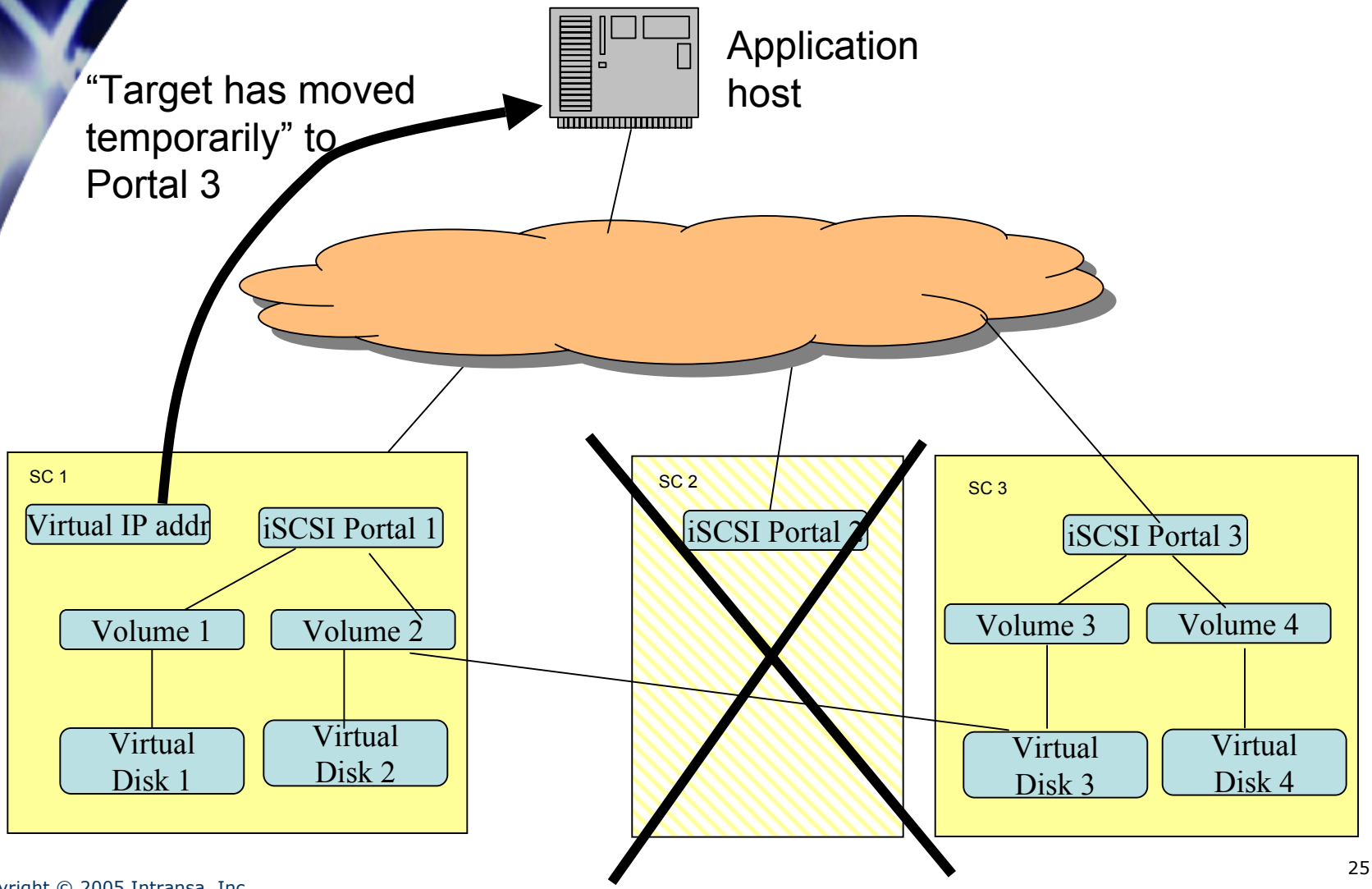
iSCSI Session Failover Example Step 3



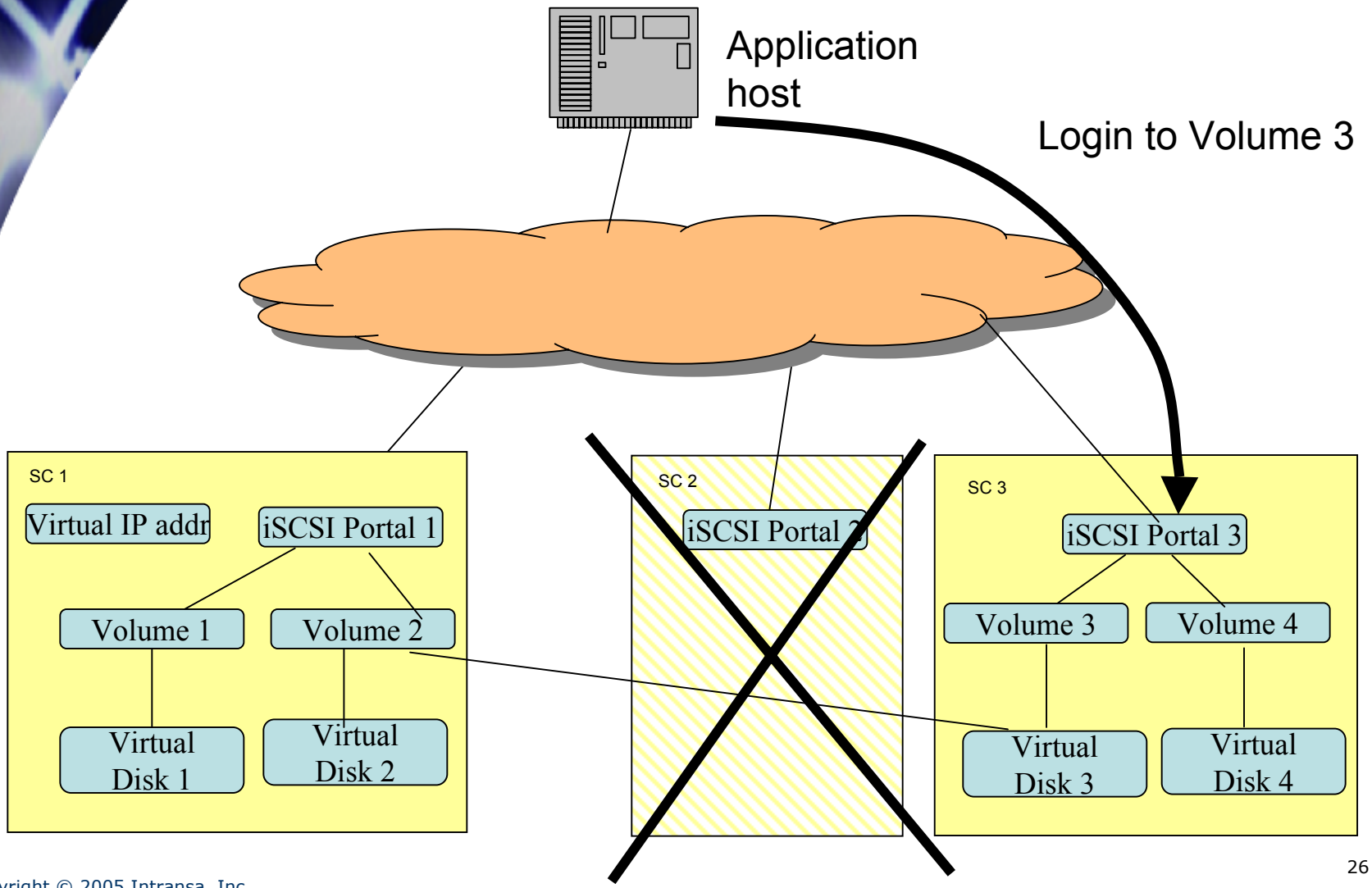
iSCSI Session Failover Example Step 4



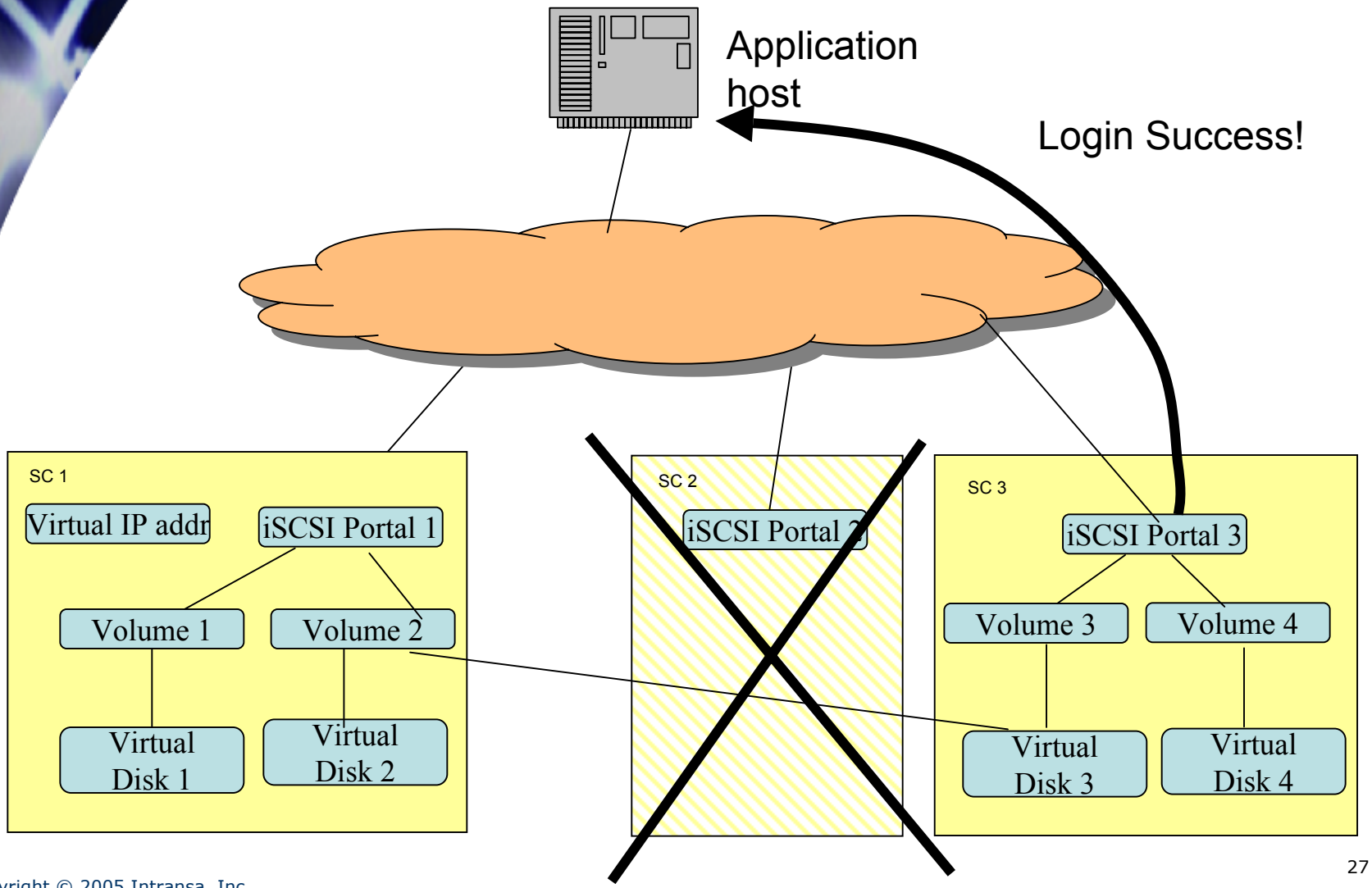
iSCSI Session Failover Example Step 5



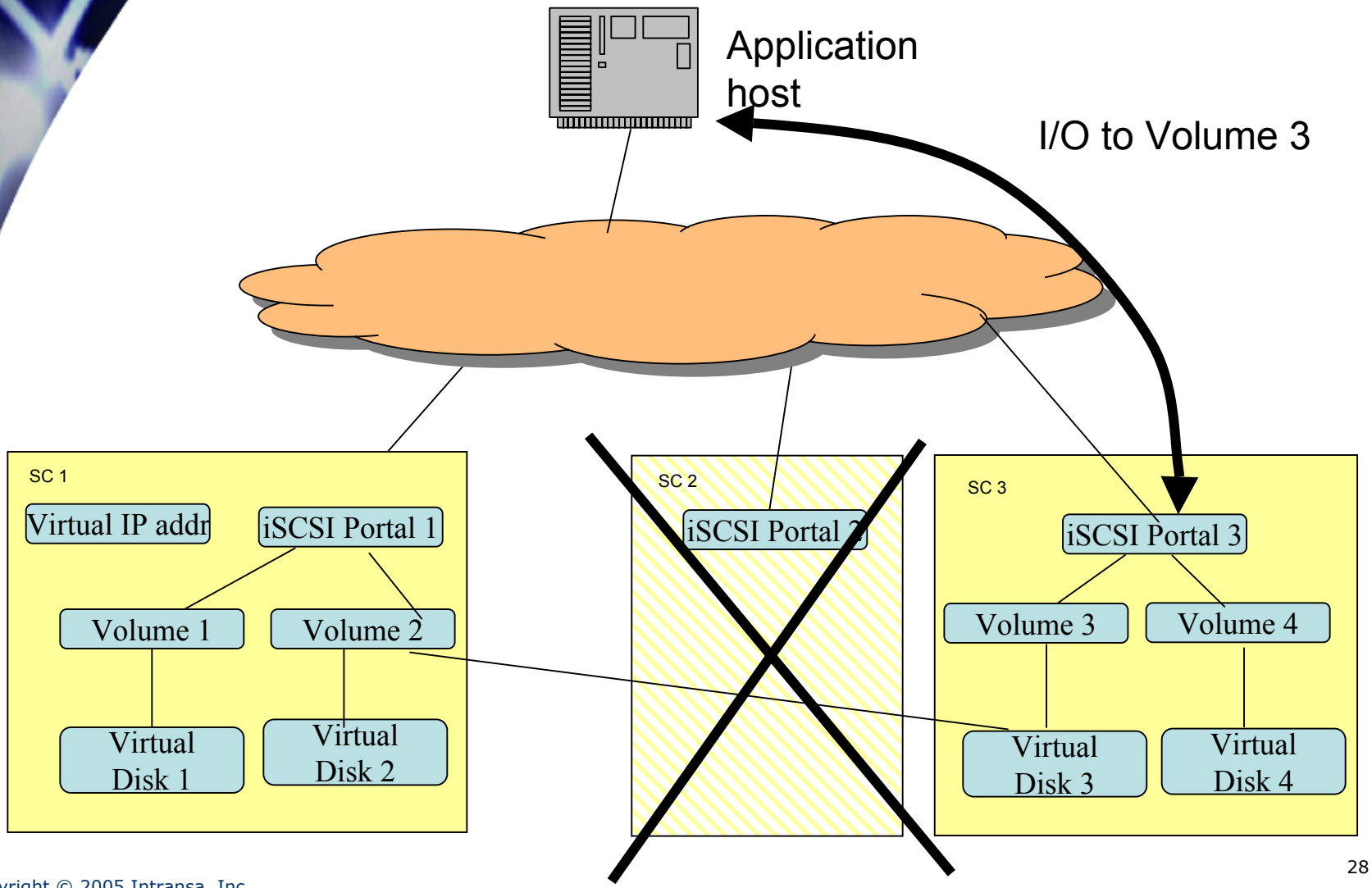
iSCSI Session Failover Example Step 6



iSCSI Session Failover Example Step 7



iSCSI Session Failover Example Step 8



Conclusions

- This architecture has been implemented and is shipping now in the Intransa IP SAN product
- Now deployed in production environments worldwide
- Scaling and High Availability design goals achieved
- Scaling has been demonstrated:
  - Routinely operate up to 6 storage controllers and eight 16-drive disk enclosures (128 drives) per cluster
  - Able to scale significantly higher
- Failover exercised in many failure scenarios
- Product is field proven and reliable