Storage agnostic end to end storage information for long distance high availability

Vijay Kumar Shankarappa

Rupesh Thota

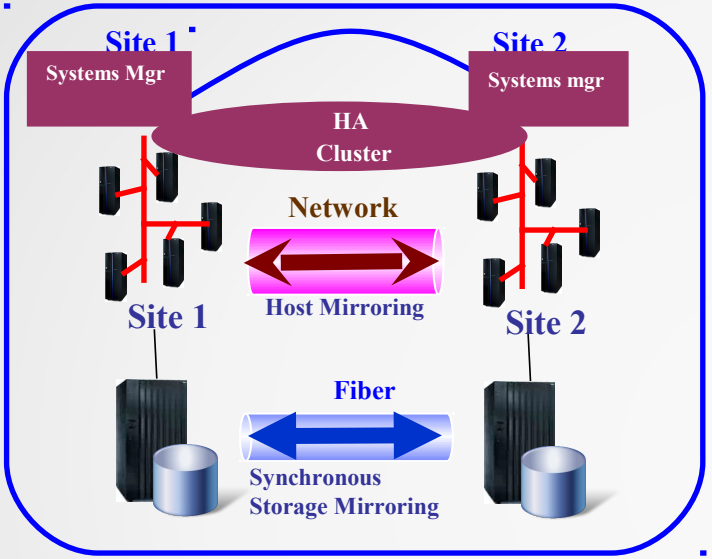
IBM India

Contents

1) High availability/Recovery solutions
2) Long distance availability challenges
3) Proposal

Cluster High availability vs VM Restart High availability

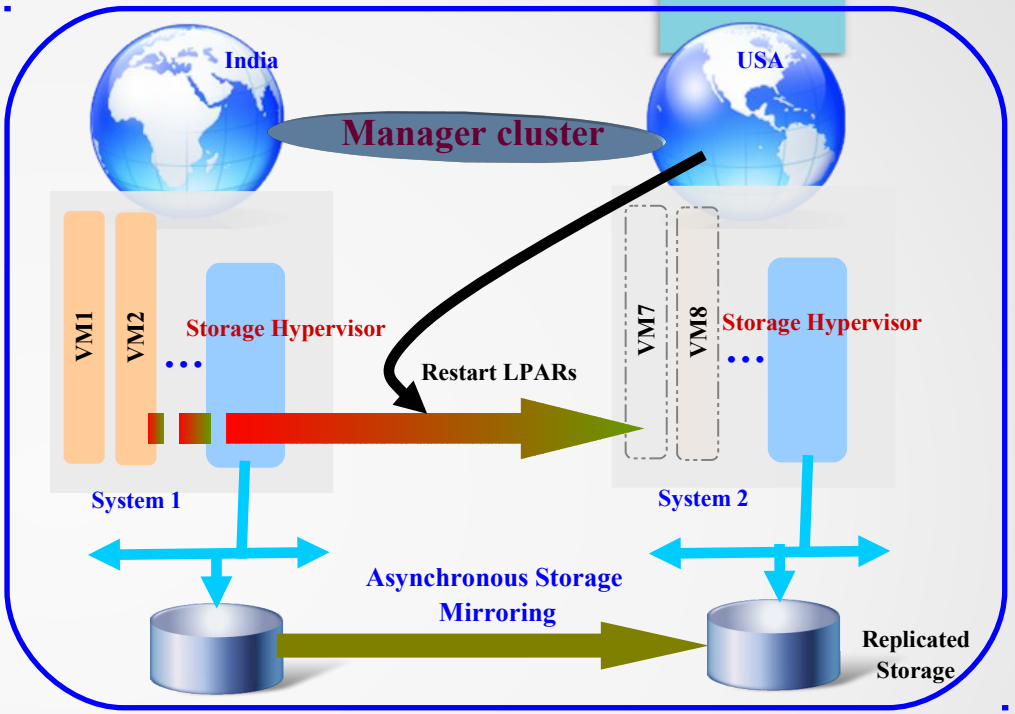
Cluster HA Solution



Sites < 100 km

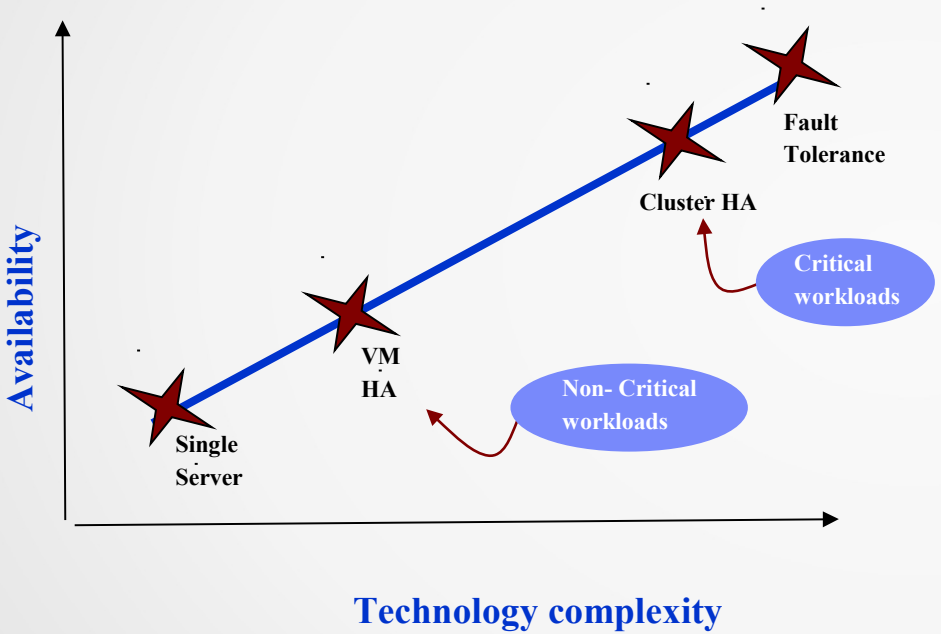
VS

VM Restart HA Solution



Long distance

Economical and Simplified HA models



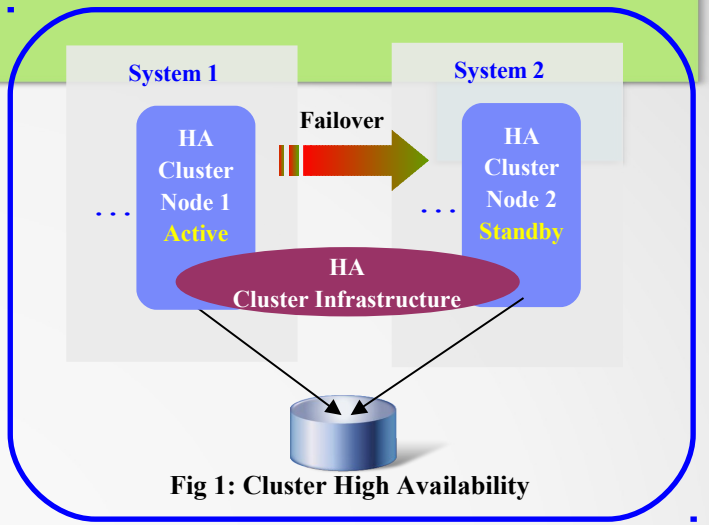


Fig 1: Cluster High Availability

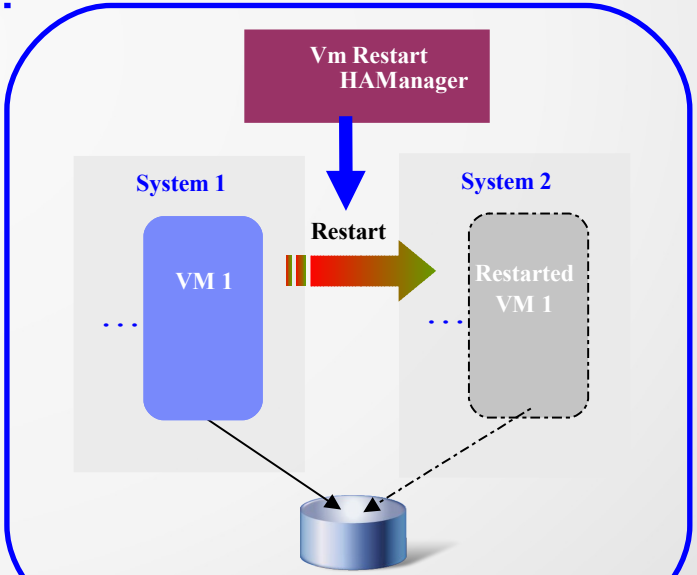


Fig 2: VM Restart High Availability

Comparison of Cluster fail-over versus VM restart

	Cluster Availability	**VM (Restart) High Availability**
Workload Startup time	Faster	Reinit & Reboot of VM
Cluster Administration (Network, Storage, Security)	Yes	No
Error coverage	Comprehensive (inside VM monitors)	Limited (outside VM Monitors)
Deployment Simplicity	Needs setup in each VM	Aggregated deployment outside VMs
License & Resource Savings	No	Yes
Workload Type protected	Critical	Non-Critical
Validation	Hard	Easily audited
Flexible failover policies	No	Yes

VM restart HA: Challenges

How to identify physical storage in use by a particular VM or a set of VMs which needs to be highly available ?

How to help admin configure physical storage data replication - Peer to Peer Remote Copy (PPRC) pairs across multiple sites ?

No SCSI standards to deal with PPRC.

PPRC is vendor specific implementation today.

Availability solutions hard to meet in a repeatable manner.

End to End flows in VM restart HA

Virtual storage: Storage hypervisors on a host system present physical storage accessible to it to the virtual machines via NPIV - Nport ID virtualization or virtual SCSI (Backed by a File or a logical volume or a complete disk or a Clustered file system Logical unit)

- **First task : Need to collect all the virtual storage mapped for a VM or VM group.**
- **Next task is to find the backing physical storage (disks)**
- **Next task is to help admin configure storage mirroring on alternate site based on consolidated virtual/physical storage information collected by storage hypervisor at source site.**
- **Next task is to validate/find the physical and virtual storage availability on alternate site.**
- **Initiate site movement by admin in case of real incident.**
- **Cleanup the virtual mappings/VMs once DR site is back up.**

Comparison: Methodology for storage data collection in VM HA

In Band - storage hypervisor	Out of band – external orchestrator/manager/agent
Single device agnostic code/module to fetch data, abstracts vendor/product/revisions	Custom code/modules for each storage vendor/product/revisions
Efficient design since hypervisor owns the virtual device mappings for a VM , gets the required data for only those backing devices quicker	Go to storage hypervisor to fetch the virtual mappings, and then query for each backing device based on storage type.
More robust as it also understands/handles MPIO for the storage it provisions.	Less scalable as it turns out to be multiple commands/scripts to get the MPIO and collate the virtual and physical mappings.
Easily extensibile with growing feature-set in virtual storage hypervisor.	Need to write new code to accommodate any changes in storage hypervisor features.

In Band data collection by storage hypervisor – SCSI standards and status as of today

Page 80h for appliance/array identifier

Page 83h to get Logical unit identifier

Vendor specific pages to get globally unique device/volume identifiers used for mirroring.

Vendor specific pages to read PPRC (copy relationships and status)

Changes for each storage vendor, models, revisions.

Dependency on vendor tools/api/cli to get the same info.

SCSI standards proposal

- **T10 SPC4 r361 onwards, proposal on Vital product data parameters**
- **http://www.t10.org/cgi-bin/ac.pl?t=f&f=spc4r36l.pdf**
- **Device constituents Page Code: 8Bh , section 7.8.5**
- **Currently optional in SPC4**

Table 581 — Device Constituents VPD page

Bit / Byte	7	6	5	4	3	2	1	0
0	PERIPHERAL QUALIFIER			PERIPHERAL DEVICE TYPE				
1	PAGE CODE (8Bh)							
2	(MSB)			PAGE LENGTH (n-3)				
3								(LSB)
	Constituent descriptor list							
4 •••	Constituent descriptor (see table 582) [first]							
	⋮							
••• n	Constituent descriptor (see table 582) [last]							

SCSI standards proposal

Constituent Device Identification VPD page code (83h) - part of Page 8Bh

If the designator type is 3h (i.e., NAA identifier), this format is compatible with the Name_Identifier format defined in FC-FS-3.

The Name Address Authority (NAA) field defines the format of the NAA specific data in the designator.

Table 597 — NAA DESIGNATOR field format

Bit / Byte	7	6	5	4	3	2	1	0
0	NAA							
1	NAA specific data							
...								
n								

The Name Address Authority (NAA) field (see table 598) defines the format of the NAA specific data in the designator.

Table 598 — Name Address Authority (NAA) field

Code	Description	Reference
2h	IEEE Extended	7.8.6.6.2
3h	Locally Assigned	7.8.6.6.3
5h	IEEE Registered	7.8.6.6.4
6h	IEEE Registered Extended	7.8.6.6.5
All others	Reserved	

Globally Unique identifier of a disk can be defined using this format using a 16 byte designator and used for configuring mirroring.

If NAA is 6h (i.e., IEEE Registered Extended), the 16-byte fixed length DESIGNATOR field shall have the format shown in table 602. The CODE SET field shall be set to 1h (i.e., binary) and the DESIGNATOR LENGTH field shall be set to 10h.

Table 602 — NAA IEEE Registered Extended DESIGNATOR field format

Bit / Byte	7	6	5	4	3	2	1	0
0	NAA (6h)				(MSB)			
1	IEEE COMPANY_ID							
2								
3				(LSB)				
4	VENDOR SPECIFIC IDENTIFIER							
...								
7								(LSB)
8	(MSB)							
...	VENDOR SPECIFIC IDENTIFIER EXTENSION							
15								(LSB)

The IEEE COMPANY_ID field contains a 24-bit canonical form OUI assigned by the IEEE.

The VENDOR SPECIFIC IDENTIFIER field contains a 36-bit numeric value that is assigned by the organization associated with the IEEE company_id in a way that combines with the VENDOR SPECIFIC IDENTIFIER EXTENSION field to make the entire DESIGNATOR field (see table 602) unique.

NOTE 72 - The EUI-64 format includes a 40-bit vendor specific identifier. The IEEE Registered Extended format includes a 36-bit vendor specific identifier.

The VENDOR SPECIFIC IDENTIFIER EXTENSION field contains a 64-bit numeric value that is assigned by the organization associated with the IEEE company_id in a way that combines with the VENDOR SPECIFIC IDENTIFIER field to make the entire DESIGNATOR field (see table 602) unique.

SCSI inquiry page/constituent to hold PPRC data

A new inquiry page to be defined to hold PPRC data in SCSI specification

- To have all the relevant mirroring information.

1) PPRC state

- Is full duplex
- Is duplex pending (Copy to establish the pair in progress)
- PPRC pair is suspended

2) PPRC status

- Status of copy operations along with partner volume id

3) Mirrored array info :

- model, vendor, revision info

Reference : IBM FICON/ESCON attachment specification has defined a page C0 to hold such data.

Takeaway: Design point

====> VM restart availability solutions easier to implement in a repeatable and storage agnostic manner if:

a) Globally unique disk identifiers are used in PPRC pairs,

b) PPRC partners and status info is standardized via SCSI inquires,

c) Adopted by all storage vendors uniformly.