Architectures for HA and DR on Power Systems

Michel Passet

IBM PSSC Montpellier

Agenda

Topics to be covered are:

- Power HA System Mirror v7.1.3 updates
- Two typical HA architectures

- Power HA Stretched / Linked Clusters
- DR with host based mirroring (two architectures)
- DR with storage based mirroring
- HyperSwap functionality

- Conclusion

High Availability & Disaster Recovery

- HA
  - When a component of the IT infrastructure fails or stops in an unplanned or planned mode (HW or SW), the service to the users is not impacted, or impacted in a very limited scope (only the in-flight transactions that have to be rolled-back)
- Examples of HA features on Power Systems :
  - LVM Mirroring, **Live Partition Mobility**, Oracle RAC
  - Power HA cluster, Metro Mirror (synchronous mode)
- Worst case will lead to a restart.
  This can be done within minutes

High Availability & Disaster Recovery

- DR
  - When there is a wide outage impacting several or all components of a location (more than IT), then we call it a disaster
  - In such situation, all the IT service is interrupted and there is a decision to be taken by management whether to restart/recover or not to the alternate remote location. This can be done within hours or days, with more or less data loss.
- Examples of DR feature on Power Systems:
  - Power HA Enterprise Edition, Global Mirror, Oracle Data Guard (asynchronous data replication)

High Availability & Disaster Recovery

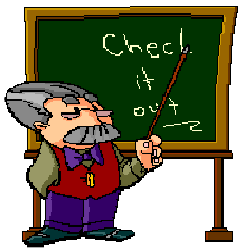
- RTO
  - Recovery Time Objective
    - Time to recover
    - Example 30 minutes. The service can be down for 30 minutes without major impact on the business.
      This value should be defined by the business, not by technology

- RPO
  - Recovery Point Objective
    - Data that can be lost during a failure
    - Zero means than no data can be lost, even in the worst case

Power HA System Mirror v7.1.3

TRADITIONAL

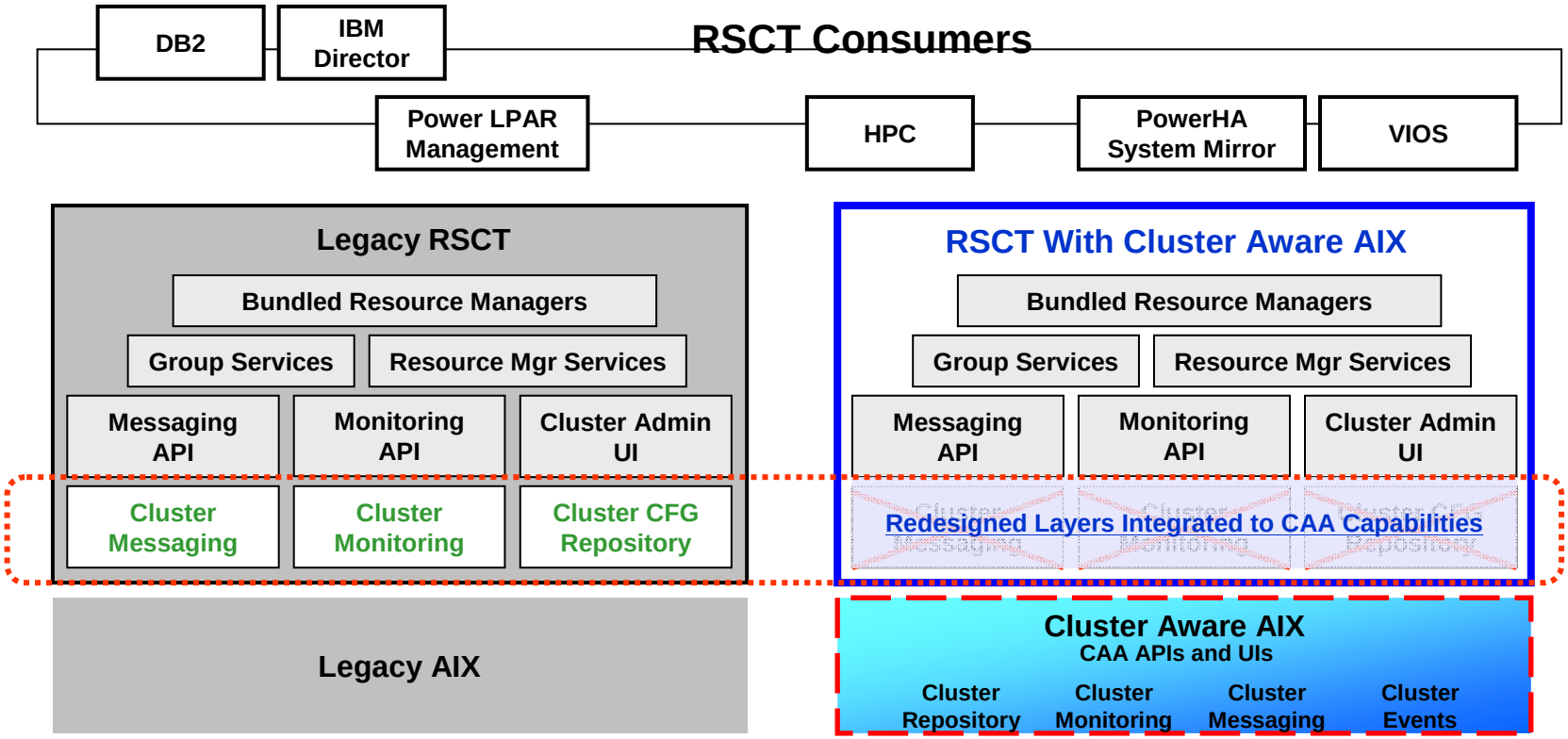
- Data Center Clustering
- Added on top of AIX
- Traditional storage configurations
- Automated 2 site failover
- Roll your own scripts
- Operations = change management



PowerHA V7

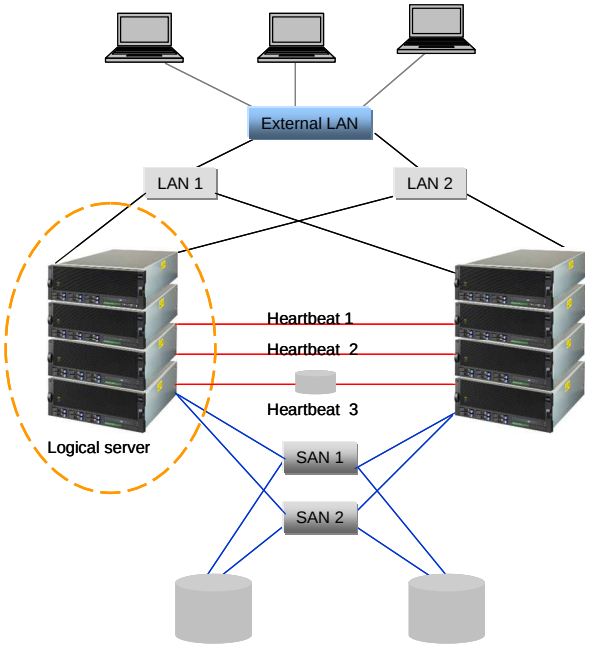
- Multi-Site Clustering
- Integrated into AIX (CAA)
- HyperSwap storage configurations
- Operator controlled 2 site failover
- Smart Assists (included at no charge)
- Operations = minimal change management

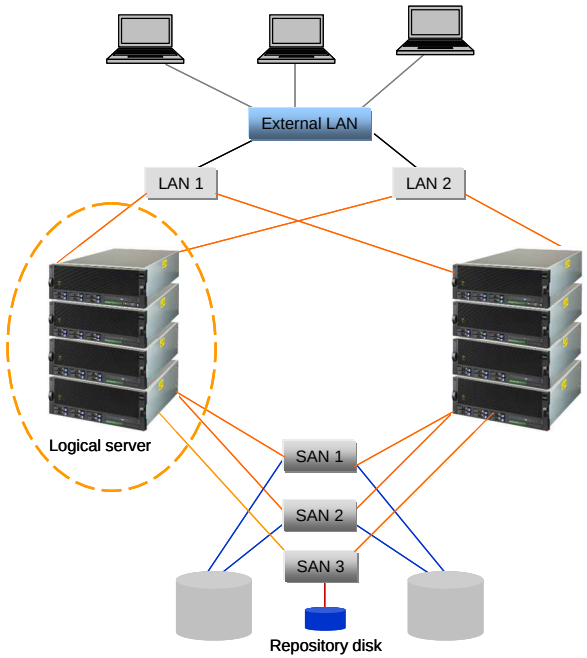
Cluster Aware AIX: Core Cluster Support



- CAA: Provides the core clustering primitives in AIX
- RSCT and Cluster Aware AIX together provide the foundation of strategic Power Systems SW
- RSCT integration with CAA extends simplified cluster management along with optimized and robust cluster monitoring, failure detection, and recovery to RSCT exploiters on Power / AIX

PowerHA 6.1 verse PowerHA 7.1





- Traditional Comm based heartbeat
- Round robin
- User space event processing
- RSCT topology management
- EOM 09/08/2013
- EOS 04/30/2015

- Comm, SAN and Repository heartbeat
- Unicast or multicast
- Kernel based event processing
- Repository disk topology management
- GUI management interface

PowerHA SystemMirror V7 Standard Edition

PowerHA SystemMirror 7.1 Standard Edition represents the next generation of High Availability for AIX

- ✓ **Based on OS integrated clustering** for simplicity and reliability
- ✓ **Systems Director-based management** for simple, centralized cluster administration
- ✓ **Smart Assists** to simplify deployment of high availability for SAP and other applications
- ✓ **Multiple redundant heartbeat** with SAN communications for robust cluster integrity
- ✓ **Advanced resource group policies** for automated recovery sequencing

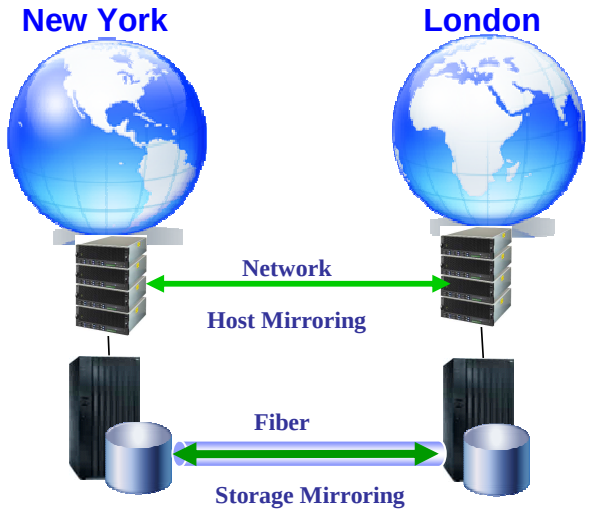
- **2013 Enhancements**
  - **Unicast Clustering**
  - **Repository disk recover config to new disk**
- **Dynamic Host Name modification support**

7.1.0 GA: Sep 2010
7.1.1 GA: Dec 2011
7.1.2 GA: Nov 2012
7.1.3 GA: Dec 2013

PowerHA SystemMirror AIX Enterprise Edition

High Availability and Disaster Recovery across multi site
Compute & Storage infrastructure

- **PowerHA SystemMirror for AIX Enterprise Edition**
  - Long distance failover for Disaster Recovery
  - Low cost host based mirroring support (GLVM)
- **Extensive support for storage array replication**
  - Short distance (~100KMs) deployment: Synchronous
  - Long distance (1000's of KM) deployment: Asynchronous



Supported Mirroring Technologies

	Replication Technology	Sync	Async
Host Replication	Geo LVM (GLVM)	✓	✓
Storage Array Replication	IBM DS8K Series Storage - PPRC	✓	✓
	SVC, Storwize,	✓	✓
	XIV	✓	✓
	EMC – SRDF	✓	✓
	Hitachi – Universal Replicator,Truecopy	✓	✓
	HP – Continuous Access	✓	✓

Time to move to PowerHA V7

PowerHA SystemMirror	6.1	7.1	PowerHA 7.1 Benefit
IBM Director based graphical user interface	NA	✓	Ease of Use
Cluster Aware AIX (CAA)	NA	✓	Reliability
Triple redundant heartbeat	NA	✓	Effectively eliminates partitioning
SAN based communications	NA	✓	Additional cluster communication path
Stretched cluster (shared repository)	NA	✓	Two-Site unicast or multicast HA/DR
Cross Site Mirroring (single site stretch cluster)	NA	✓	LVM mirroring with CAA
Linked clusters (separate repositories)	NA	✓	Two-Site HA/DR separate networks
HyperSwap with DS8800, DS8870	NA	✓	Two-Site continuously available storage
Active-Active HyperSwap & single node HyperSwap	NA	✓	Options for continuous app and storage availability
Multi-Site set up wizard	NA	✓	Speeds up implementation
Two site linked cluster operator managed failover	NA	✓	Operator decides whether or not to failover
Federated Security	NA	✓	Cluster wide security management
Live Cache SAP hot standby	NA	✓	Fast failover for APO SCM
Smart Assists for LiveCache and Netweaver	NA	✓	Faster, customizable deployment
Root Vg failure handling	NA	✓	Avoid downtime due to inactive OS

PowerHA V6.1 EOS: 4/30/2015

Three year service extension planned

PowerHA SystemMirror 7.1.X Editions for AIX

High Level Features	Standard Edition	Enterprise Edition
Centralized Management CSPOC	✓	✓
Cluster resource management	✓	✓
Shared Storage management	✓	✓
Cluster verification framework	✓	✓
Integrated disk heartbeat	✓	✓
SMIT management interfaces	✓	✓
AIX event/error management	✓	✓
Integrated heartbeat	✓	✓
PowerHA DLPAR HA management	✓	✓
Smart Assists	✓	✓
Multi Site HA Management	✓	✓
PowerHA GLVM async mode		✓
GLVM deployment wizard		✓
IBM Metro Mirror support		✓
IBM Global Mirror support		✓
OEM Copy Services		✓
Hyperswap Support		✓

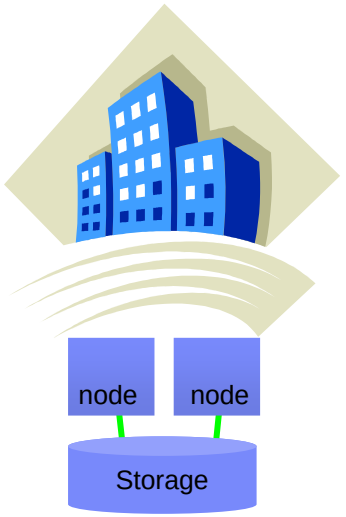
12

Highlights:

- New Editions to optimize software value capture
- Standard Edition targeted at datacenter HA
- Enterprise Edition targeted at multi-site HA/DR
  - Stretched Clusters
  - Linked Clusters
- Per processor core used + tiered pricing structure
  - Small/Med/Large

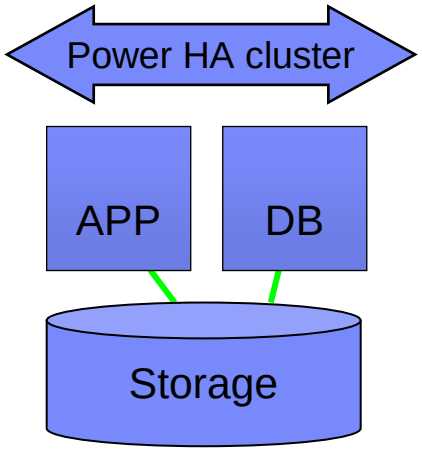
High Availability & Disaster Recovery

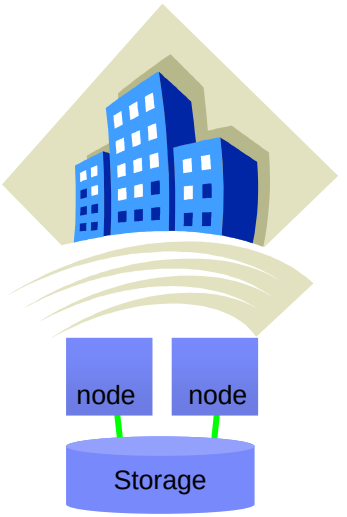
- High Availability : one secured site
  - Typically two nodes and a storage bay in a data center
- HA architecture
  - All components are doubled
    - Servers
    - The storage itself is entirely redundant
- Does not protect against
  - Storage failure
  - Data center failure



High Availability & Disaster Recovery

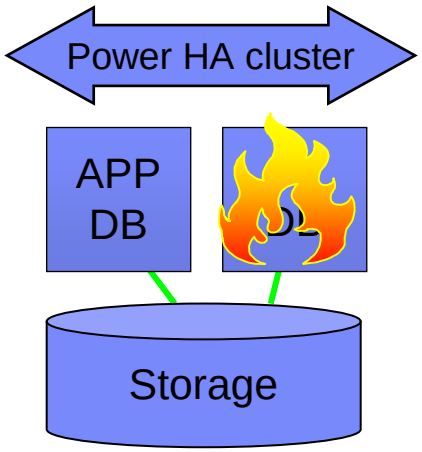
- Typical architecture with IBM Power HA System Mirror
  - Apps and DB are both secured by automatically failing over the other node
  - There is a service outage





High Availability & Disaster Recovery

- In case of node failure
  - Both APP and DB are hosted by the same node
- In case of storage or data center failure
  - No service any more

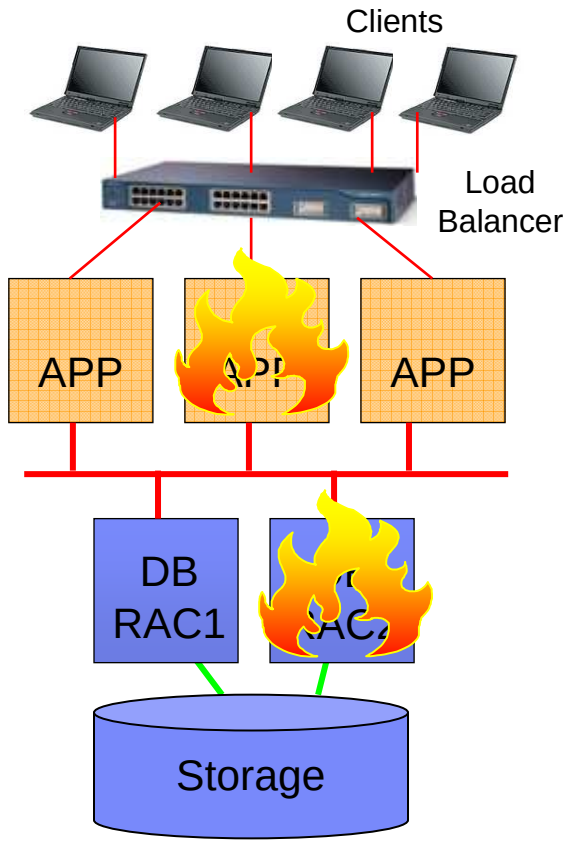


- Micro partitionning and Capacity On Demand (CoD) can be used to avoid performance concerns at the client end

- RTO = few minutes
- RPO = no data loss

High Availability & Disaster Recovery

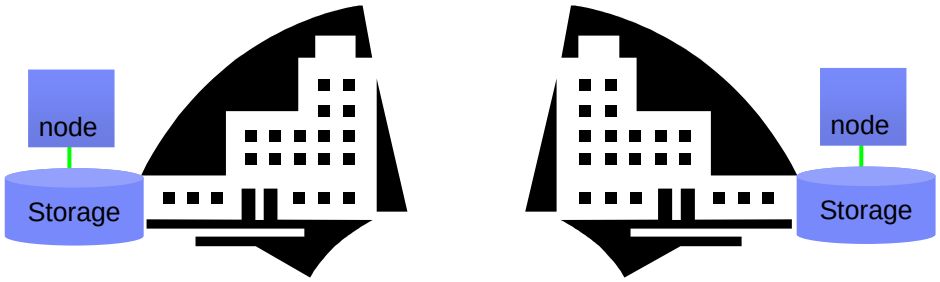
- Apps – DB typical architecture with Oracle RAC
  - Apps are protected by multiple clones
    - If one APP server is failing, the clients are directed to the others APP servers
  - DB is protected by Oracle RAC
    - If one instance is failing, the database remains available



- RTO = zero
- RPO = no data loss

High Availability & Disaster Recovery

- Disaster Recovery : two sites
  - Typically two data centers in different buildings of the same company
    - Each data center hosts one node and one storage unit
  - Some kind of mirroring has to be defined between the two storages units
- Small distance, to lower latency concerns (< 50 Km)

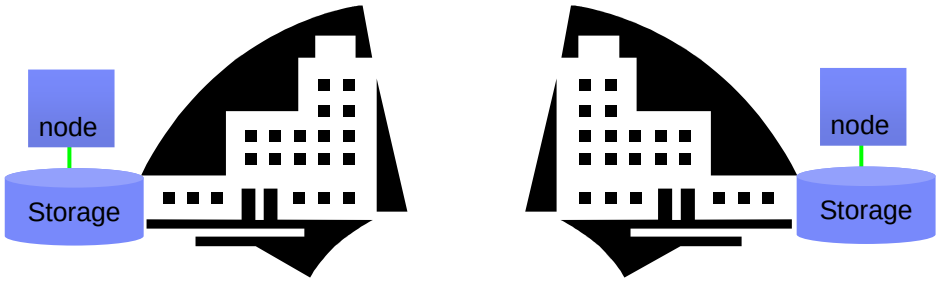


High Availability & **Disaster Recovery**

- Data Mirroring can be achieved at different levels

  – Operating System layer
    - LVM mirroring, GLVM, GPFS replication, …

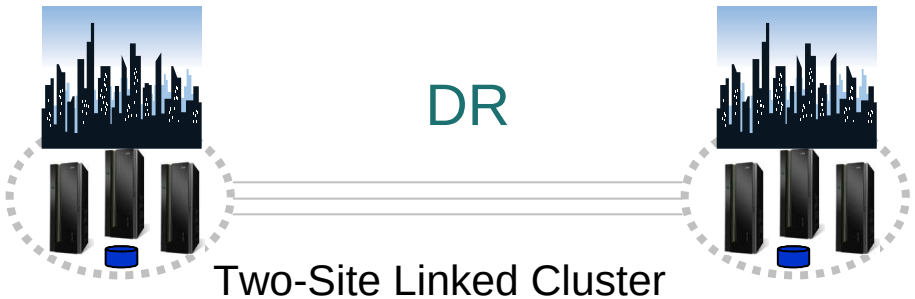
  – Storage layer
    - Metro Mirror, Global Mirror, SRDF, TrueCopy…, SVC vdisk mirroring



PowerHA 7.1 Stretched / Linked Clusters

- PowerHA SystemMirror Enterprise Edition
  - Two for two-site deployments
  - Each option provide configurations optimized for customer requirements

Two-Site Stretched Cluster



Two-Site Linked Cluster

- Stretched Cluster
  - Supports unicast (default) or multicast communications
  - Triple redundant heartbeat
  - Campus/Metro deployments
  - Targetted for enhanced HA (automatic failover)
  - Use of LVM mirroring
- Linked Cluster
  - Enables two sites with independent networks (campus or cross country)
  - For cross country deployments (suitable also for Campus/Metro)
  - Targetted for DR (manual failover)
  - Use of Metro Mirror

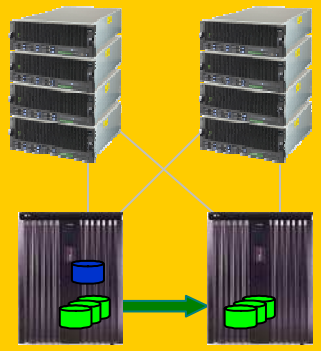
PowerHA SystemMirror 7.1 Stretched Clusters

Cross-Site Mirroring
Campus

Hyper Swap Mirroring
Metropolitan Region
Active – Active
Active - Passive

Single Data Center
Applications remain active

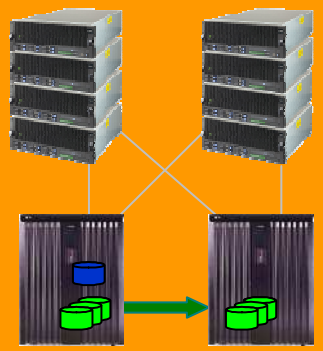
Continuous access to data in the event of a storage subsystem outage



PowerHA-Standard Edition
LVM Mirroring
RPO=0, RTO = 0

Two Data Centers
Systems remain active

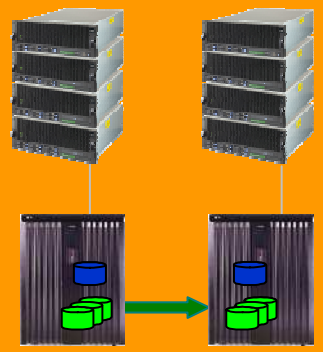
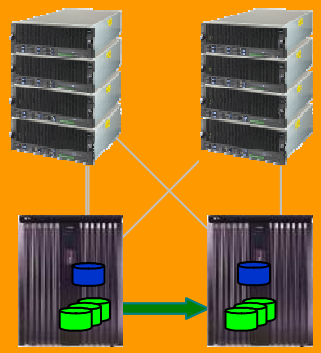
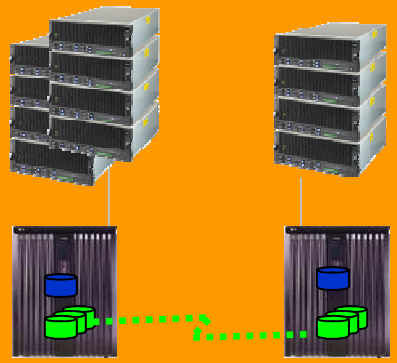
Continuous access to data in the event of a storage subsystem outage



PowerHA Enterprise Edition
DS8800 and Metro Mirror
RPO =0 RTO <1 hr
Storage RTO..minutes

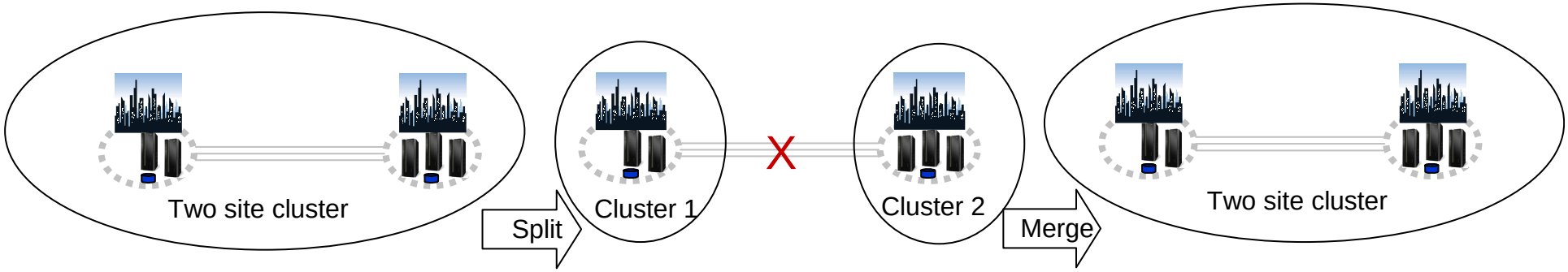
Repository Disk Standard Edition Enterprise Edition

PowerHA SystemMirror 7.1 Linked Clusters

Recovery within a Metropolitan Region	Hyper-Swap Mirroring Metropolitan Region Active – Passive only (synchronous)	Disaster Recovery at Extended Distance
Two Data Centers Systems remain active Multi-site workloads can withstand site and/or storage failures	Single Data Centers Applications remain active Continuous access to data in the event of a storage subsystem outage	Two Data Centers Rapid Systems Disaster Recovery with "seconds" of Data Loss Disaster recovery for out of region interruptions
		
PowerHA Enterprise Edition GLVM, Metro Mirror, SRDF,TrueCopy RPO=0 & RTO<1 hr	PowerHA-Entertprise Edition DS8800 and Metro Mirror RPO=0, RTO ~ 0	PowerHA Enterprise Edition GLVM/PPRC/SVC/SRDF/TrueCopy RPO secs & RTO <1 hr

Repository Disk Standard Edition Enterprise Edition

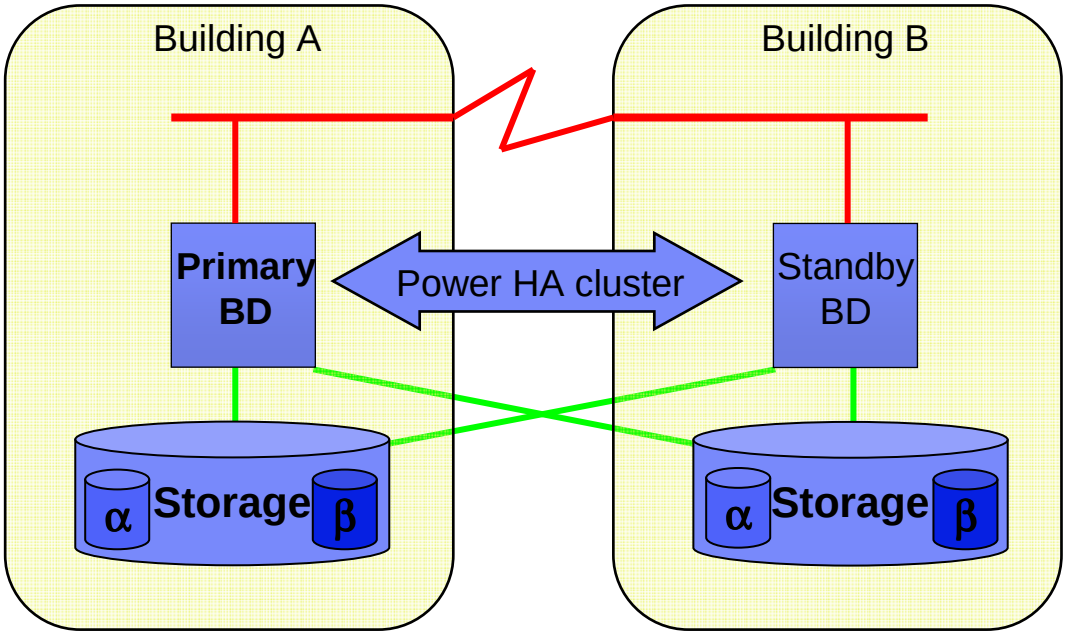
Linked Clusters Split/Merge handling



Policy Setting	Split	Merge	Approach
Manual	✓	✓	Manual steps needed for recovery to continue
Tie Breaker	✓	✓	Tie break Holder side wins
Majority Rule		✓	Greater of N/2 side wins Else, side that includes node with the smallest node id wins
Priority		✓	Operator chooses a numerical value such as "largest serial number"

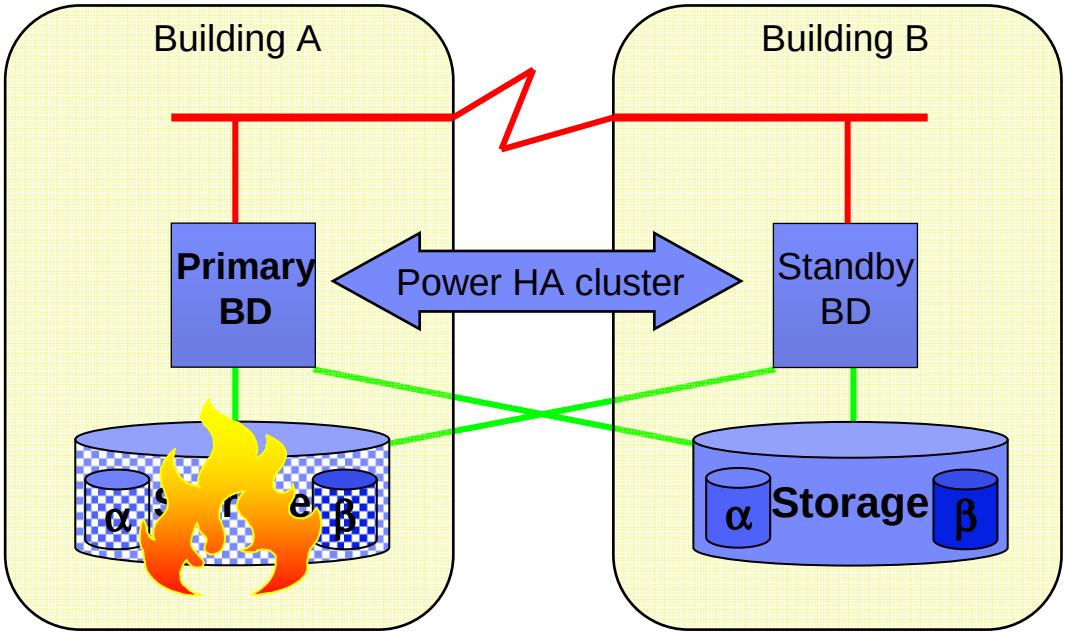
DR & LVM mirroring (Stretched Cluster)

- Example with a standalone database
  - DB node protected by Power HA System Mirror (Standard Edition)
  - Data mirrored using AIX LVM mirroring



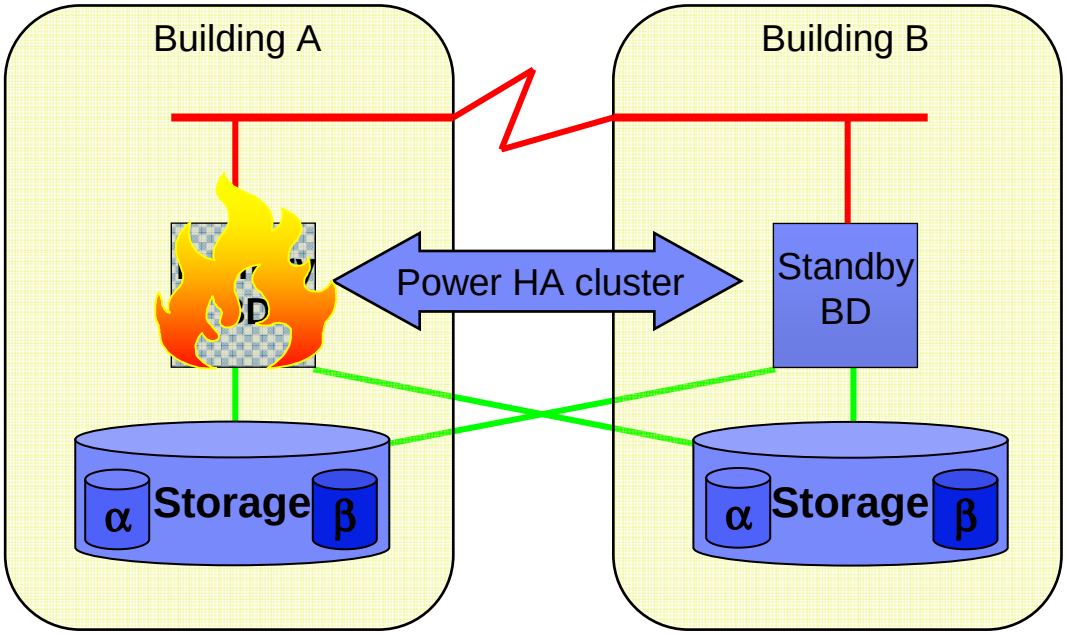
DR & LVM mirroring (Stretched Cluster)

- Storage failure: no service outage
  - AIX LVM still have one good copy
  - Failover automatic, fallback needs mirror resync



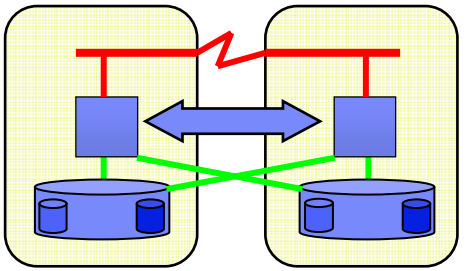
DR & LVM mirroring (Stretched Cluster)

- Primary node failure
  - Power HA System Mirror is restarting automatically the application on the standby node
  - There is a short service outage



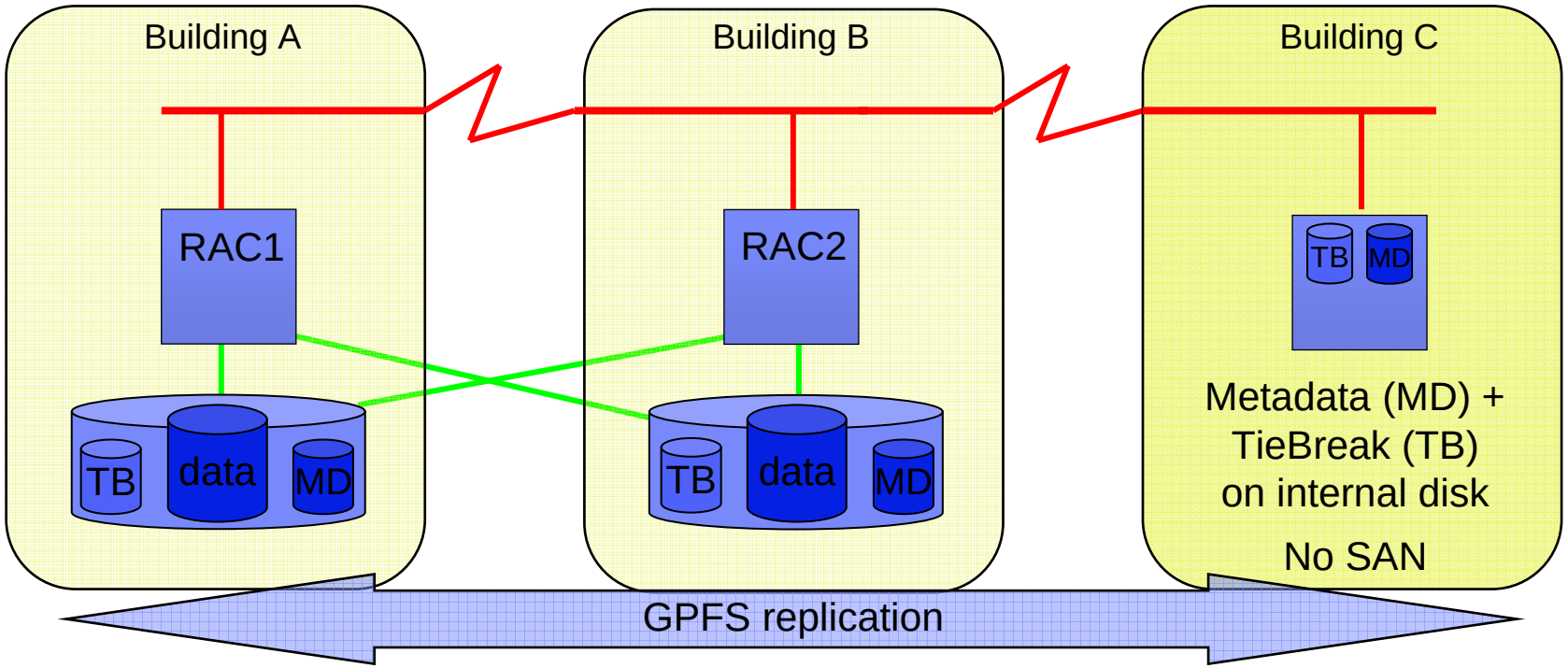
DR & LVM mirroring (Stretched Cluster)

- + Advantages
  - Both mirrors are used for read/write
  - LVM is using the local copy (preferred) for its reads -> no performance degradation
  - LVM is writing in parallel on the two copies (limited impact)
  - No outage in case of storage failure. Managed by LVM, not by Power HA
- - Disadvantages
  - Distance limited by latency / bandwidth
  - Quorum has to be managed
  - Short outage in case of node failure
- RTO = Time to restart the DB
- RPO = 0



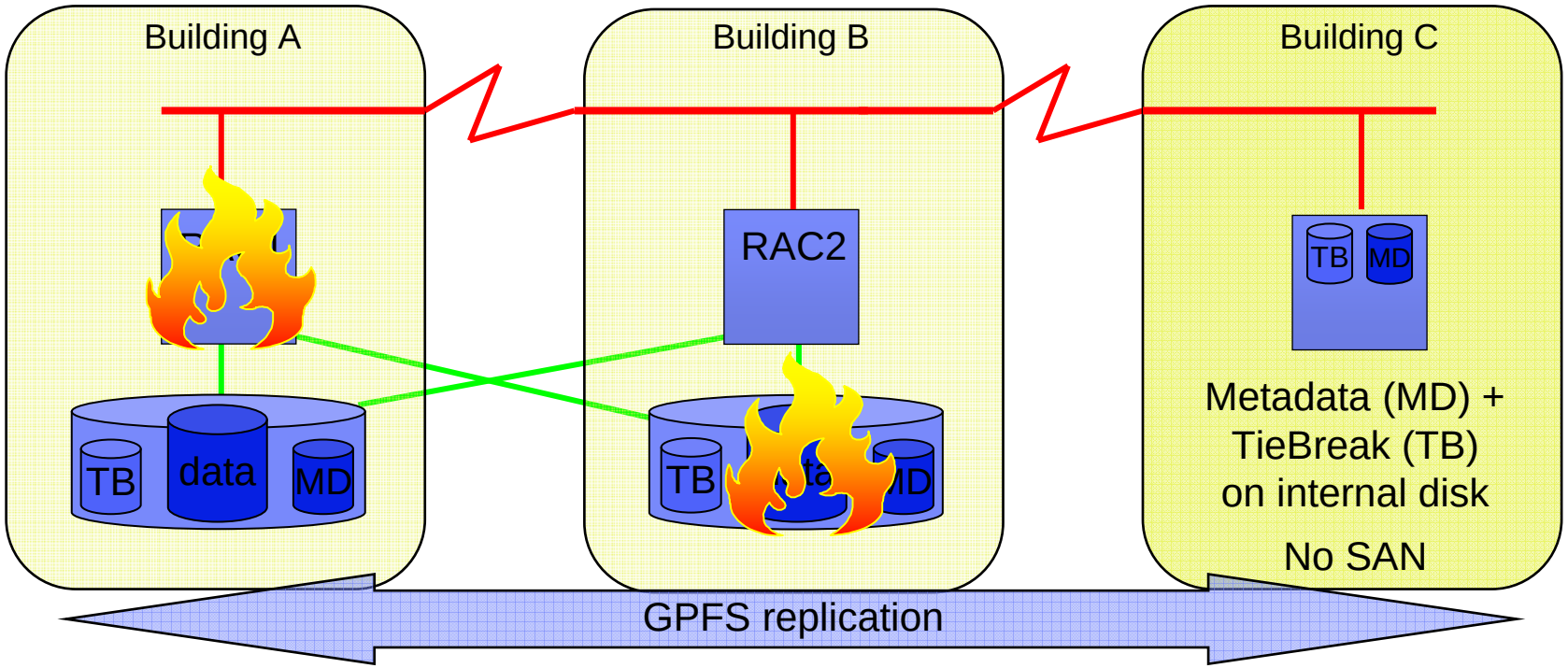
DR & GPFS replication

- Example with a Oracle RAC cluster database
  - RAC DB on top of GPFS replicated cluster file system
  - Three independent sites are required by GPFS replication



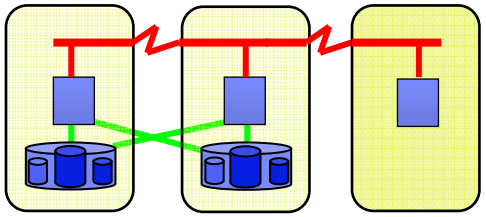
DR & GPFS replication

- Storage failure on one site
  - No outage, GPFS manages
- Node failure
  - No outage, Oracle RAC manages

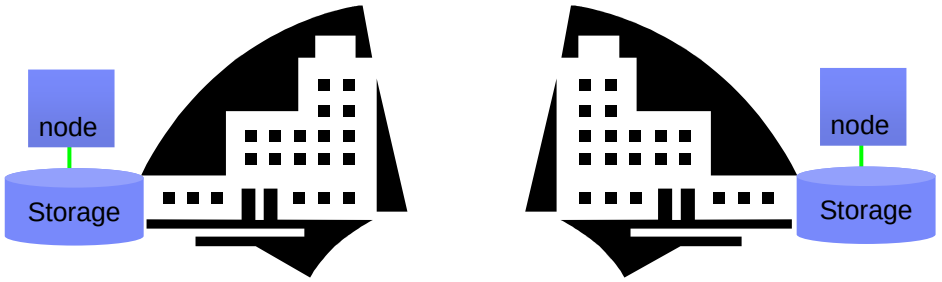


DR & GPFS replication

- GPFS provides a data replication mechanism that protect against disks or bay failure
  The replication is done by GPFS itself and not by the AIX LVM
- + Advantages
  - Storage failures are managed automatically by GPFS, with no database outage
  - Node failure (RAC instance) does not stop the DB service
  - No outage in case of storage or site failure
- - Disadvantages
  - Three sites are required
- RTO = 0
- RPO = 0

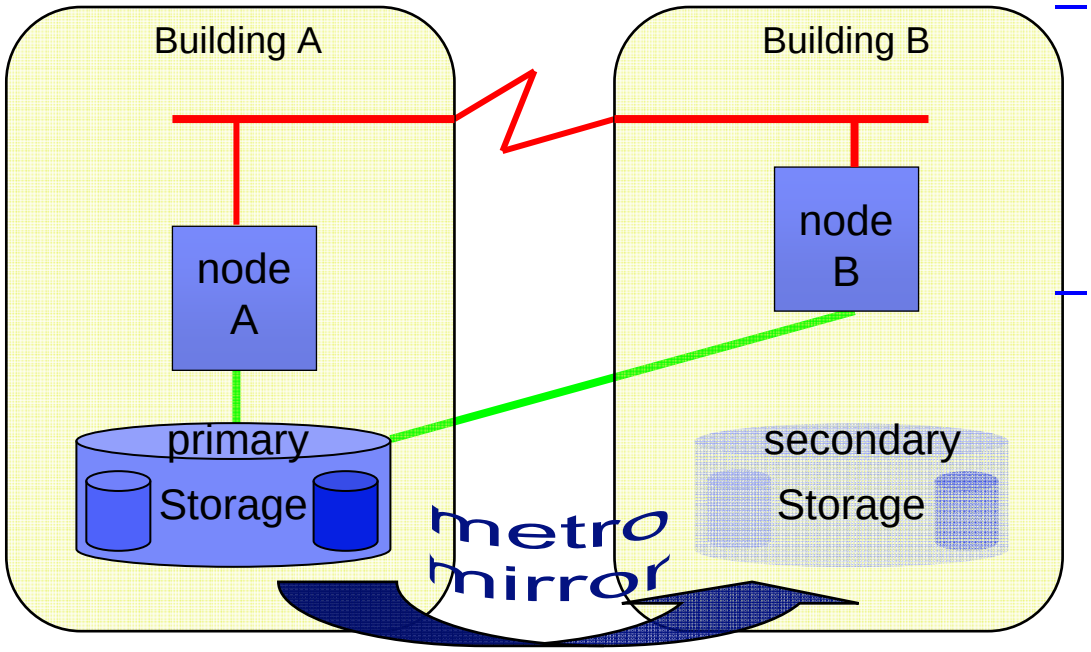


- Data Mirroring can be achieved at different levels
  - Operating System layer
    - LVM mirroring, GPFS replication
  - Storage layer
    - Metro Mirror, Global Mirror



DR & Storage Metro Mirror (Linked Cluster)

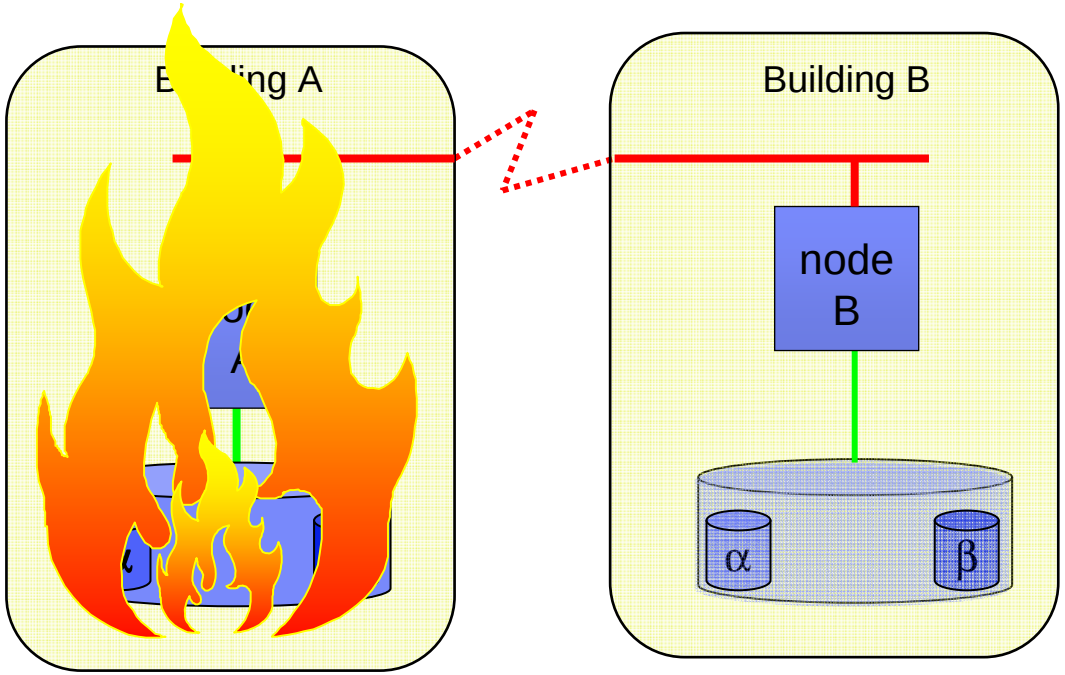
- The mirroring is done by the storage itself
- AIX and the application is not aware about this mirroring



- The secondary storage is a synchronous copy of the primary (Metro Mirror activated)
- The disks of the secondary storage are not visible from the nodes

DR & Storage Metro Mirror (Linked Cluster)

- Upon disaster, the service stops, and has to be restarted on the secondary site / storage
- IBM Power HA System Mirror EE can automate

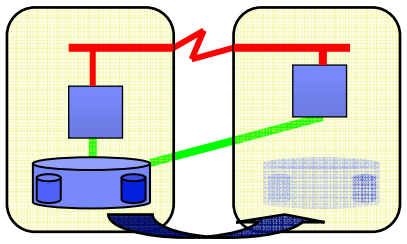


– Storage administration :
  - Stop the metro mirror
  - Assign the secondary LUNs to node B
  - For node B, the storage is back again. The node is not aware that it is a copy
  - Restart service
  - The secondary storage becomes primary
  - Metro mirror have to be resynchronized after disaster is repaired

DR & Storage Metro Mirror (Linked Cluster)

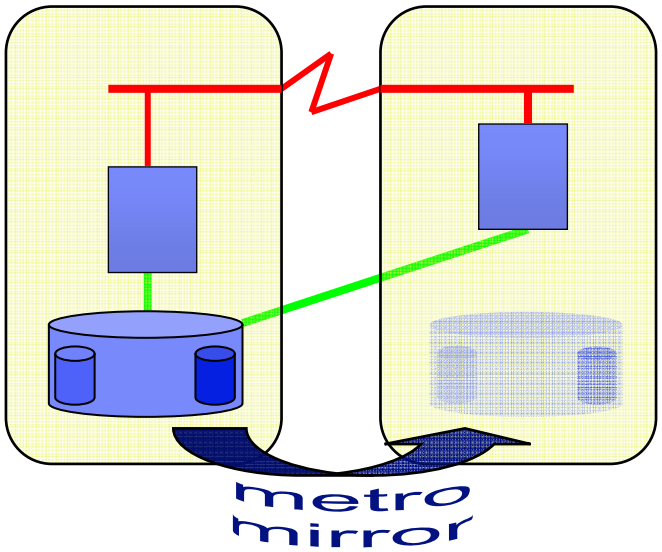
- + Advantages
  - Easy to setup
  - Common way to protect globally all the data (system also)
  - Defined at the storage level, for all kind of applications and all operating systems (AIX, Linux, i, Windows, etc…)

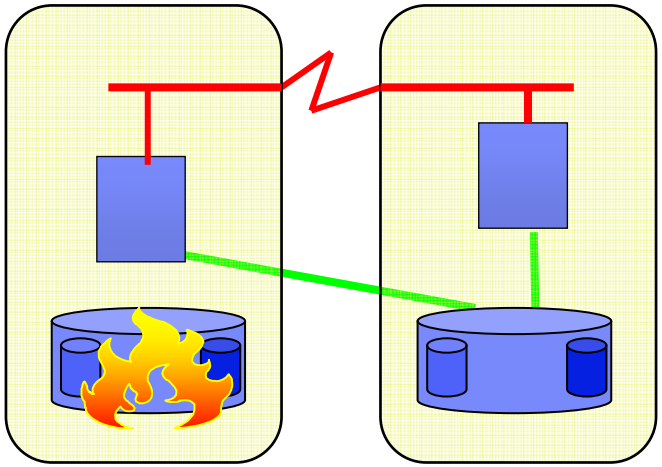
- - Disadvantages
  - In case of primary storage failure, there is an outage. Then, the application is restarted on the secondary



HyperSwap Support by AIX-PowerHA

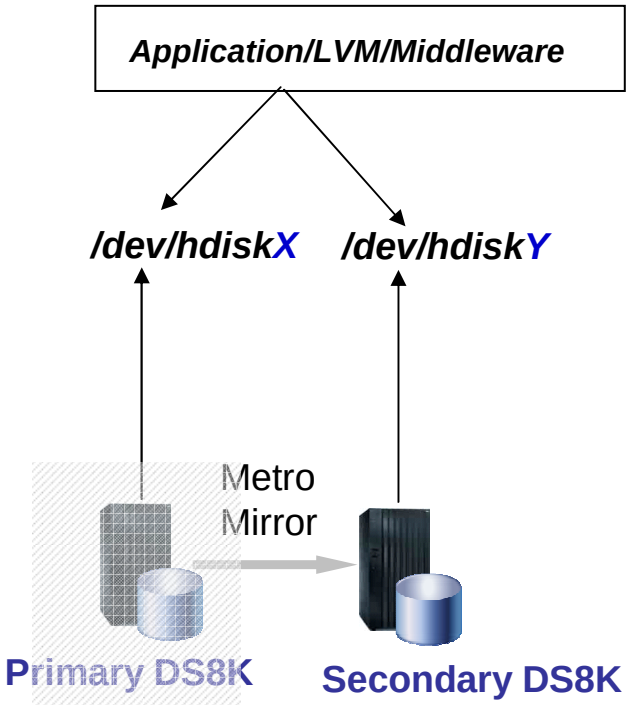
- HyperSwap coordination across hosts and sites
  - Planned or unplanned HyperSwap
  - Multi host synchronization
- Consistency group management across DS8K systems
- Typical swap times less than few seconds
- HyperSwap Support for critical system disks
  - Rootvg
  - Paging device
  - Dump Devices
  - Repository disk
- Disk Grouping Support
  - Groups disks and establish consistency groups
- Support for both AIX LVM and Raw disks
  - Disk or VG preparation
  - Disk Error handling
  - Oracle can be deployed with LVM or ASM disks
- VIOS: NPIV only. No vSCSI support
- Requires DS8800 or above storage



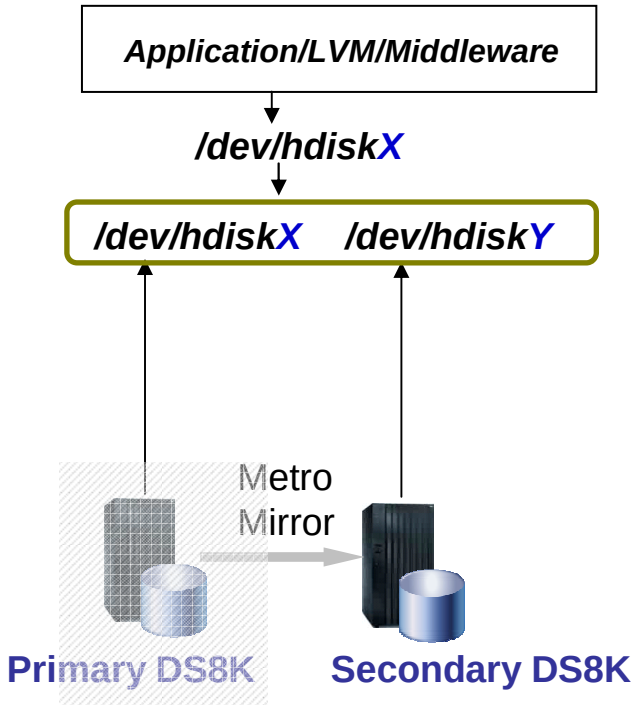


HyperSwap Support by AIX-PowerHA

HyperSwap device configuration transparent to application
Applications continue to use the devices as usual - storage switching is fast …seconds



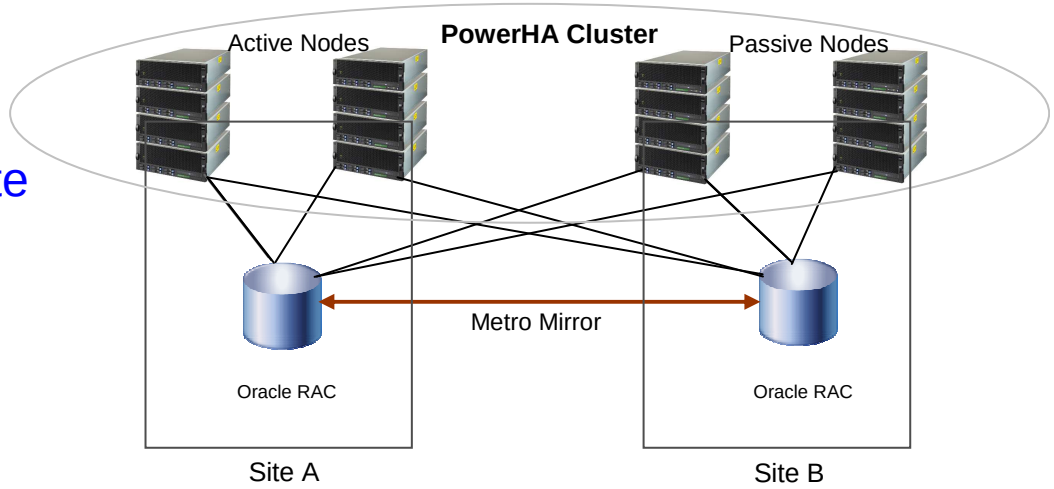
Traditional Metro Mirror Cluster



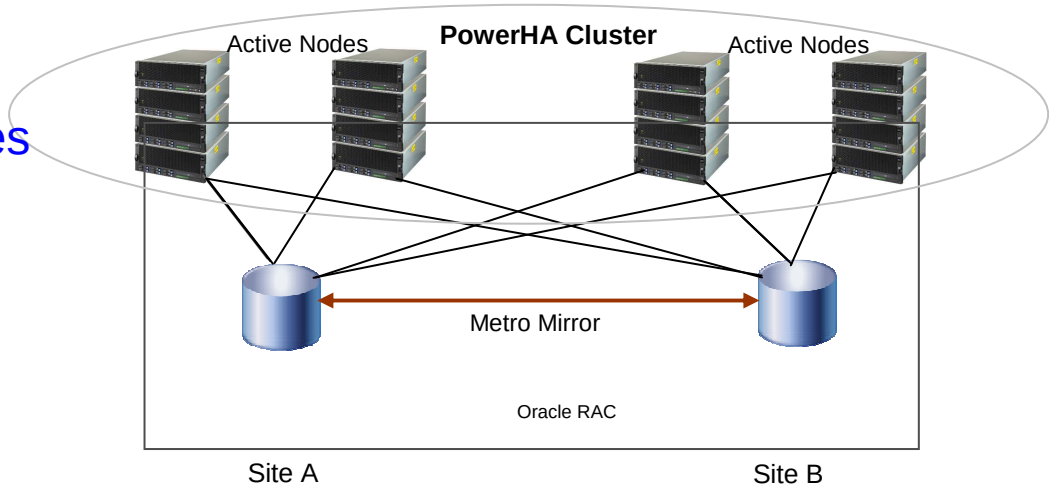
HyperSwap Cluster

Continuous Availability With PowerHA HyperSwap

- Active-Passive Sites
  - Active-Active workload within a site
  - Active-Passive across sites
  - Storage continuous availability across sites



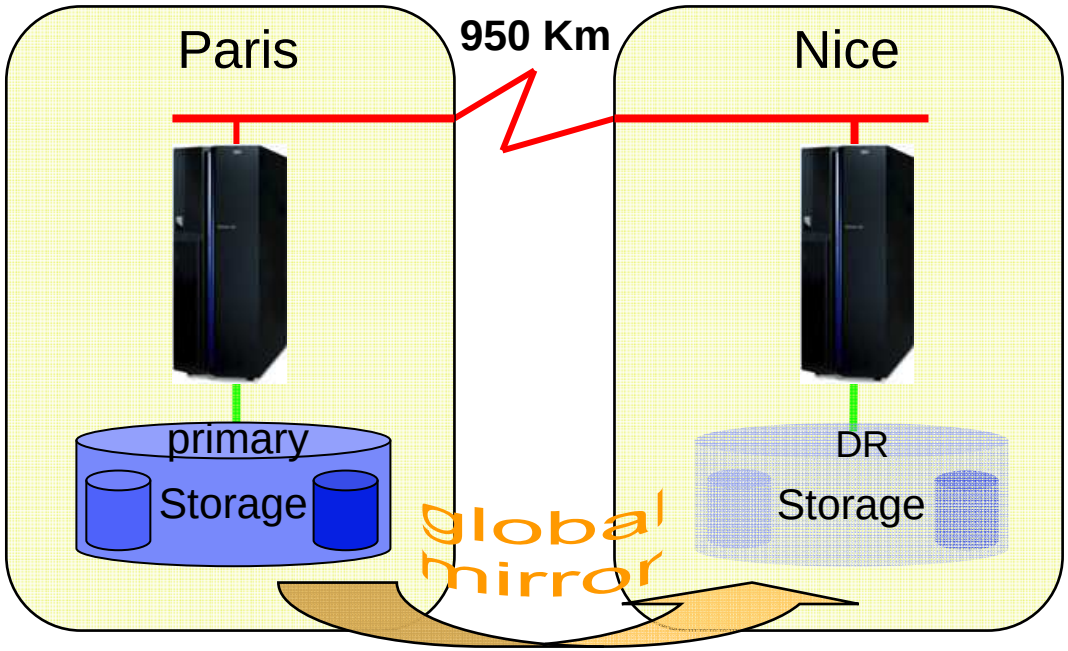
- Active-Active Sites
  - Stretched clusters only
  - Active-Active workload across sites
  - Continuous availability
  - Oracle RAC long distance deployment



- Mirroring at the AIX layer (LVM, GPFS)
  - + No outage in case of any storage failure
  - + Limited administration to recover
  - – Valid only for the data using LVM or GPFS, but not for database with raw devices (Oracle ASM mirroring still possible)
  - – Not applicable for non AIX application
- Mirroring at the Storage layer (Metro Mirror)
  - + A single mechanism for all OS (AIX, Linux, …) and all applications (databases, others), including system disk (rootvg boot on SAN)
  - – Outage in case of the primary storage failure
  - – Storage administration to recover after failure

Long distance Disaster Recovery

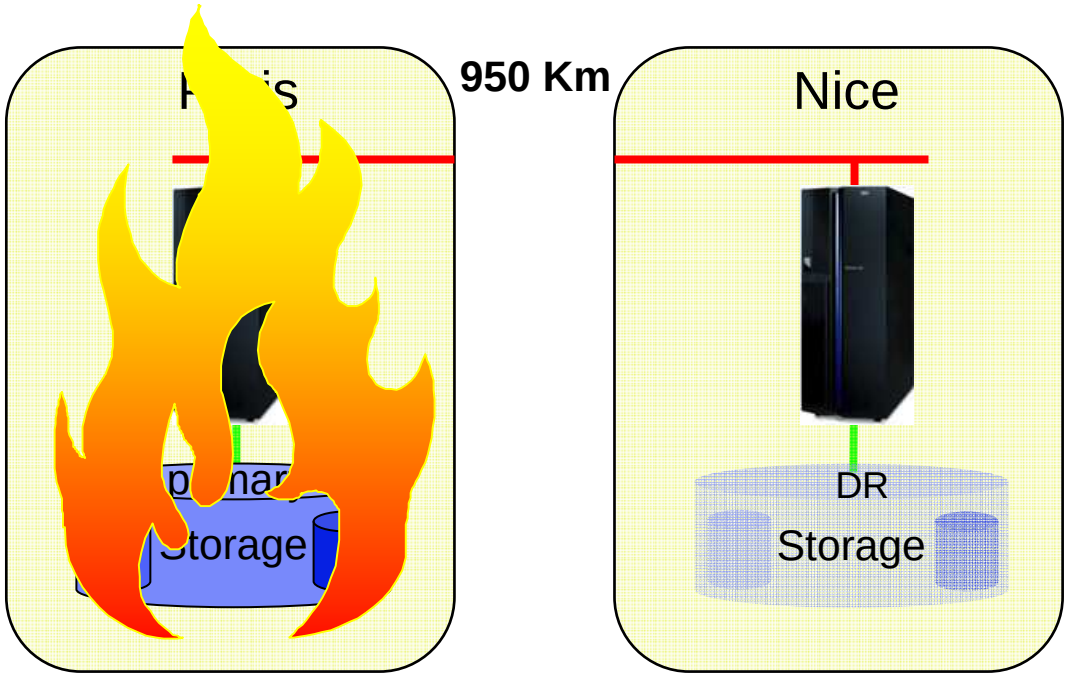
- Use of Global Mirror (asynchronous mirroring)
- No performance degradation due to distance



- Same hardware on both sites
- Same I/O cards, same slots
- Boot on SAN
- Power HA System Mirror EE deals with Global Mirror

Long distance Disaster Recovery

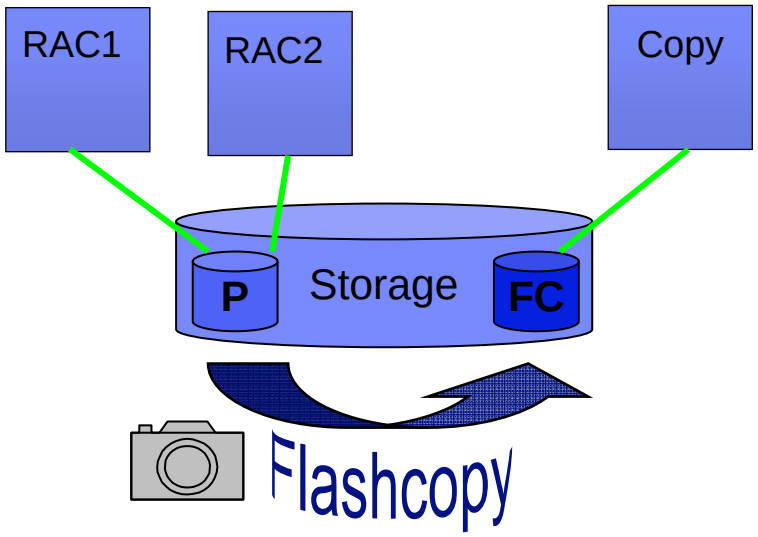
- Global Mirror keeps data integrity
- Manual startup on DR site



- The same LPARs are restarted on a copy of the system disk
- Not easy to reuse DR hardware in normal operations
- RTO can be days
- RPO depends on the line

Point in time copy (Flash Copy)

- More for online backup than for HA, but can help to increase the availability by reducing the planned outages (for backups)



  - Provides an **instant** copy of a whole set of data
  - For Oracle, suspend IO, make flash copy, resume IO (1 second only)
  - Used for time consumming tape backup
  - Used for cloning a production environment for dev / test
  - Compatible with AIX LVM, GPFS, Oracle

Disaster Recovery means distance ?

- Not necessarily
- A long distance is mainly for earthquakes, floods.
  - Risk has to be evaluated. Can be very low.
- Distance leads to performances concerns
  - 100 Km = 1 to 2 ms of added latency round trip
- Amadeus has secured a single site in Munich (called little pentagon)
  - Data centers isolated, building can resist a plane crash and atomic bomb
  - Power supply, networks, flood control, physical access are completely separated

Disaster Recovery means distance ?

- In Montpellier, IBM has a secure second data center, at 20Km away from the main site
  - For efficient disaster recovery, without the problems due to long distance
  - Good balance between risks and constraints

High Availability & Disaster Recovery

- Keep it simple
  - Processes are part of HA
  - If the HA/DR architecture is too complex, it can be too difficult to manage with normal administrators skills
  - Avoid the necessity of expert skills to manage a disaster situation
- Refrain from wanting everything
  - RTO=0 (no outage)
  - RPO=0 (no data loss)
  - Long distance DR without performance concerns
- Find the good balance
  - Business needs vs costs and complexity

High Availability & Disaster Recovery

- Think globally, for all your IT
  - Avoid to use a different HA or DR solution for each of your application
- Create and update HA/DR scenarios and procedures
  - Do not improvise during a crisis