IBMCHAMPION

Belisama

Session: Power migrations (101)

Antony Steel

Power Systems

Thanks to:
Various customers who allowed testing

#include <std_disclaimer.h>
These notes have been prepared by an Australian, so beware of unusual spelling and pronunciation.
All comments regarding futures are probably nothing more than the imagination of the speaker and are IBM Confidential till after GA.

Why migrate?

Belisama

- You may find yourself facing the following challenges… If so, a migration may be the best solution
  - Upgrade of physical environment (Server / Storage) (new features, performance, reduce costs…), including MES upgrade
  - Upgrade of operating system
  - Moving Data Centre / location within DC
  - Moving in new workloads (for example migrating Linux workloads to Power)
- This session is an introduction to migrating workloads, including between Data Centres and will:
  - Examine the various ways to migrate Power systems, both online and offline – working a customer case if we have time
  - These options include using features such as:
    - LPM; Enterprise Pools; mksysb; Simplified Remote Restart; as well as in place upgrades of the OS (NIM)
  - Required OS and firmware levels for the various platform levels will be discussed.
  - Planning considerations for the upgrade and/or the migration are also addressed.

Options available

Belisama

- Options
  - Concurrent migration – ie without an outage
    - LPM (with or without flexibility of Enterprise pools)
      Can be done in conjunction with Live Kernel Update (AIX)
  - Non concurrent migrations – during a scheduled outage
    - In place MES upgrade
    - Restore a mksysb
    - Inactive LPM (to get around changes in VLANs, storage, etc)
    - Use *alt_disk_copy*
    - Use *alt_disk_mksysb*
    - Use *nimadm* to create new OS image from mksysb
    - Move LUNs to another system
    - Use Simplified Remote Restart

First steps

Belisama

- Tips
  - Planing, Planing and more planing
    - Is System clean? Check error logs, ifixes, is TL/SP complete?
  - Backup everything, even your backups!
  - Tools: FLRT, SSIC
  - Set allow migrations with inactive VIOS (flexibility with LPM in troubled times)
  - Doing MES upgrade (topic in itself, ping me for info)
    - MES is an acronym used by IBM which stands for Miscellaneous Equipment Specification.   Any server hardware change, which can be an addition, improvement, removal, or any combination of these. The serial number of the server does not change. (in most cases)
  - Use AIX live update to update AIX to be compatible with new frame / Firmware

https://www14.software.ibm.com/webapp/set2/flrt/power

https://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss

OS Planning – some considerations

Belisama

- The following needs to be considered in the planning phase
  - Starting with AIX 7.2 TL3, the default SMT level is SMT8
  - AIX 7.2 TL3 on POWER9 lpars will boot in SMT8 regardless of the compatibility mode set, older versions in SMT4.
  - AIX 7.2 TL3 on POWER8 lpars will boot in SMT4
  - LPM will migrate with whatever the active SMT mode is
  - POWER9's performance benefits in SMT8 mode are noted across a wide variety of workloads
  - Hardware performance is extremely robust in SMT8 mode
  - HMC v9r1m930 and POWER9 now allow for bidirectional concurrent migrations. Prior to these levels, migrations could only go in one direction at a time.
  - As of HMC v9r1m920 it's possible to use LPM to migrate LPARs between servers that have the same serial number although they must be different machine types and models. Each machine must be on a different HMC as well. This allows you to use LPM to migrate LPARs to a new server when using the same serial number as part of an upgrade; e.g., when upgrading from an E880 to an E980 using serial number protection.

OS Planning (cont)

- Other considerations that may impact performance after move
  - Placement of LPARs (see *lsmemopt* on HMC)
  - Correct capacity planning for actual workload on new system
  - There are some known issues with migration to Power9 to do with VPM folding / Virtual I/O performance (see tips white paper in reference)
  - Plan your entitlement and virtual processor count as rules have changed for Power9 (a good starting point is to use peak utilisation to determine VP Count and average to determine entitlement)
  - Check the IBM Power Virtualisation Best Practices for details on tuning changes
  - Heritage configurations
    - Always wise to review OS tuning, storage layout / configuration, virtualisation layer for any legacy design. Be aware that some of your tuning may have been to better handle older / slower devices. Particularly relevant for systems that have been upgraded over the years.
    - Engage a Partner or Lab Services to assist with an Audit or a benchmark.

Power Enterprise Pools

Belisama

- The capability to freely move core and memory activations between servers in a pool provides a new degree of architecture. Application areas to consider using Power Enterprise Pools include:
  - Rebalancing server capacity - Mobile activations can be moved between servers to make best use of core and memory resources. Mobile activations can also be use to temporarily relocate resources for period end processing or full-scale performance testing.
  - Live Partition Mobility (LPM) - When using LPM for scheduled maintenance or partition relocation, mobile activations can be shifted from the LPM source server to the LPM target server. The IBM Lab Services LPM Automation Toolkit automates the movement of the LPARs using LPM and activates the target mobile resources and then deactivates source mobile resources.
  - PowerHA clusters–If the primary server fails, you can move mobile activations from the primary server to the backup server.
  - Disaster Recovery-With a single HMC network across sites mobile activations can be moved between servers in different sites

Warning

- There are a number of supported and “un-supported” or “Non-preferred” ways of moving, migrating or cloning AIX

https://www.ibm.com/support/pages/node/670623

  - Supported
    - LPM
    - Cloning via restore of a mksysb onto the new system via tape, file, usb, VIO repository, NIM.. Note: Set recover devices to “NO” when it is restored to keep the ODM clean.
    - Cloning via alt_disk_copy. Assign a LUN to the current system, clone AIX with the alt_disk_copy, then move the disk to the new system. (there are flags to remove devices from the new ODM)
    - Use alt_disk_mksysb to install an existing mksysb onto a new LUN, which is then assigned to the new system. Or use nimadm to update / apply fixes at the same time
  - Non-preferred
    - These methods include making a copy of a rootvg (via storage) or moving a disk from one system to another. In many cases it will work, but IBM cannot issue a blanket statement as the device tree in the rootvg will not necessarily match the new system.
    - Using the setting *chdev -a ghostdev=1 -l sys0* you may be able to design a workable solution
      - 1 will delete customised ODM when booting in a different LPAR / System is detected
      - 2 will change rootvg VGID and its disks PVID

Worked example

- Data Centre migrations …
  - Is this a common problem (migrating workloads from one DC to another)?
  - Procedure
  - Perform
  - Post migration review / lessons learnt

Data Centre (DC) migrations …. is it a common problem?

Belisama

- Migrating workloads from one DC to another
  - Out grow existing DC
  - DC closes
  - Review or renew of contracts
  - Customer moves
  - Consolidating DC / hardware upgrade
    - Move p5-p8 to p9
    - Move “Out of support” AIX levels to in support
  - Upgrade of hardware / move to private cloud
    - Migration of mixed Power to POWER8/9 for traditional workloads / HANA
    - Migration of mixed x86 linux to Linux on Power
    - New linux workloads on POWER
  - ….

DC Migrations Procedure - Analyse customer requirements

Belisama

- Review the current infrastructure
  - RTO, RPO (and classify environment, use for planning)
  - Availability and redundancy configuration
  - Availability and redundancy requirements during the migration
  - Sufficient CPUs / licensed cores
  - Tools – FLRT; LPM verification; iostat
- What constraints around timings
  - Outage windows – fit into plan
  - Change control – prepare to avoid delays
  - Access/security requirements
- Managing the process
  - How many customer teams?
  - Fit your migrations in with other parts of the organisation (you are unlikely to be the only migration)
  - Coordination and communication issues
- Testing
  - Reproducible tests
    - I/O performance for storage (# mirror copies, Different storage systems / sites)
    - Simulated load for LPM

DC Migrations Procedure – Audit environment

Belisama

- Conduct an audit of the current environment
  - What Operation Systems and patch levels
  - What virtualisation in place
  - What management & monitoring
  - What applications & dependencies
  - Current network design
  - Current storage layout
- Audit demonstrate that current environment meets customers standards?
- Audit expose any problems for migration?
  - We found both storage and network inconsistencies during the audit and were able to highlight some OS exposures that could be patched prior to migration

DC Migrations Procedure – Review options

Belisama

- Review possible LPAR migration options
  - What limitations do the customer requirements impose
    - Outage? Availability? Timing windows? Network / storage access?
- Review possible Storage migration options
  - What limitations do the customer requirements impose
    - Outage? Availability? Timing windows? Bandwidth and data volumes
  - What are the options generally for DR?
    - Storage mirroring (sync, async, hyperswap)
    - OS mirroring (LVM, GLVM)
    - Application mirroring (Log shipping)
- For each option review:
  - Timings
  - back-out plans
  - OS / firmware / application dependencies

DC Migrations Procedure – develop plan / coordination

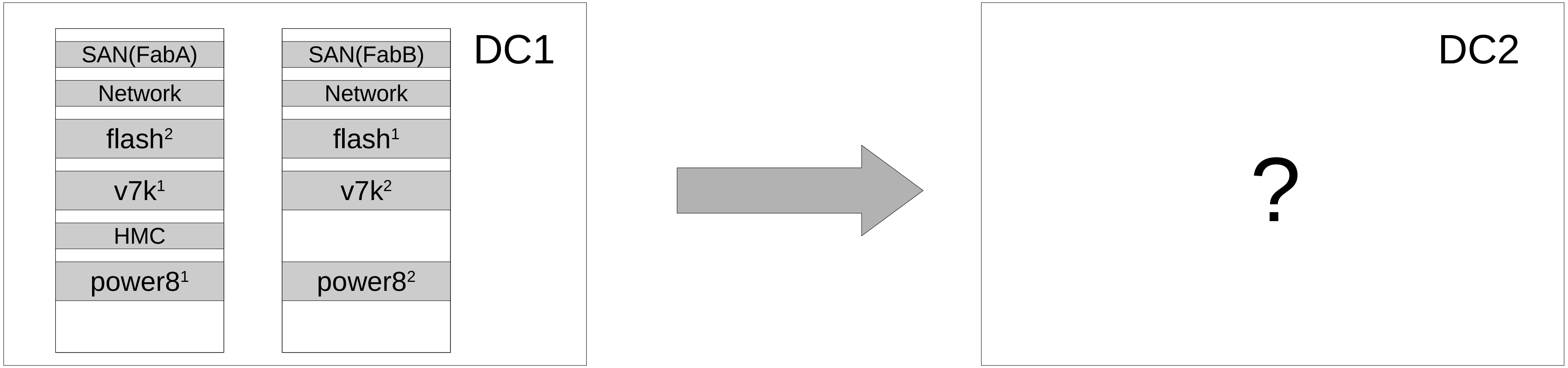
- Develop plan
- Coordination / communication across teams
- What can go wrong? What will go wrong? Back-out / Rollback / contingency plans

DC Migrations - Perform

- Perform migration(s)
- Post migration review
- Lessons learnt

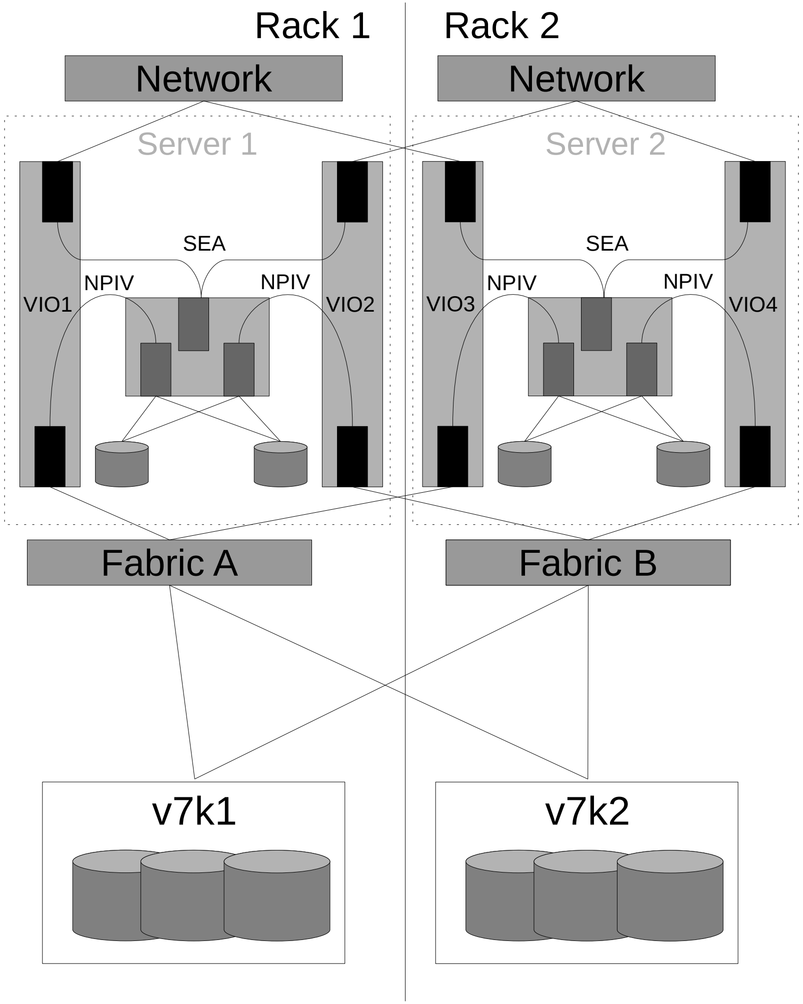
Example

- Customer requirement
  - Move operations from existing DC to new DC in same city
  - 2 racks of power and storage
    - Fully virtualised; mixed patch level (OS / f/w)
    - LVM mirroring between racks
  - Issues
    - Needed to apply some patches (HMC; VIO; AIX; firmware)
    - Some mirrors inconsistent
  - No outage highly desired; work on production only during restricted hours on Sunday morning
  - Backup no redundancy and ran nightly (therefore physical move during the day)



Customer starting position

Belisama



Options to migrate OS instances

Belisama

- Hardware
  - Need to maintain customer’s availability requirement (2 copies of storage; 1 standby server)
  - Need inter-site link to handle data replication within window mixed with system activity, while not impacting the system activity
- Options to migrate OS
  - LPM
    - Pro: No outage; easy to roll back
    - Con: HMC; Network; Storage visible; Network latency
  - Simplified Remote Restart
    - Pro: faster if memory access high; less requirements; easy to roll back
    - Con: Short outage
  - Create new instance at new site, test; move data LUNs for migration
    - Pro: Less network/storage requirements; Can perform OS updates in move
    - Con: Longer outage
  - PowerHA
    - There was no PowerHA in the customers environment

Used this option for one customer, nimadm migration for OS LUNs rsync for data LUNs Change freeze for OS, regularly sync to keep last sync to a minimum for cutover.

Migrations

Belisama

- Take the opportunity to perform AIX migration / install code or updates?
  - If migrating the OS: use NIM *nimadm*
  - If installing code etc, and OS level etc correct – AIX Live Kernel update
- *nimadm**
  - The process does the following:
    - Create a copy of the rootvg on a spare client LUN (similar to alt disk install *alt_disk_copy*)
    - Migrate the newly created copy of rootvg to the new version of AIX, installing additional filesets as required – While the system is still running on the current version of AIX, ie no disruption
    - At convenient time schedule a reboot of the client – choosing the new rootvg as the target.
  - The advantages are:
    - Minimum disruption on the client (just a reboot) – no outage for the migration
    - As the process of the migration is run from NIM, it primarily uses NIM resources (and some network bandwidth)
    - Easy back-out and debugging. Can roll back to the original copy of the rootvg and investigate the updated rootvg to resolve problem – In a fully virtualised environment can even assign the rootvg to another LPAR!

* Used for customer in Singapore AIX 5.3 on old h/w and migrated to POWER8

Migrate storage

Belisama

- Options to maintain at least 2 copies of storage
  - Use Storage subsystems mirroring/replication
    - Pro: consistent across all platforms; offload data copying so less impact on CPU
    - Con: Expensive; license requirements
  - Use LVM
    - Pro: cheap and easy (already use LVM mirroring within site)
    - Con: mirroring is fine for AIX/Linux..
  - Use GLVM
    - Pro: only requires IP link
    - Con: AIX specific
  - Log shipping / replication
    - Pro: Less bandwidth requirement
    - Con: Application specific

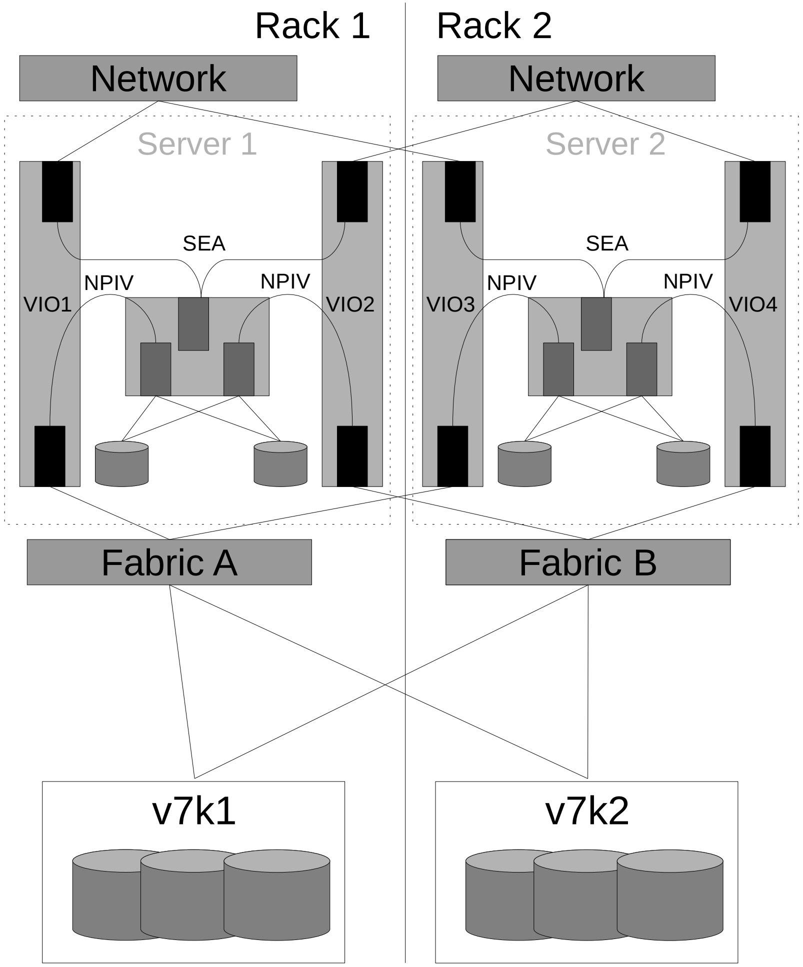
Options

Belisama

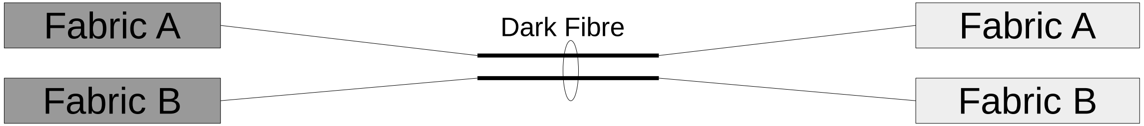
- Option 1: Lift and shift - Move 1 rack at a time
  - Con: As we shift each rack, we do not have redundancy
- Option 2: Storage replication – 2 way Storwize Remote Copy
  - Con: The cutover from source storage on site 1 to target storage on site 2 is disruptive
  - VMRM / Simplified Remote Restart could be used to reduce the failover time
  - IBM PowerHA-XD could be used to reduce the failover time
  - Need to use either native IP replication or stretch the fabric using FCIP routers or dark fibre
- Option 3: Storage replication – 2 way Storwize HyperSwap
  - Con: The cutover from source storage on site 1 to target storage on site 2 is disruptive
  - IBM PowerHA-XD could be used to reduce the failover time
  - Need to use either native IP replication or stretch the fabric using FCIP routers or dark fibre
- Option 4: Server replication – 3 way Logical Volume Mirroring
  - Compatible with Live Partition Mobility = online migration
  - Options to present LUNs from new DC to OS images (Stretch SAN; FCIP Routers; iSCSI; GLVM)
- Option 6: Introduce iSCSI as an option for mirroring
  - Con: Major change in operation of site, keeping 2 copies of data at all time difficult to manage

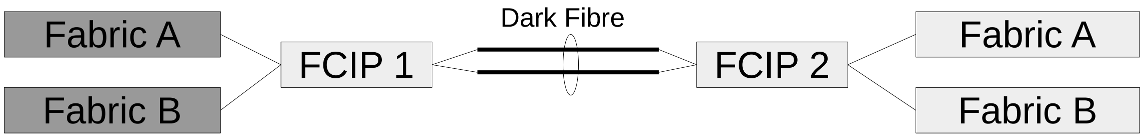
Extending the fabric

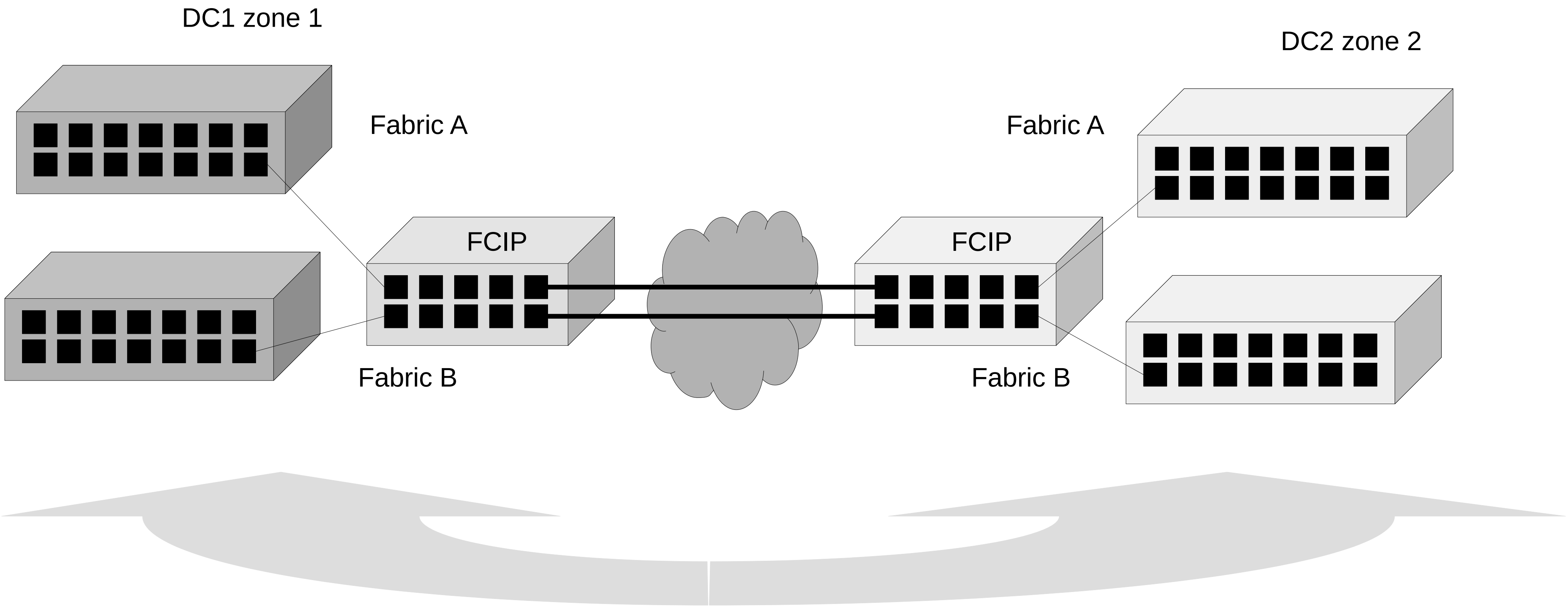
Belisama



- Few options need to be considered:
  - Leverage iSCSI host to storage protocol instead of the FCP protocol currently used
  - Leverage FCIP routers to interconnect FCP across both data centres without merging SAN fabrics (FCIP routers might only be used to simply prevent the SAN fabrics on each site from merging… they do not have to be used to encapsulate FCP frames into TCPIP frames)
  - Decision: 2 loan switches to maintain redundancy






DC1 zone 1
DC2 zone 2
Fabric A
Fabric A
FCIP
FCIP
Fabric B
Fabric B

Summary

Belisama

	iSCSI	FCIP Router	Cross site interconnect (dark fibre w/wo WDM)		
			SAN switch	Host Connection	Storwize Remote Copy
Compatibility with LVM	Yes	Yes	Yes	Yes	No
Disruptive	No	No	No	No	Yes
Redundancy	Yes	Yes	Maybe[1]	Maybe[2]	Maybe[3]
Performance	*	****	***	**	*****
Cost	$$	$	$$$$	$$$$$	$$$
LW SFPs	No	No	Yes	Yes	Yes
Migration steps	+++++	++++	++	+	+++
Stability/Resiliency	**	****	*	**	***

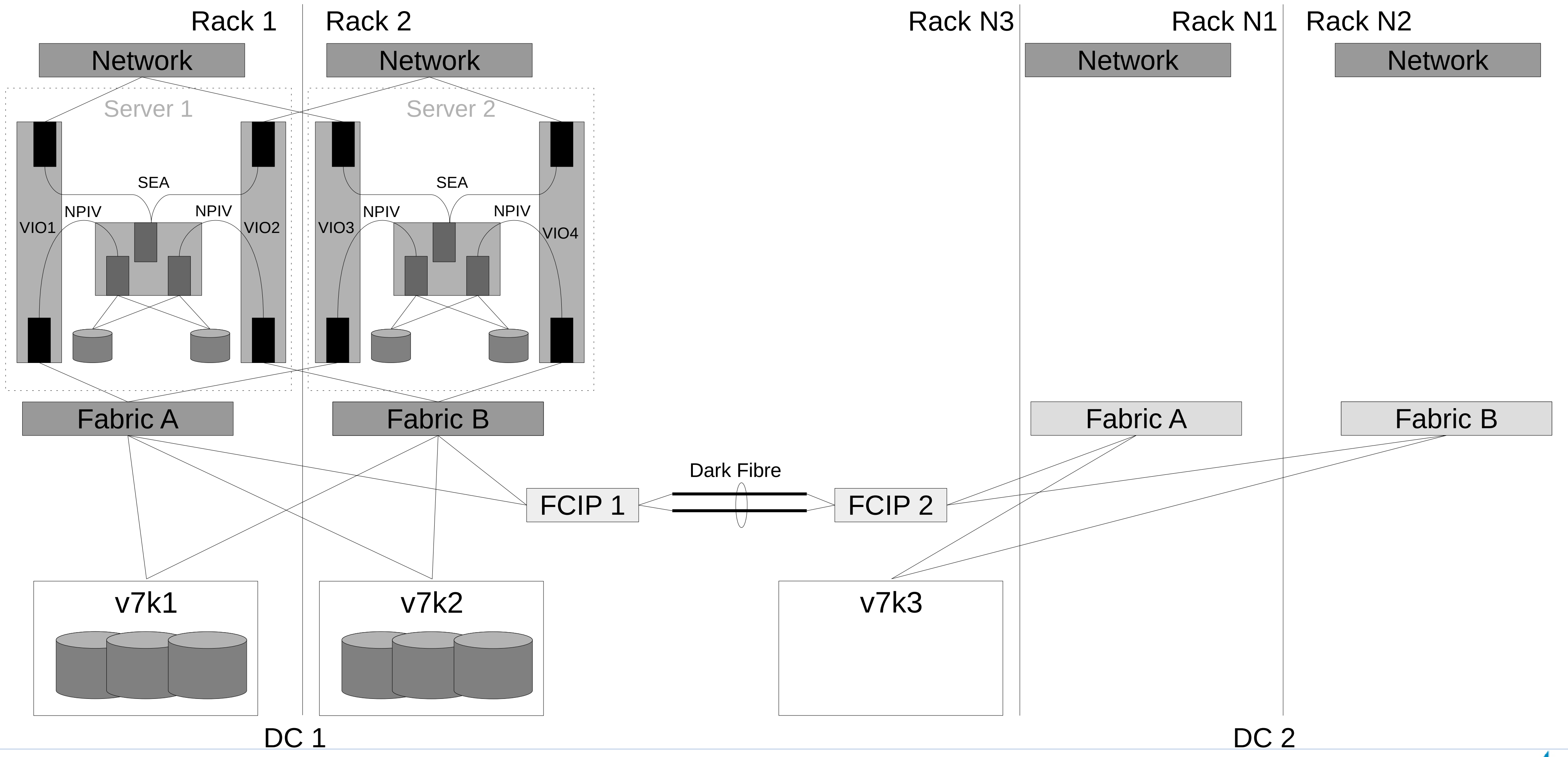
1. If you add a SAN switch per site during the relocation, than SAN fabrics A & B can be stretched, otherwise it would mean interconnecting both SAN fabrics.
2. If you have many dark fibres or active DWDM with many channel then all host connections on site 1 can be connected to the SAN switch B on site 2 and reversely all host connections on site 2 can be connected to the SAN switch A on site 1.
3. Recommendation is at least 2 ports per Storwize node canister

Considerations

- Why use FCIP routers
  - The key advantage is because of the integrated routing function that isolates 2 SAN fabrics while still interconnect them. That is connecting but not merging, so any topology change on one side does not propagate to the other side (for example a registered state change notification (RSCN) storm does not propagate outside the EX or VEX (routed) port.
  - It was easy to harvest both the SAN switch configuration and the Storage / LUN configuration to script a consistent build of the remote SAN switches and storage.
- Bandwidth requirements
  - Audit of environment
    - Audit of LVM and Filesystem layout provided the total amount of data to be moved
    - Performance audit of the virtual machines gave an indication of the amount of data updated
  - This gave us some flexibility, as depending on cable type, link speed and distance, we have a choice of SFPs and may require mulitple links across multiple protocols.
    - Depending on your fibre cable type, the link speed and the distance, you might need to use LongWave SFPs as well as WDM equipment. For this customer, using LW SFPs was a must, and depending on how many dark fibre links we were provided and our expected throughput, we may need to share one or multiple links across multiple protocols.

Extending the fabric

Belisama



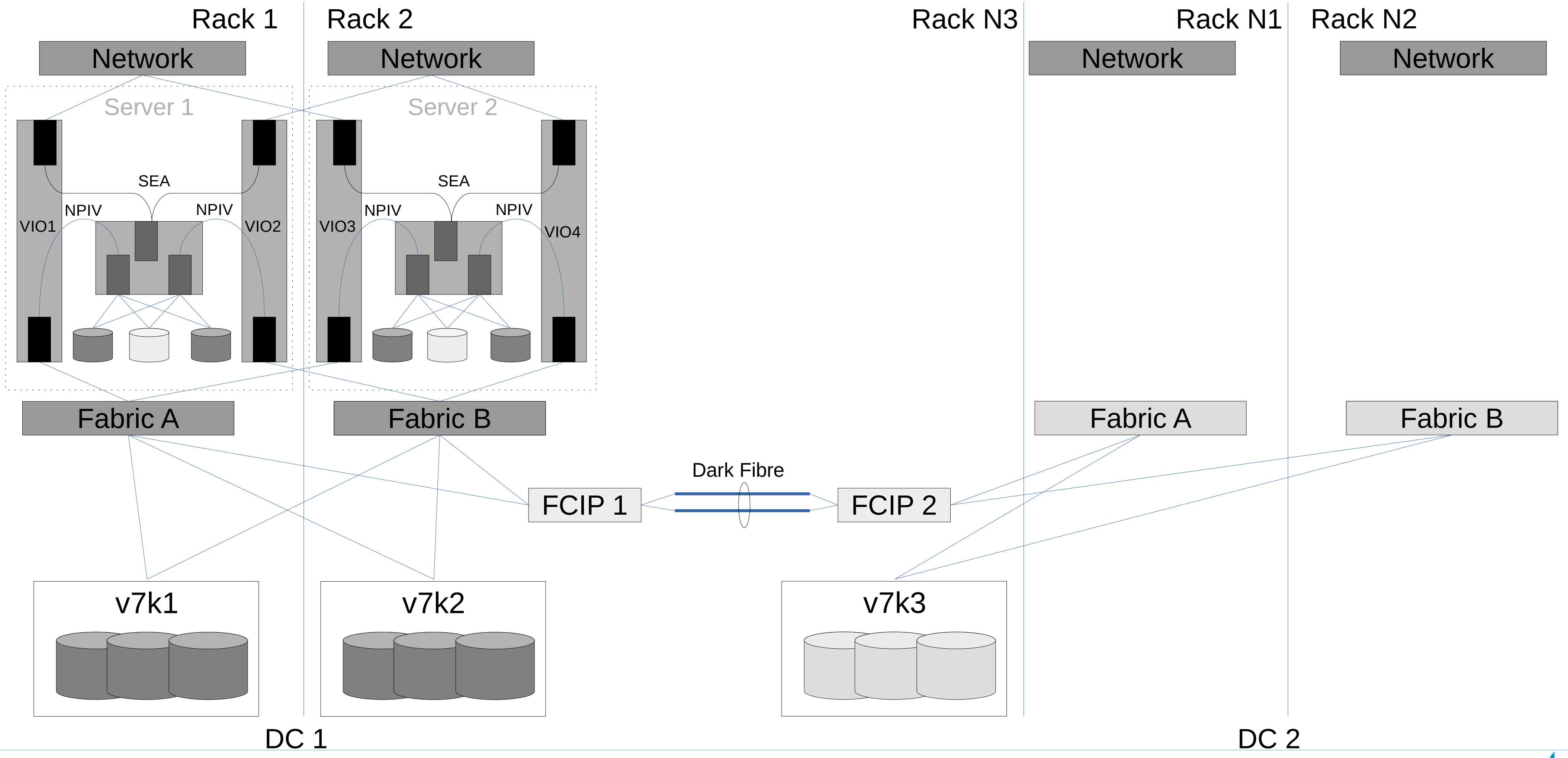
Build and testing

Belisama

- Confirm the configuration of the loan switches (copy and edit from the clients switches)
- Confirm the creation of all the LUNs on the loan storage (scripted).

Add third copy to all LUNs

Belisama



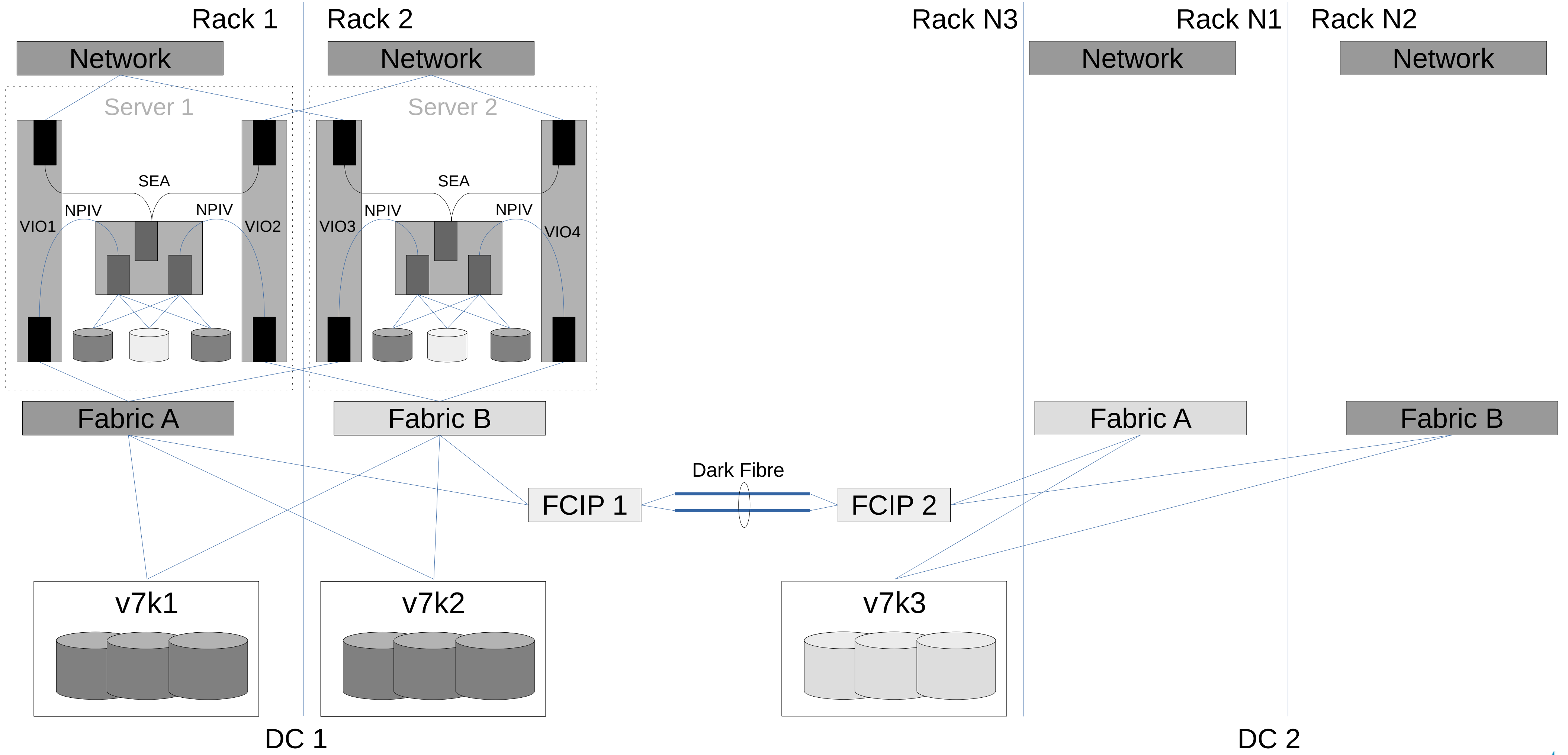
Build and testing

Belisama

- Add LUNs to test LPAR (testing configuration end to end)
  - Scripts were used on each LPAR to manage the 3rd copy of the mirror and manage each missing copy and re-sync as the underlying storage was removed and then became available during the hardware relocation
- Build I/O baseline performance for test LPAR with 2 way mirror
  - Create hosts and mappings on loan storage for test LPARs
- Add additional LUN(s) to each VG on test LPAR and sync (time to confirm performance)
  - Re-scan the storage, confirm presence and size of new LUNs
  - Run script to add LUNs to appropriate VGs and resync each LV
  - Build table of multiple test runs to establish mean and standard deviation
  - Confirm plans (time to resync data between successive rack moves and fit within backup windows, while reducing impact on day to day operations
- Test I/O performance for test LPAR with 3 way mirror and compare
- After sign-off add third mirror to all LPARs and script sync during quiet times.

Swap switches (Fabric A stable)

Belisama

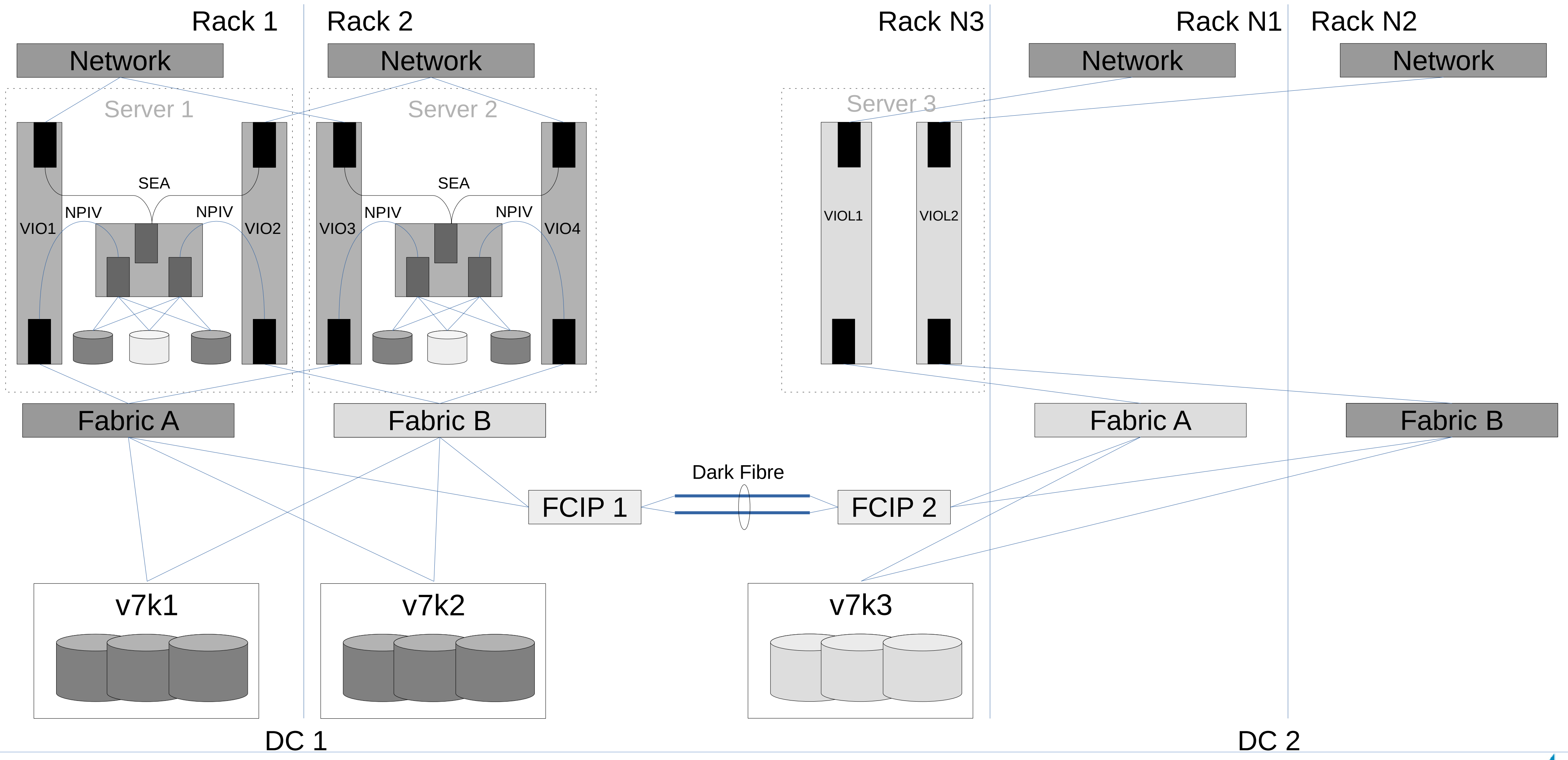


Test

- Confirm all paths, mappings and performance

Adding redundancy for virtual machines

Belisama



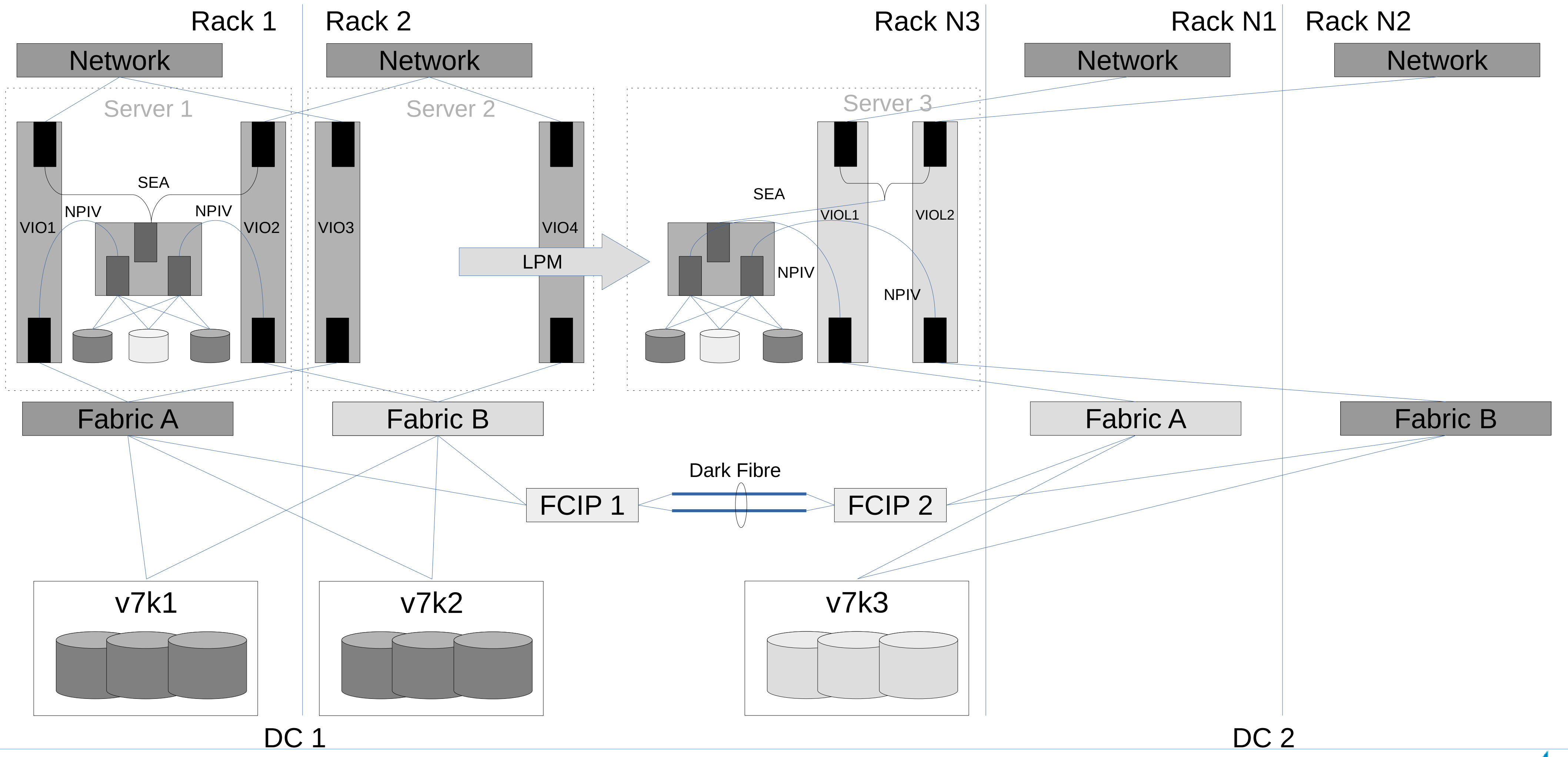
Test

Belisama

- After establishing a baseline for LPM (measure from Rack 1 to Rack 2 in DC 1 with simulated I/O and memory operations), test LPM from DC1 to DC2 and back again
- Test scripts for removing mirrored LUNs from specific storage units and re-adding / resync'ing. Confirm timings from different Host Servers / sites
- Moving test LPAR between Rack 1, Rack 2 and the loan rack also confirmed zoning / mapping and physical paths.

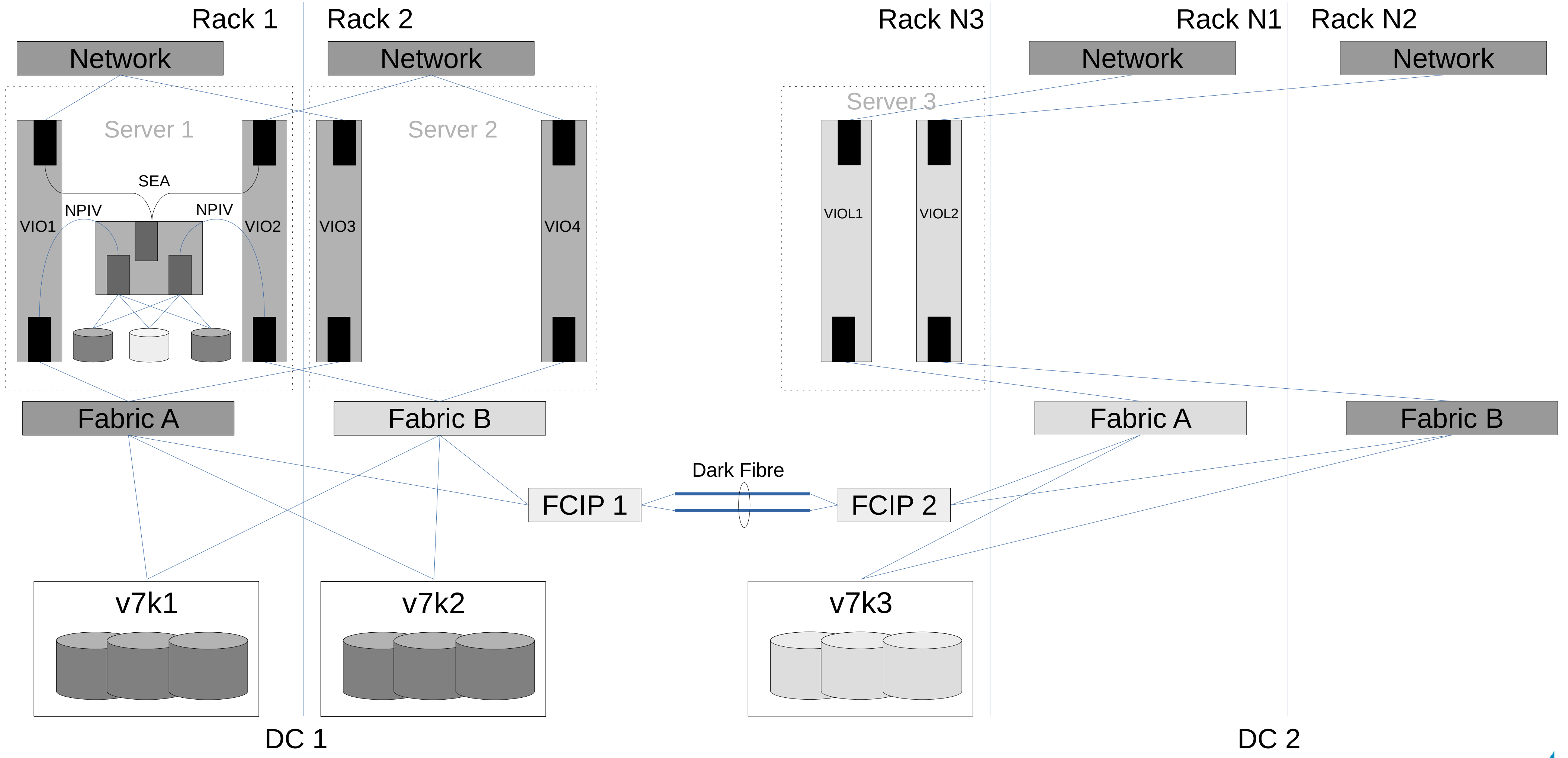
Adding redundancy for virtual machines

Belisama



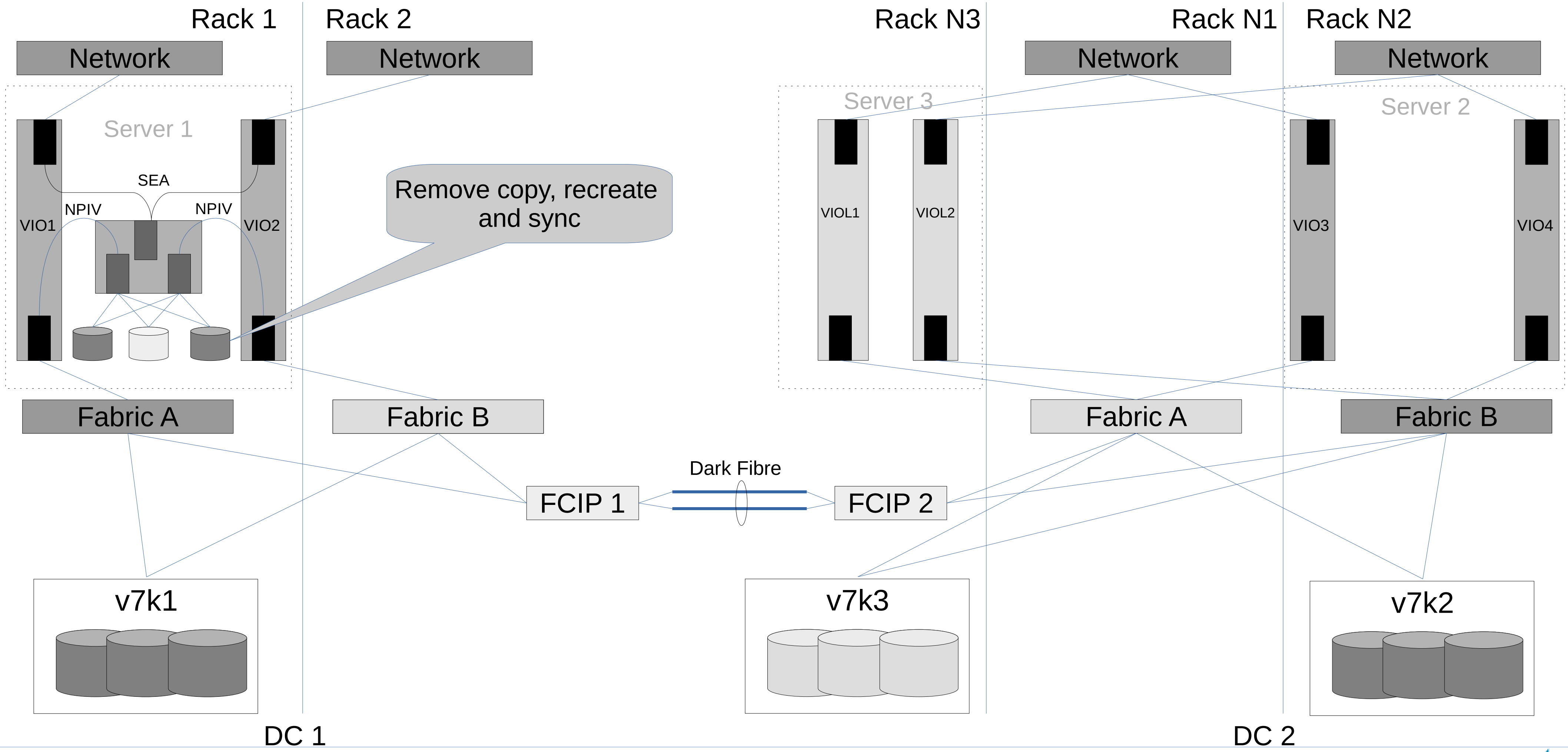
Remove workload from 2nd rack

Belisama



Move Rack 2

Belisama

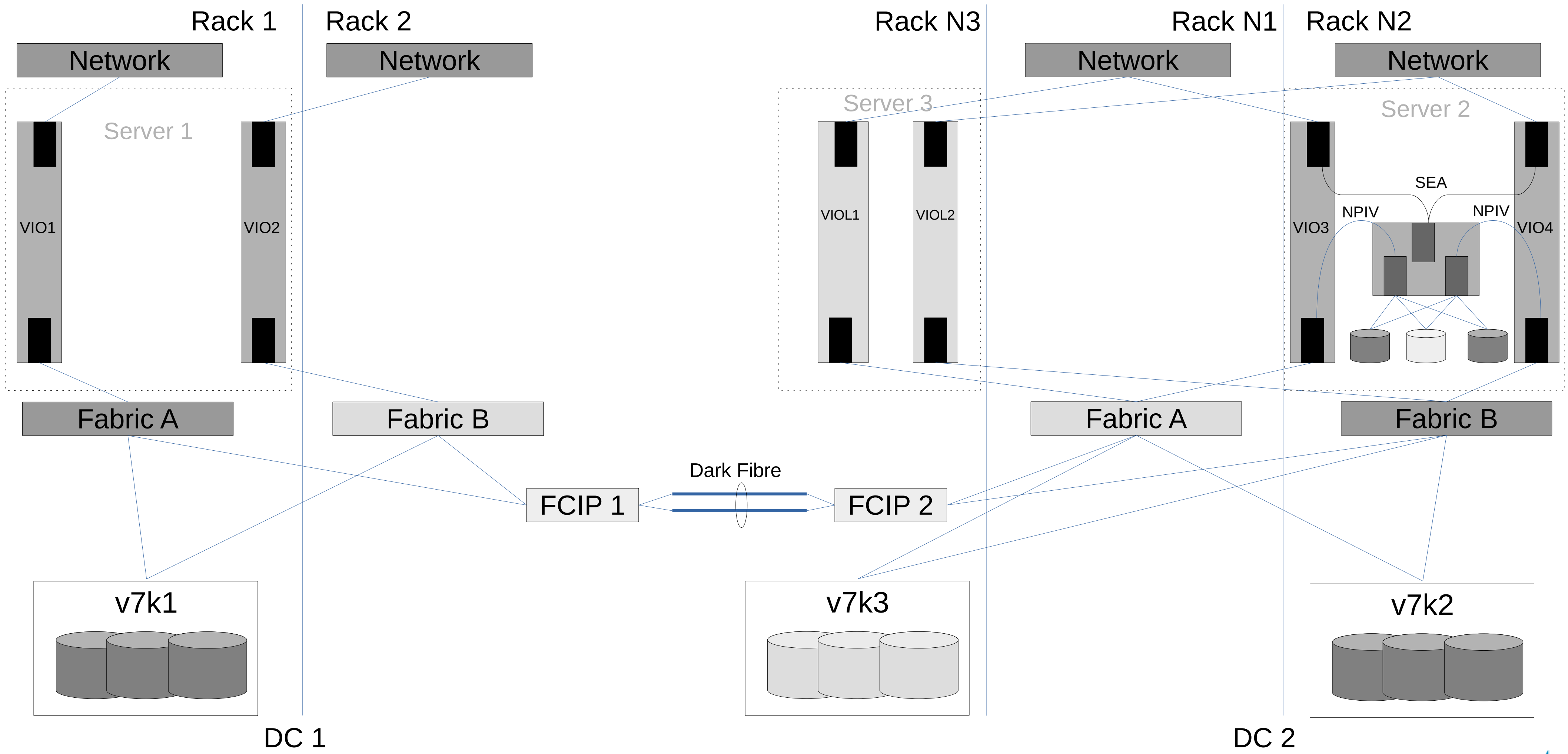


Test

- Confirm all paths
- Run hardware diagnostics
- Check all error logs
- Confirm test LPAR LPM to Rack 2 in DC2

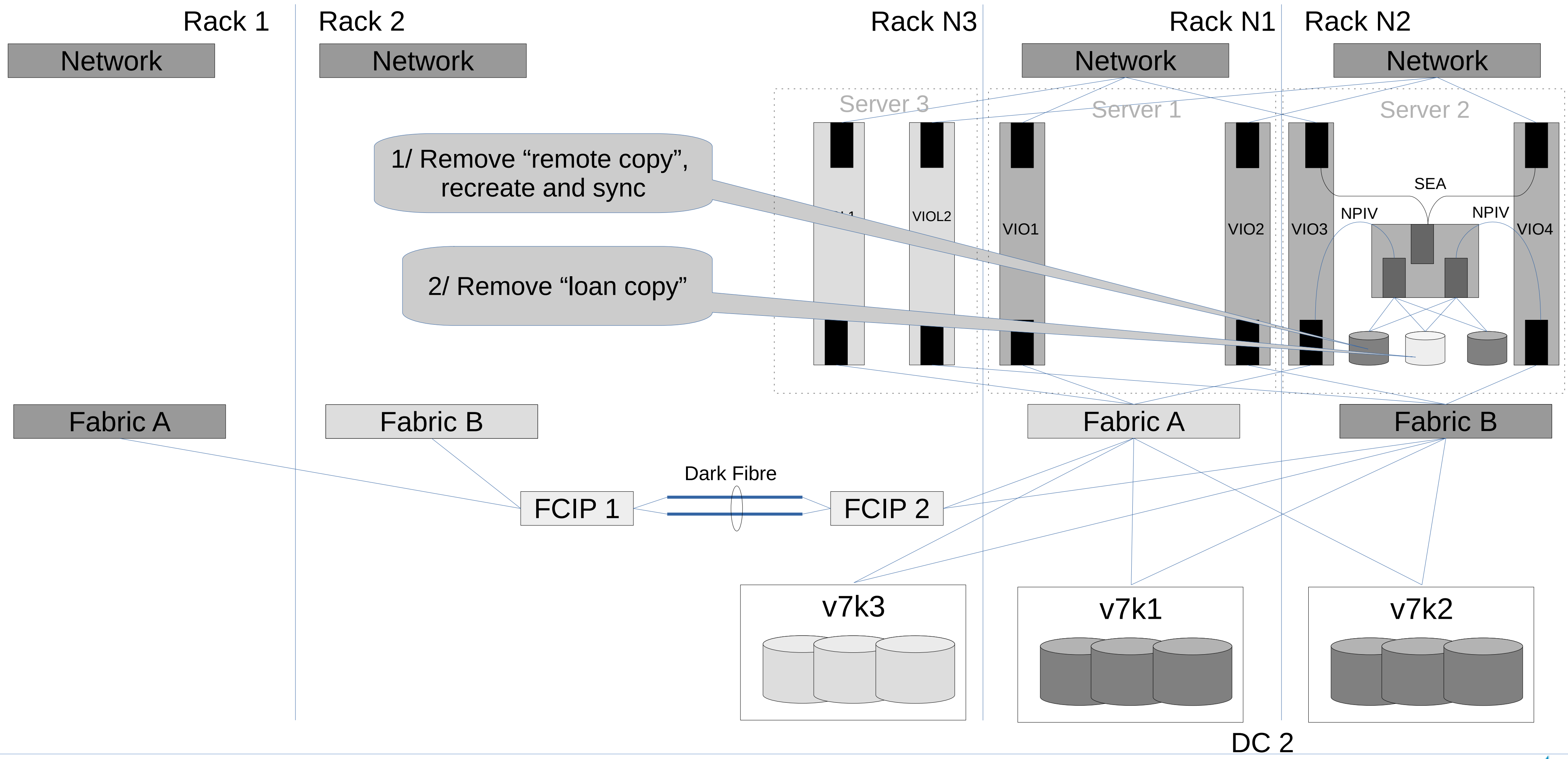
Move virtual machines

Belisama



Move rack 1

Belisama



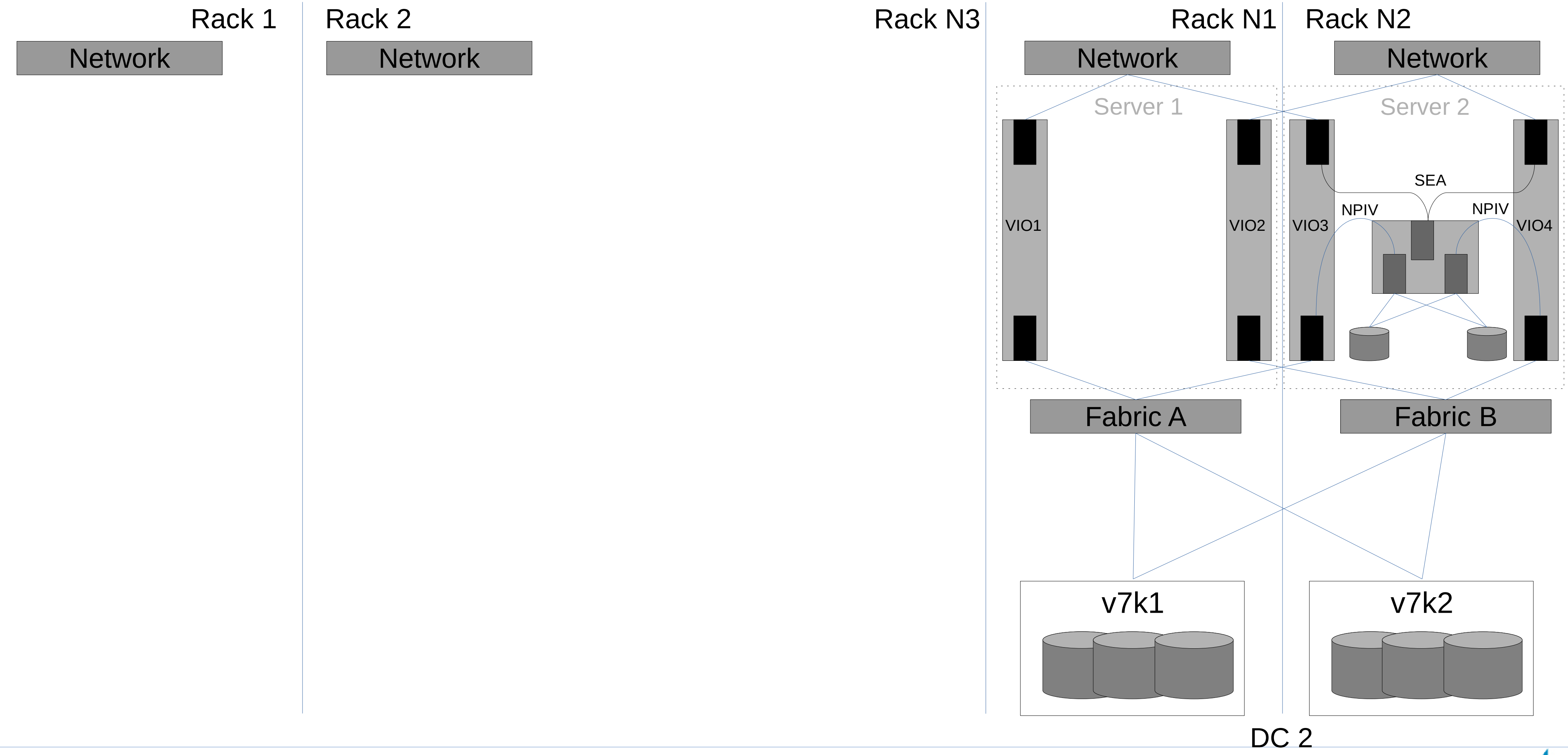
Test

Belisama

- Confirm all paths / LUNs available
- Confirm all Logical volumes syncd and available
- Confirm baseline performance and LPM between 2 new racks

Swap switches (Fabric B stable) and remove loan equipment

Belisama

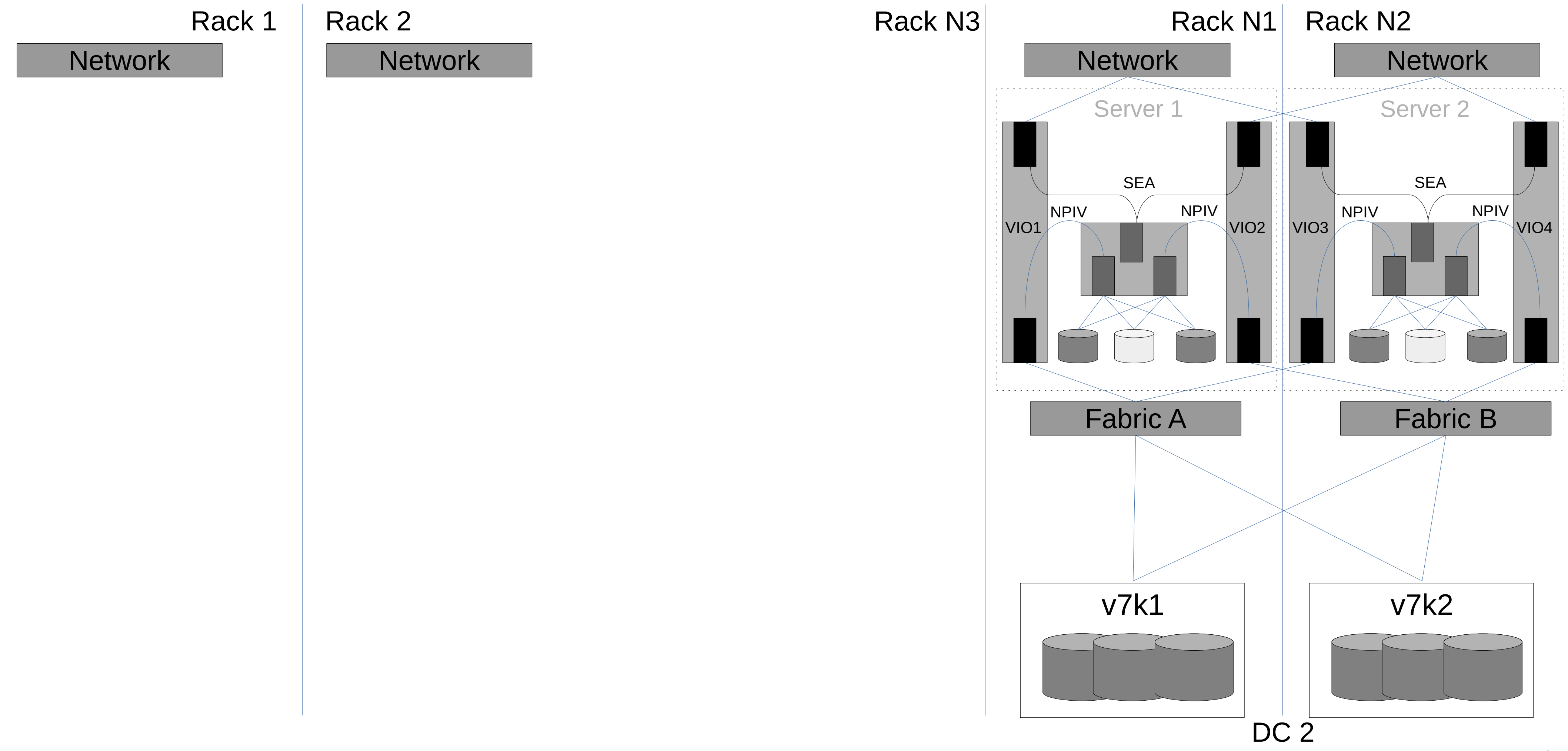


Test and cleanup

- Confirm baseline I/O performance on test LPAR with 2 way mirror matches pre-migration performance
- Confirm LPM performance between the two new racks
- Remove all customer configuration from the loan equipment
  - Scrub all the disks on the Loan storage (learned lessons about performance of *urandom* and *yes*)
  - Remove all configurations from the loan switches and factory reset
  - Remove all configurations from the FCIP routers and factory reset

Balance workload in DC2

Belisama



Conclusion

- Customer happy (Power was the only architecture that was migrated without an outage – or impact on performance
- Lessons learned
  - While technical team is comfortable with the technology and plans, not all of the customers teams are that familiar with the features we took advantage off – communicate!
  - Control communications – While it is important for everyone to be kept up to date, don’t swamp with too much information
  - You can never test to much
  - You can never plan too much (and have a good imagination when dreaming up potential issues – access as DC if it is still a worksite?)

References

- Methods of duplicating the OS: https://www.ibm.com/support/pages/node/670623
- Tips migrating workloads to POWER9 https://www.ibm.com/downloads/cas/39XWR7YM
- IBM Power virtualisation best practices https://www.ibm.com/downloads/cas/JVGZA8RW
- Please feel free to contact me for more specific details about anything mentioned above.

Session: Power migrations (101)

¿ Questions ?

Thanks!

For further information….
Contact:

Antony (Red) Steel antony.steel@belisama.com.sg
+65 9789 6663

Power Systems