PowerVM virtualization features overview

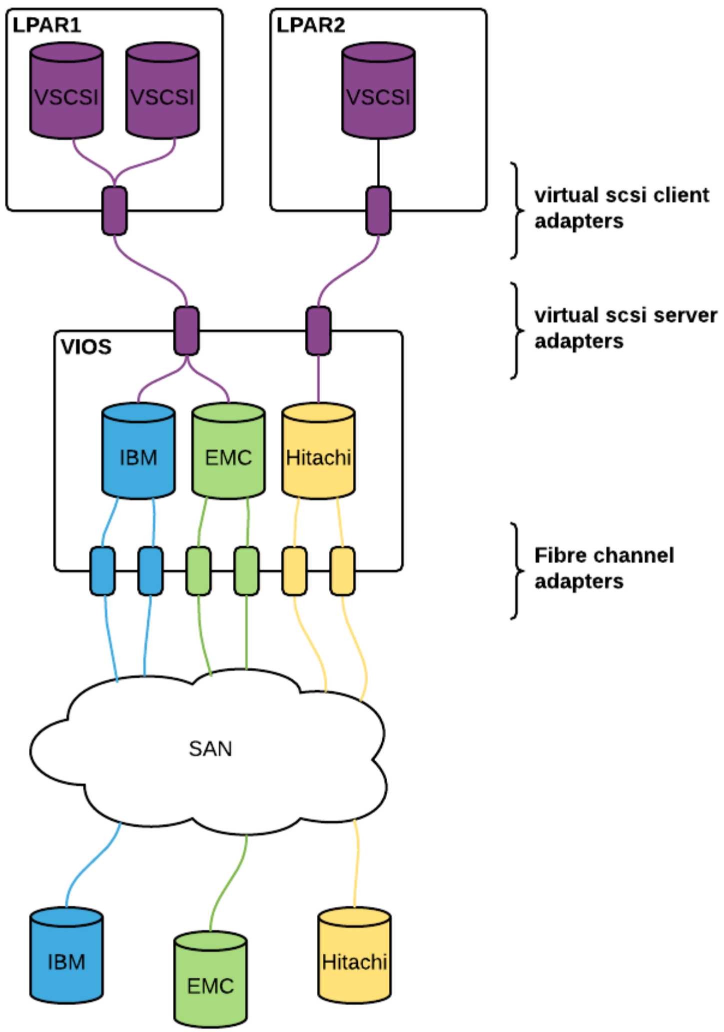
Version 1.0

Alain Dejoux
IBM Systems Lab Services consultant
Email: adejoux@fr.ibm.com
Twitter: @adejoux
Web site : djouxtech.net

Agenda

- Disk virtualization
  - **VSCSI**
  - NPIV
  - Shared Storage Pool
  - Tuning
- Network virtualization
  - Shared Ethernet Adapter
  - SR-IOV
  - VNIC
- DPO

VSCSI overview



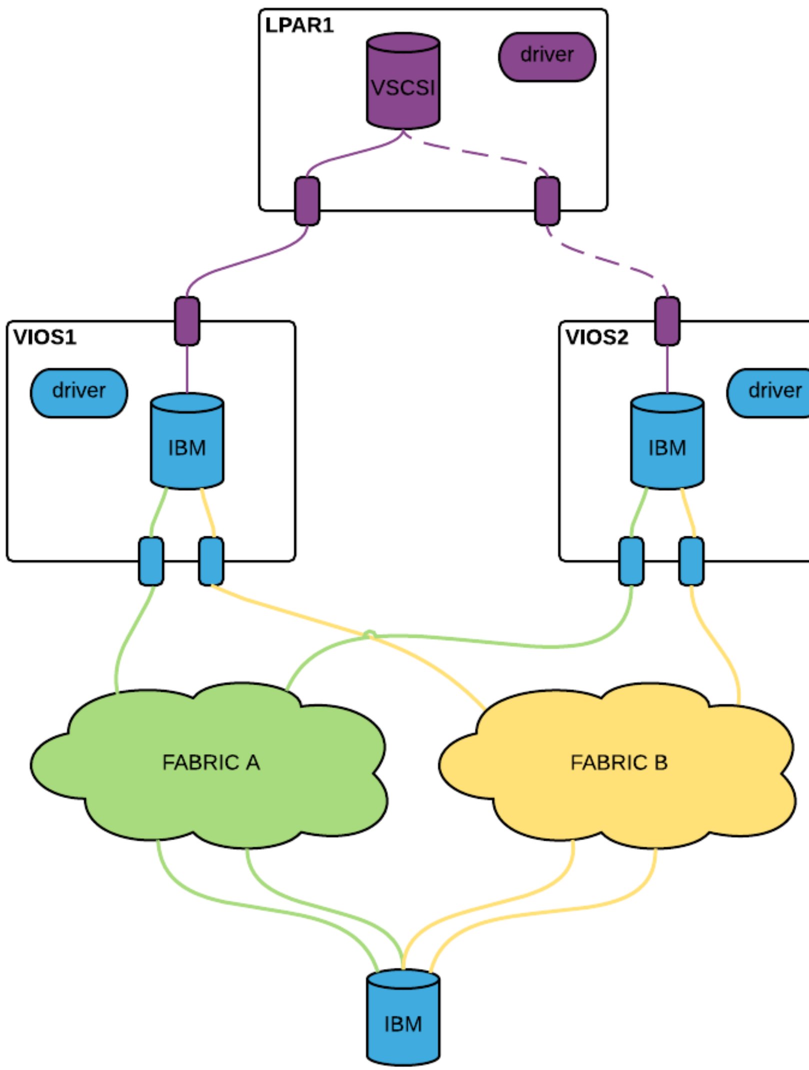
Vendor SAN storage associated with the FC adapter port is managed by the VIOS.

Disk drivers are installed at VIOS level.

Each SAN disk is individually mapped on a client partition.

Client partition use standard MPIO driver.

VSCSI dual vio servers configuration



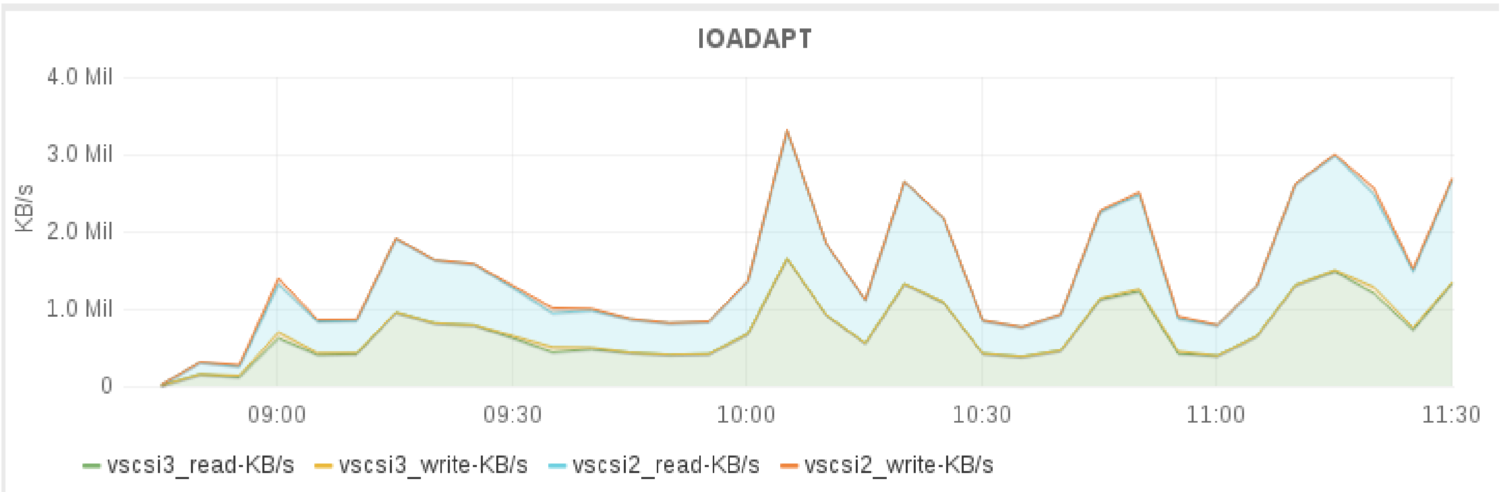
Standard availability solution.

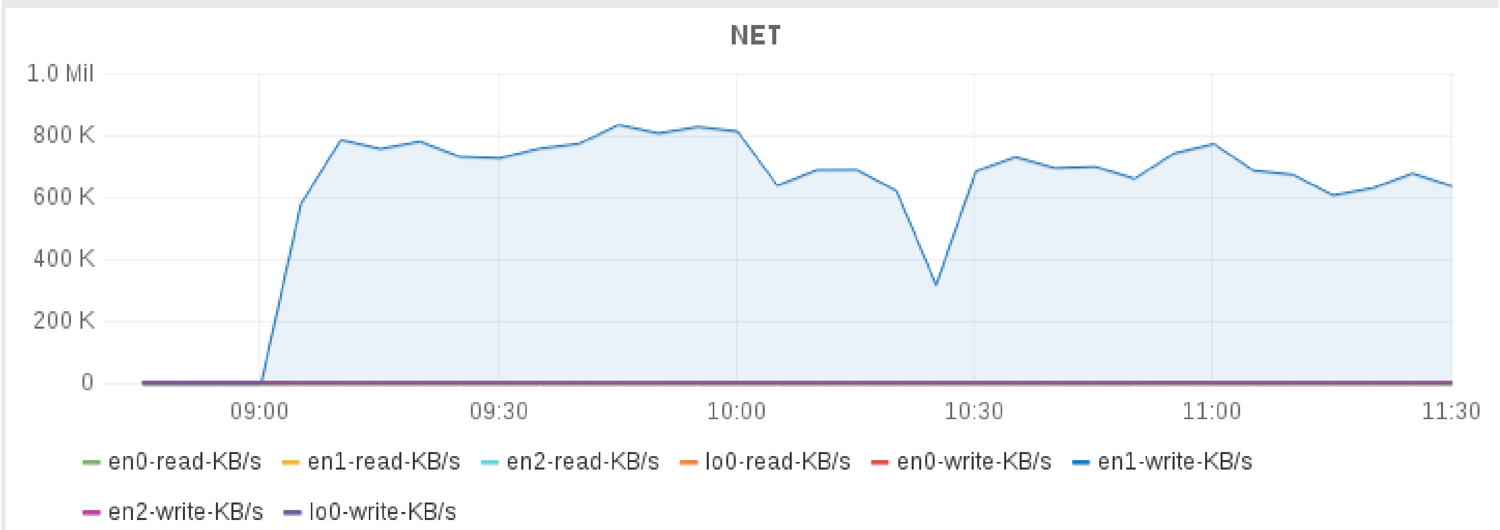
Failover mode only.

This configuration allows vendor disk driver updates without outage at partition level.

Drivers updates are performed only at VIOS Level.

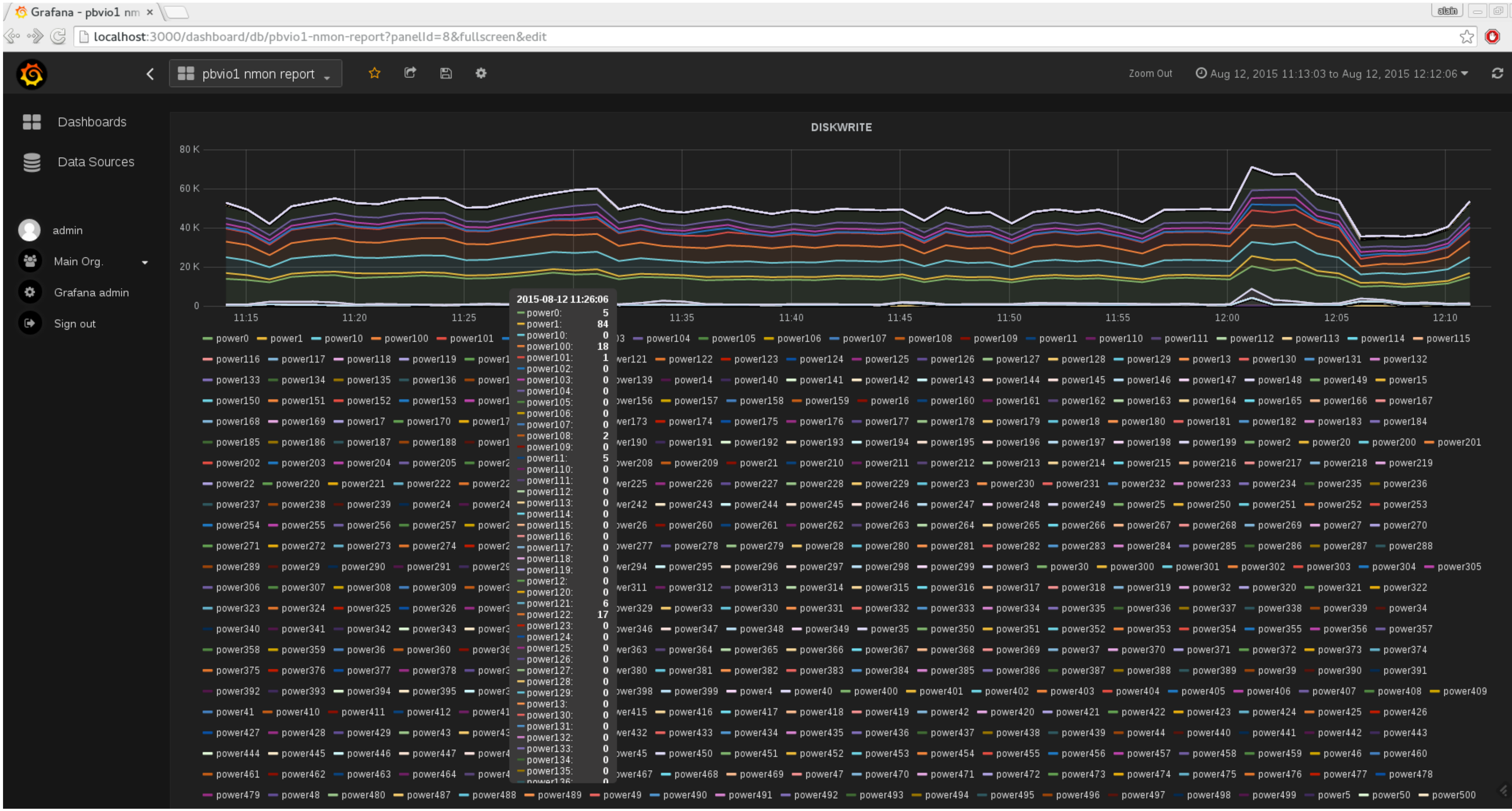
VSCSI performance: Power8 Switzerland customer example





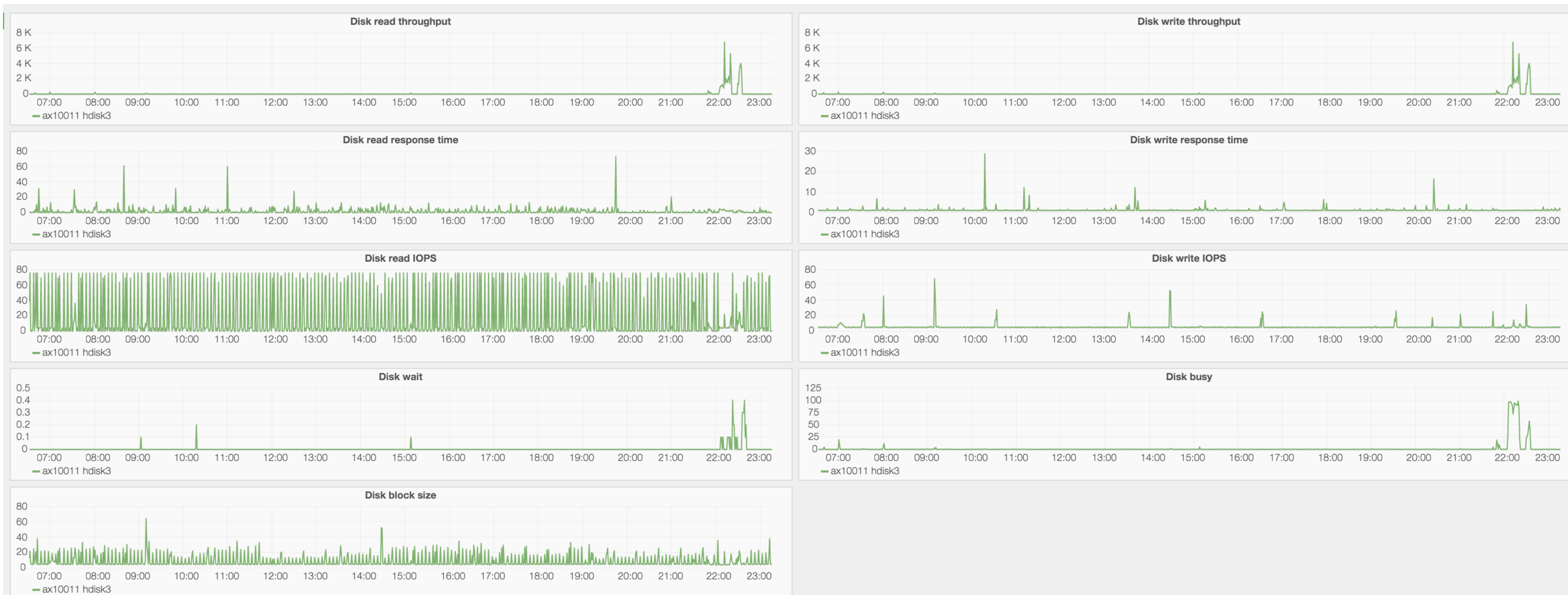
Some customers was expecting lower performance with vscsi. It's wrong.
Fully virtualized network sustains 800MB/s during backups.
Virtual SCSI adapters are regularly reaching 3 GB/s.

VSCSI: too many disks at vios level(customer example)



It's the main "issue" we see with vscsi. Disk management doesn't need to be difficult on Power.
In this, example : 100+ partitions using vscsi. 600+ luns
=> Very difficult for system management, boot time, performance analysis or problem determination.

performance: disk metrics

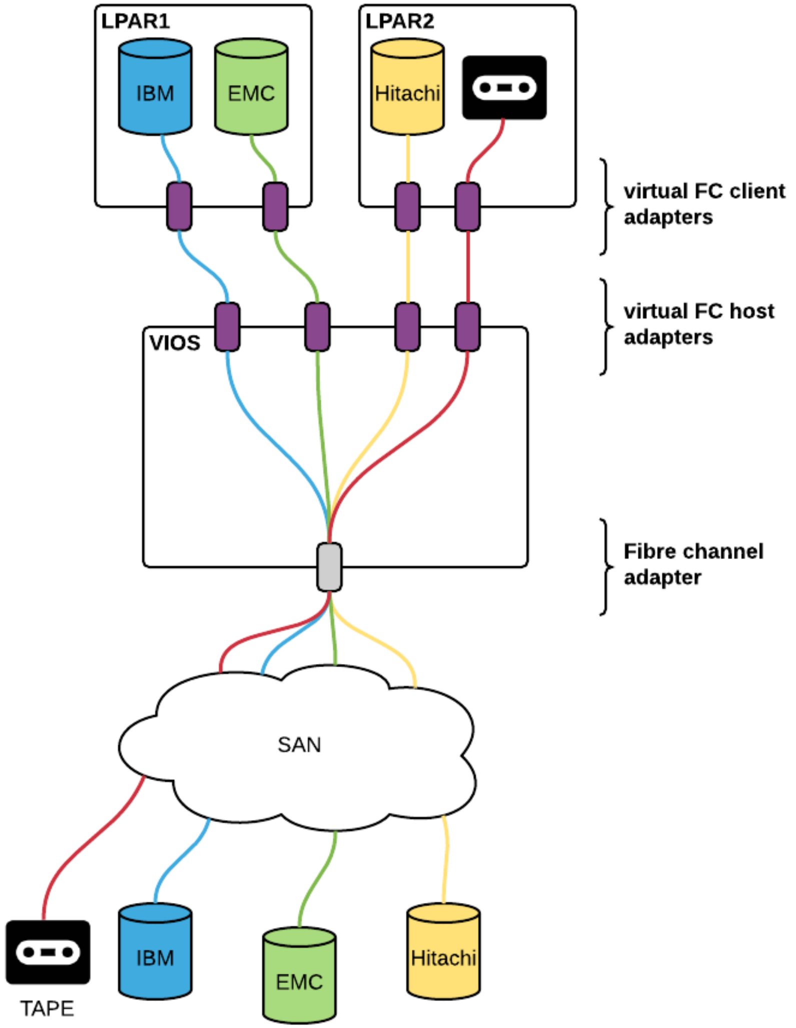


IO performance analysis needs to correlate multiple metrics from the same disk.
Work become harder when you have thousands of disks.

Agenda

- Disk virtualization
  - VSCSI
  - **NPIV**
  - Shared Storage Pool
  - Tuning
- Network virtualization
  - Shared Ethernet Adapter
  - SR-IOV
  - VNIC
- DPO

NPIV overview



NPIV: N_Port ID Virtualization.

It's a Fibre Channel hardware feature. Multiple Fibre Channel node port IDs can share a single physical N_Port.

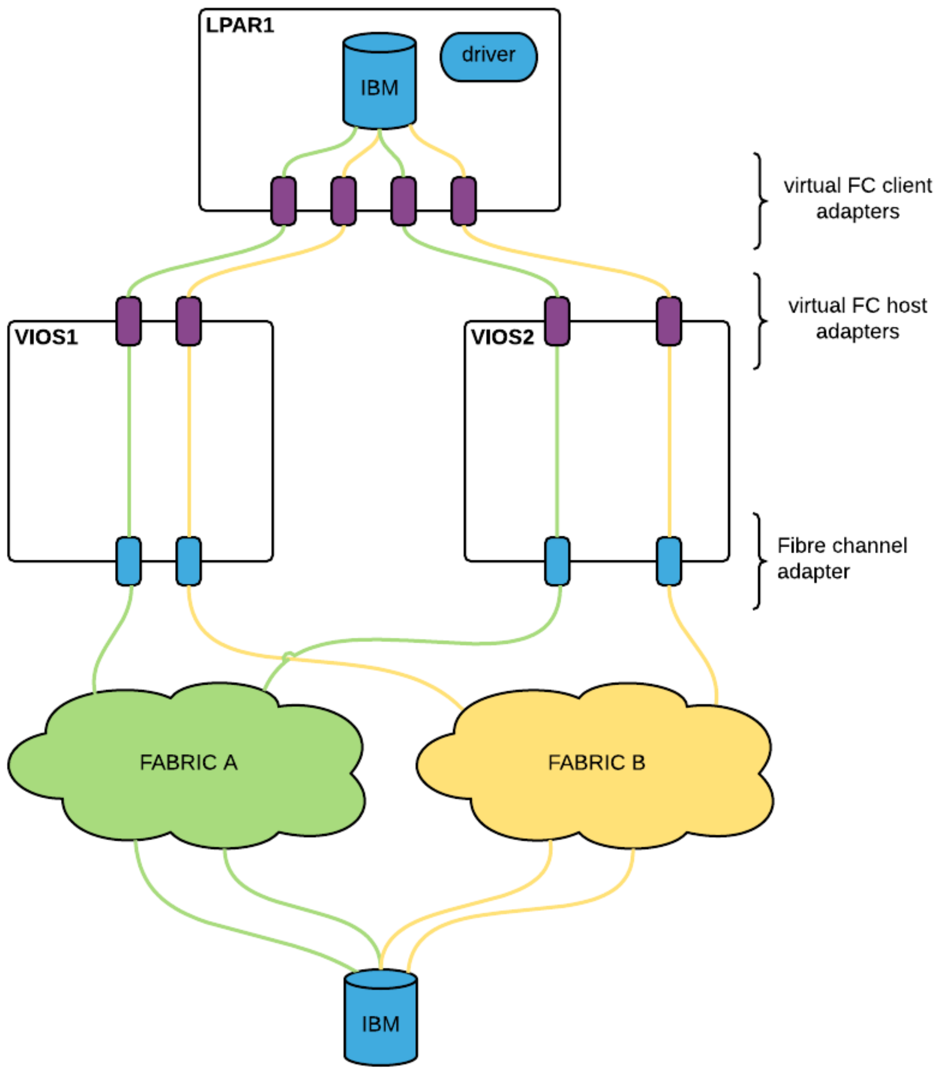
Each client partition has dedicated wwpns.

Disk drivers are installed at lpar level.

Load balancing available at lpar level.

Easiest virtualization solution for lan free backup.

NPIV dual vio servers configuration

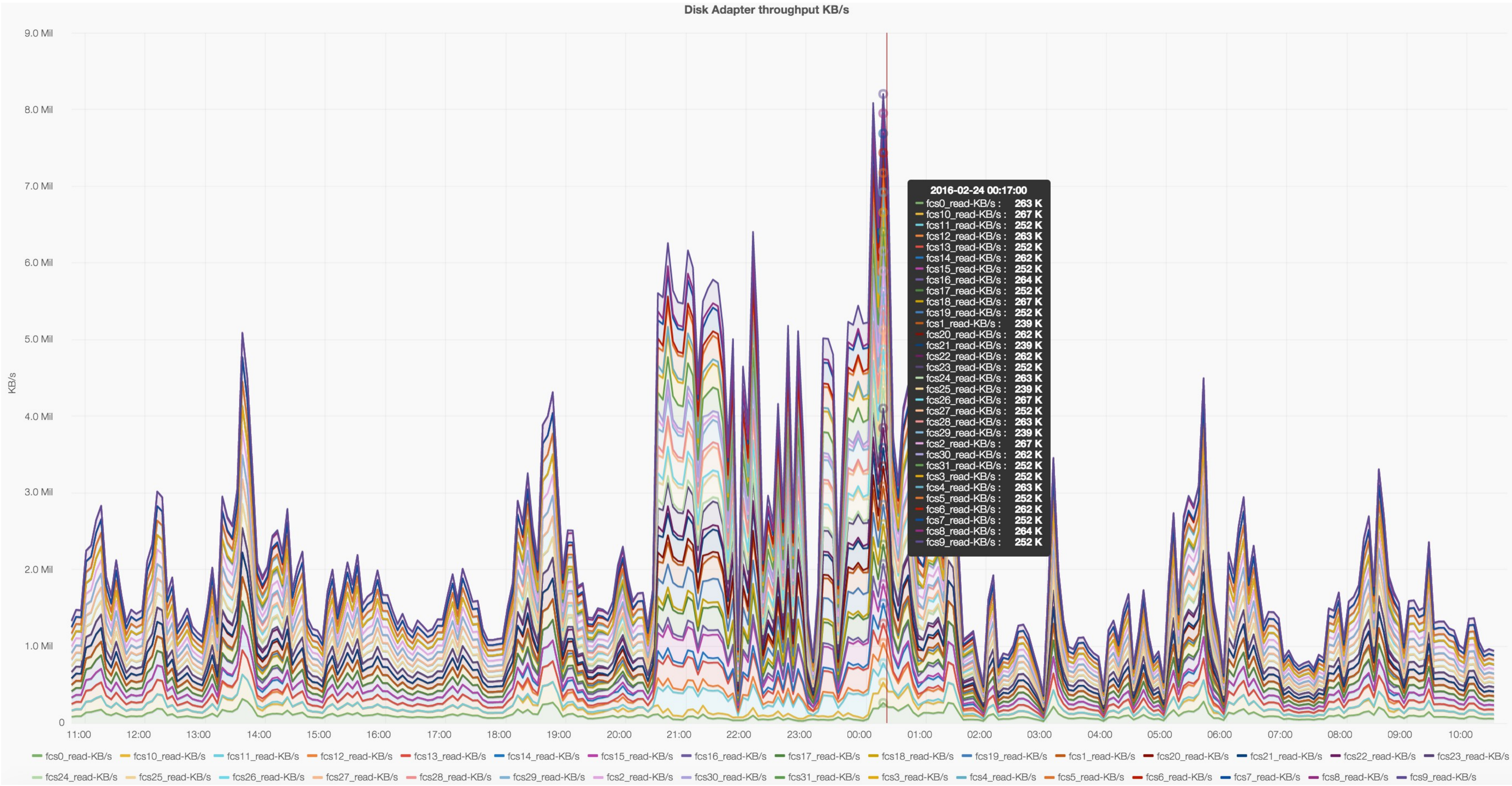


Load balancing at partition level.

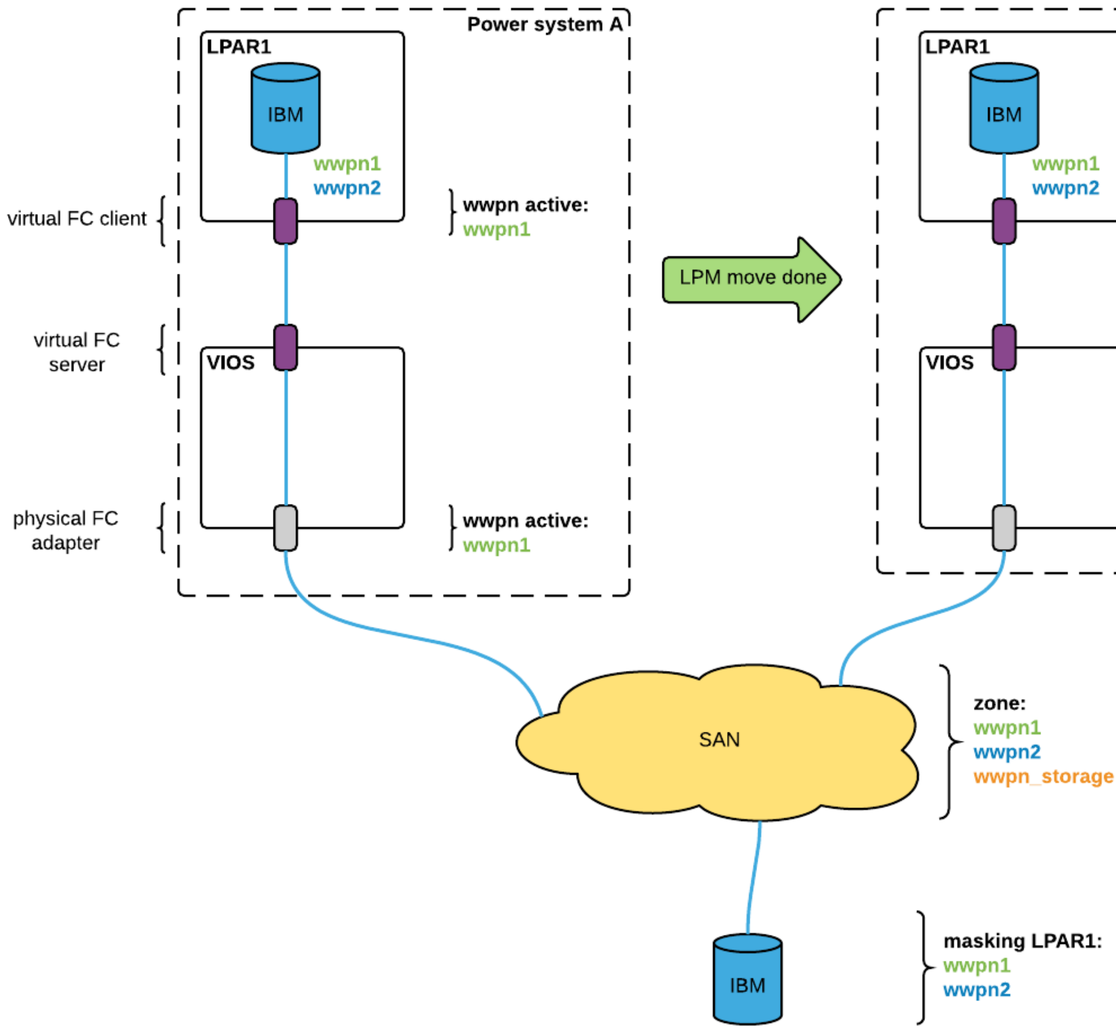
Most common setup is 4 paths.

Disk drivers are installed at vios level.

NPIV load balancing



NPIV and Live Partition Mobility



To see what is the current active wwpn:
fcstat fcsX

Each virtual fibre channel adapter has 2 wwpns. Only one is active at one time. When LPM operation is performed, the partition will switch to use the inactive wwpn on the destination system. It will keep using it after the operation.

NPIV and Live Partition Mobility: SAN disk level validation

Before VIOS 2.2.4(Dec 2015), LPM validation was only validating if the inactive wwpn was able to connect on the same storage port than the active wwpn on each virtual fibre channel adapter.

If host mapping was not exactly the same for the inactive wwpn, problems arise.

If it's not possible to ensure the zoning and mapping recommendations are applied, use disk level validation. The validation will take more time but will ensure all disks are visible on the target system. It need to be enabled on all source vios :

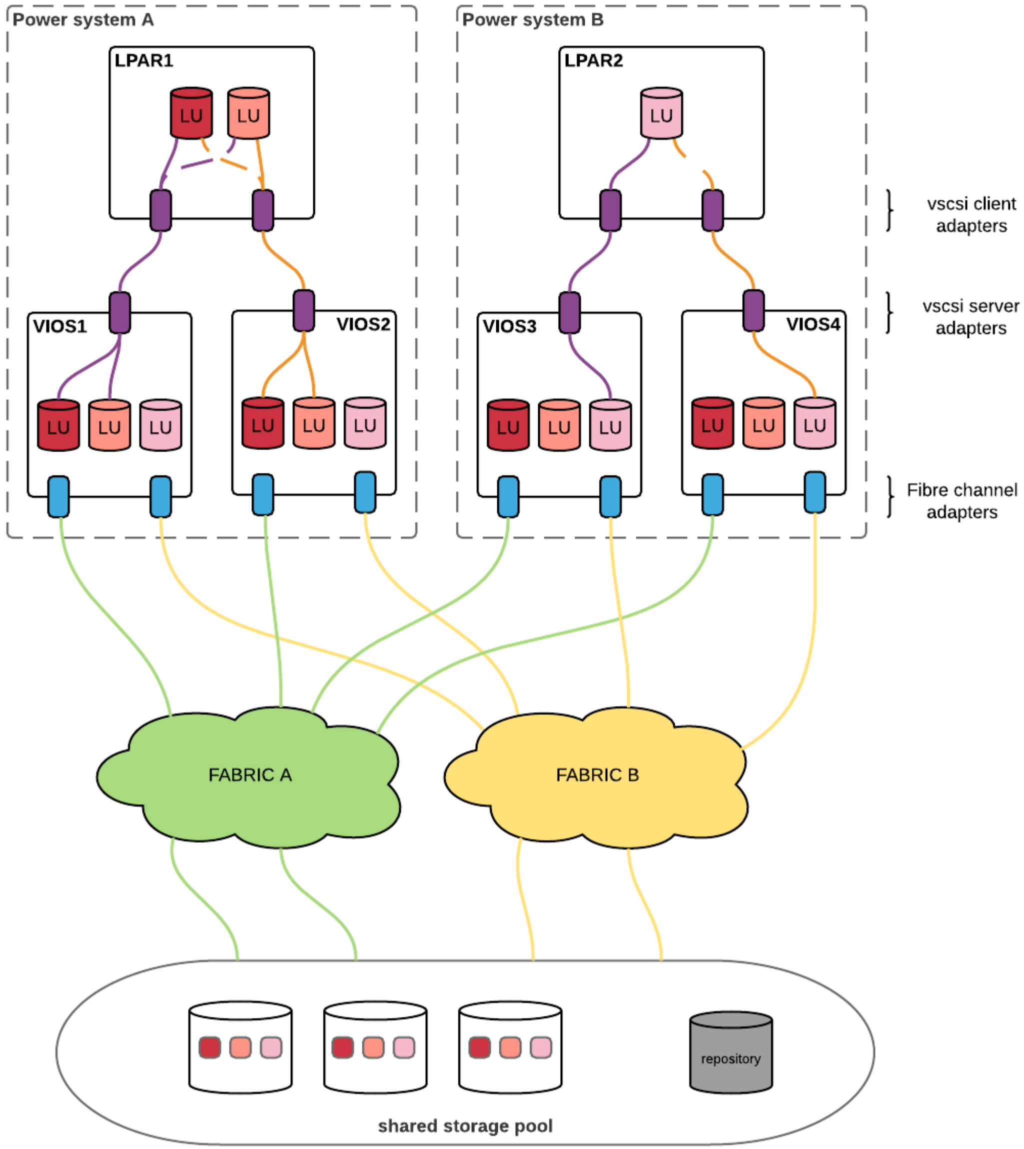
```
chdev -dev vioslpm0 -attr src_lun_val=on
```

A good description of this feature is available here:
disk-level-validation-for-lpm-of-npiv-lpar-2

Agenda

- Disk virtualization
  - VSCSI
  - NPIV
  - **Shared Storage Pool**
  - Tuning
- Network virtualization
  - Shared Ethernet Adapter
  - SR-IOV
  - VNIC
- DPO

Shared storage pool overview



A shared storage pool is a shared filesystem.

Logical Units are files in this filesystem.

This files are mapped through vscsi on client lpars.

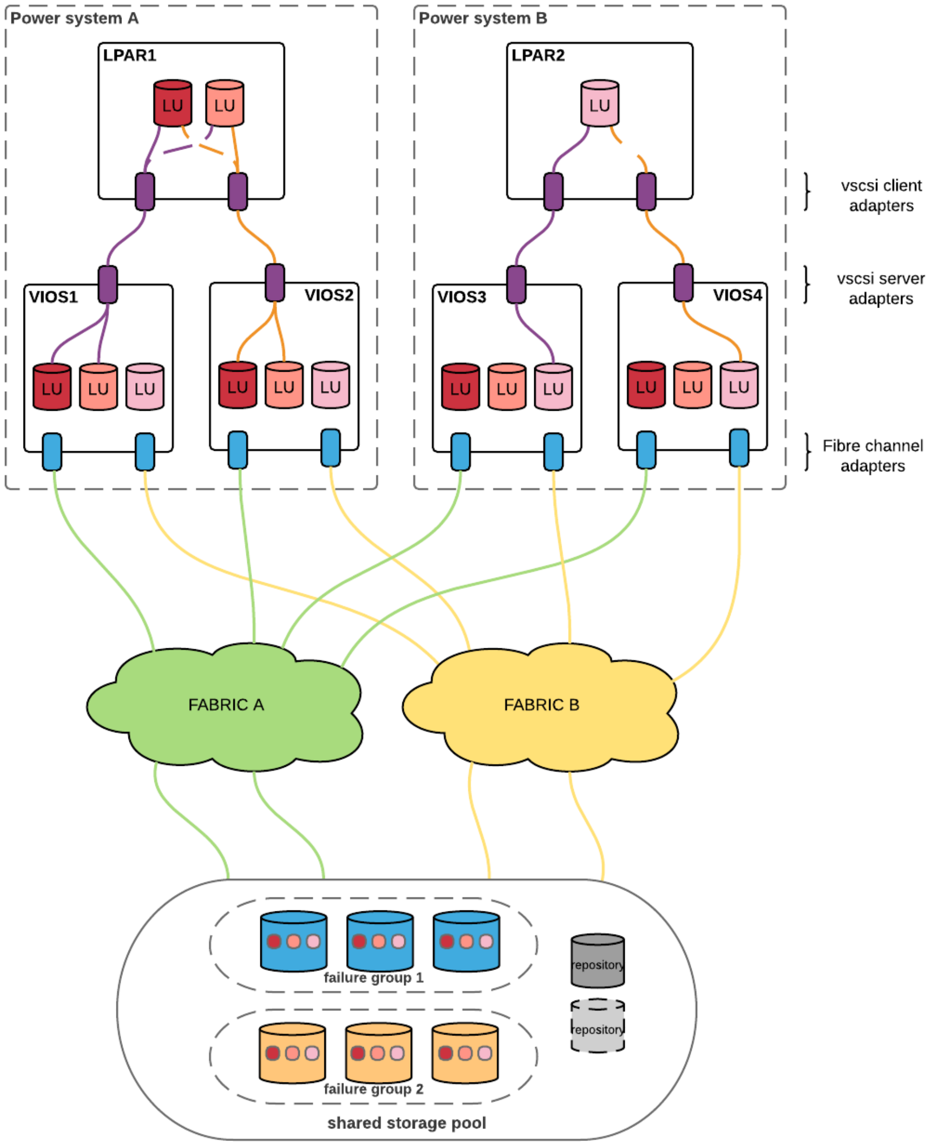
Client lpars manage them like standard MPIO disks.

Vendor disk drivers are installed at VIOS level.

VIOS are in cluster.

Cluster Aware AIX infrastructure is used, like in PowerHA.
IP network is important in the cluster.

Shared storage pool failure group



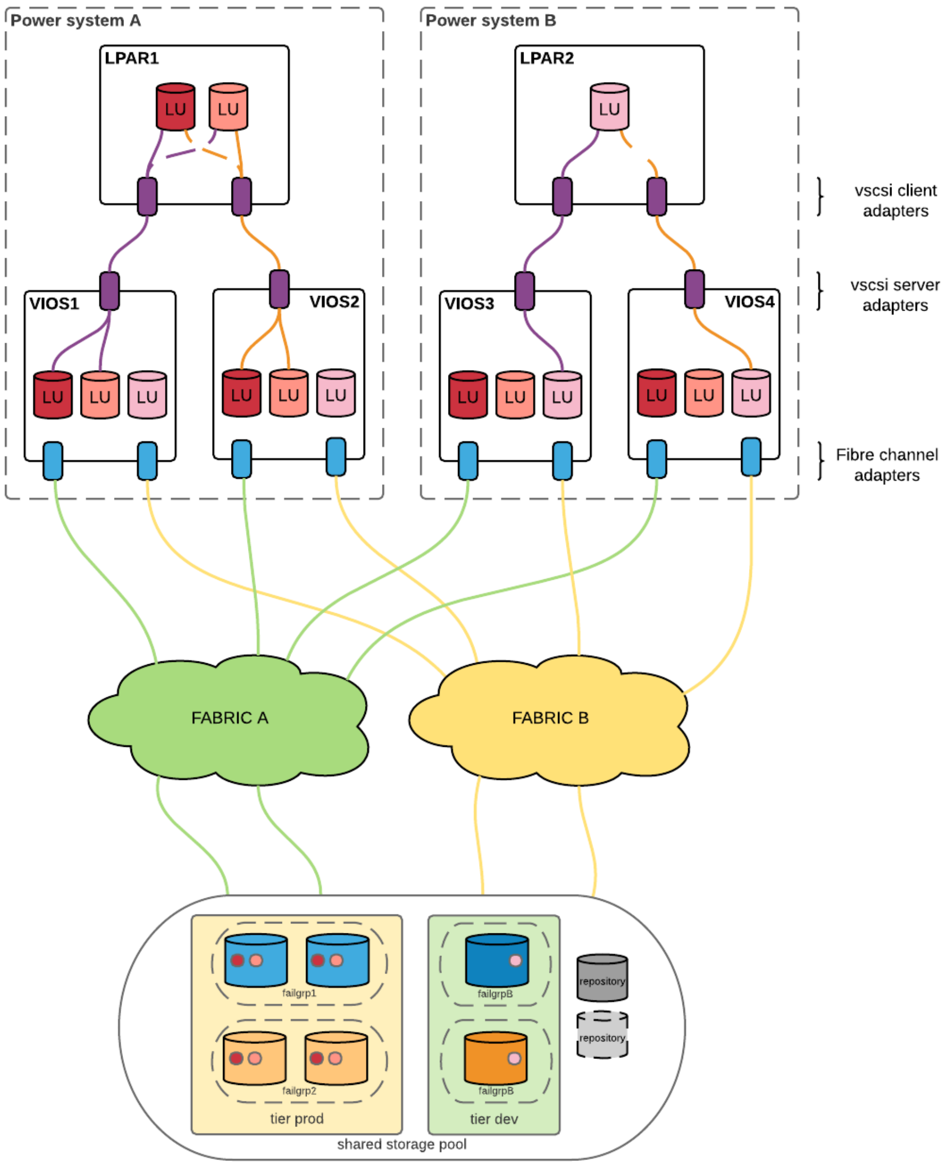
Failure groups was introduced to provide mirroring feature to SSP.

Allow to mirror data between physical storages.

It's **not** disk to disk replication.

Easy setup.

Shared storage tiering



Tiering was the most wanted feature.

Allow to dedicate disks to workload.

Easy data migration between tiers.

Compatible with failure groups.

Available in VIOS 2.2.4.10 and later.

Agenda

- Disk virtualization
  - VSCSI
  - NPIV
  - Shared Storage Pool
  - **Tuning**
- Network virtualization
  - Shared Ethernet Adapter
  - SR-IOV
  - VNIC
- DPO

Disk queue depth and FC adapter number of command elements

- The disk **queue_depth** parameter define the number of slots available for in-flight IOs the disk can have at one moment.
  One in-flight IO is an IO request initiated to the storage which is still not completed.
  When the IO is completed, the slot in the queue is freed.

- If the queue is full, new IOs are put on the wait queue until a slot is freed.
  => performance impact.

- Allowable values for **queue_depth** range from 1 to 256.
  Review with the storage vendor the recommended/supported value.

- Similarly a FC adapter has a queue for in-flight IOs. The number of slots is defined by **num_cmd_elems**.

Disk queue depth tuning

- Review the queue usage with iostat:
- **# iostat -D hdisk2 2 2**

```
System configuration: lcpu=8 drives=13 paths=43 vdisks=22
hdisk2      xfer:  %tm_act    bps    tps    bread    bwrtn
                   0.0      0.0    0.0    0.0      0.0
            read:      rps  avgserv  minserv  maxserv  timeouts   fails
                   0.0      0.0    0.0    0.0      0        0
            write:     wps  avgserv  minserv  maxserv  timeouts   fails
                   0.0      0.0    0.0    0.0      0        0
            queue: avgtime  mintime  maxtime  avgwqsz  avgsqsz   sqfull
                   0.0      0.0    0.0    0.0      0.0      0.0
```

Look at :

- **avgwqsz**:
  average size of wait queue size. It needs to be 0 for best performance.

- **avgsqsz:**
  average service queue size. Allows you to see how many in-flights IOs are ongoing on the disk.

- **sqfull**:
  Indicates the number of times the queue was full per second.

If possible, keep the default queue depth value used by your storage provider drivers and add disks.
But keep a low number of disks if possible. Need to balance both requirements.
It simplify performance problem troubleshooting.

Number of disks by FC adapter

- For best performance, you can apply this formula :

$$\frac{FC\ adapter\ number\ of\ command\ elements}{disk\ queue\ depth}$$

It will give you a maximum number of disks you can assign without never overloading the physical adapter.
It's a conservative value.

- If your storage driver support load balancing the formula become:

$$number\ of\ FC\ adapters * \left(\frac{FC\ adapter\ number\ of\ command\ elements}{disk\ queue\ depth}\right)$$

Reference document : AIX/VIOS Disk and Adapter queue tuning

We cannot cover the full topic on queue tuning in this presentation.

For more informations and deeper technical details, read this document :
AIX/VIOS Disk and Adapter queue tuning

IBM Americas Advanced Technical Skills

AIX/VIOS Disk and Adapter IO Queue Tuning

Dan Braden

IBM AIX Advanced Technical Skills

Vio servers rules 1/2

- Rules was introduced in VIOS 2.2.4.
- It allow to change default and current values for devices for better performance and availability.
- For example, rules related to fscsi set dynamic tracking and fast fail over:
  padmin# rules -o list|grep fscsi

driver/iocb/efscsi	dyntrk	yes
driver/iocb/efscsi	fc_err_recov	fast_fail
driver/qliocb/qlfscsi	dyntrk	yes
driver/qliocb/qlfscsi	fc_err_recov	fast_fail
driver/qiocb/qfscsi	dyntrk	yes
driver/qiocb/qfscsi	fc_err_recov	fast_fail

- For example, it will set **num_cmd_elems** and **max_transfer_size** based on the physical adapter capabilities:
  padmin# rules -o list

adapter/pciex/df1060e21410410	max_xfer_size	0x400000
adapter/pciex/df1060e21410410	num_cmd_elems	4096
adapter/pci/df1080f9	max_xfer_size	0x400000
adapter/pci/df1080f9	num_cmd_elems	2048

Vio servers rules 2/2

- Applying default rules:
  rulescfgset

- Changing a default rule for a specific device:
  rules -o modify -l hdisk0 -a reserve_policy=single_path

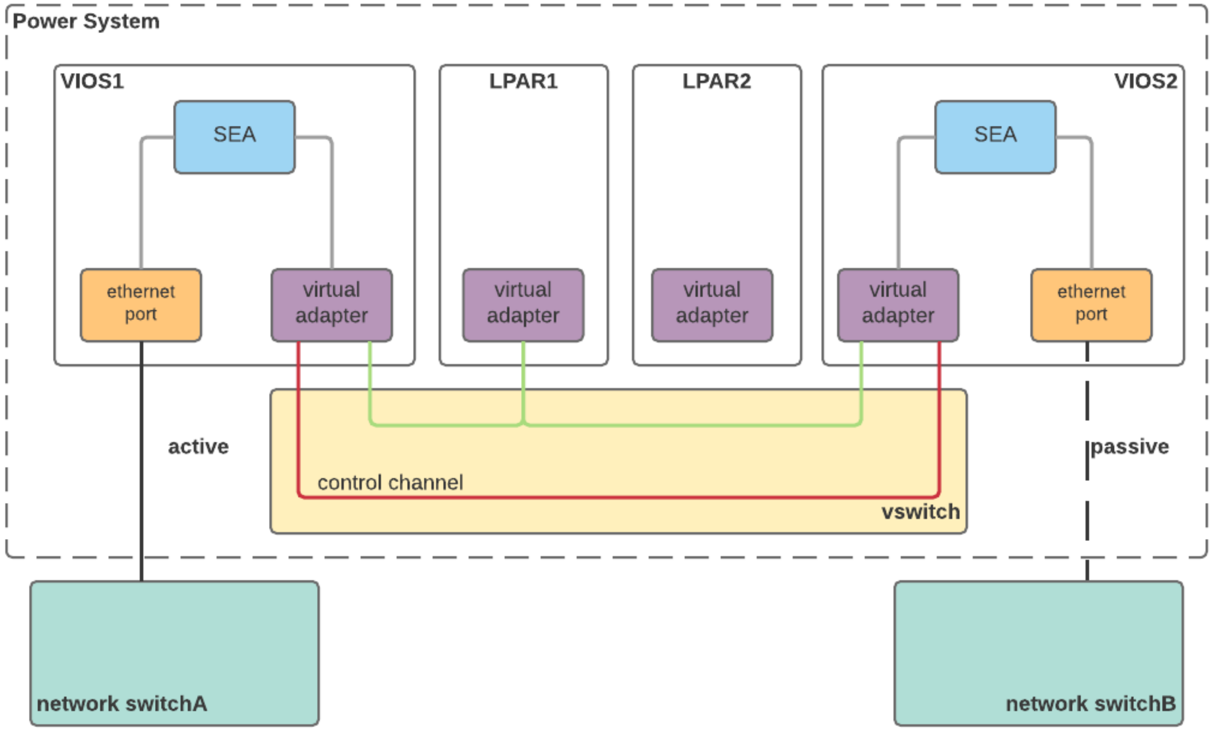
- View differences between default rules and current settings:
  padmin# rules -o diff -s -d
  devParam.disk.fcp.mpioosdisk:reserve_policy device=disk/fcp/mpioosdisk single_path | no_reserve
  ...
  devParam.PCM.friend.fcpother:algorithm device=PCM/friend/fcpother fail_over | round_robin
  ...
  devParam.adapter.pseudo.ibm_ech:hash_mode device=adapter/pseudo/ibm_ech default | src_dst_port
  devParam.adapter.pciex.df1000fe:num_cmd_elems device=adapter/pciex/df1000fe 200 | 2048
  devParam.adapter.pciex.df1000fe:max_xfer_size device=adapter/pciex/df1000fe 0x100000 | 0x400000
  ….
- **Rules man page is very well documented and explain all the capabilities.**

Rules are still new and don't cover all devices and settings but it's already a great improvement and simplify a lot the vio server tuning configuration.

Agenda

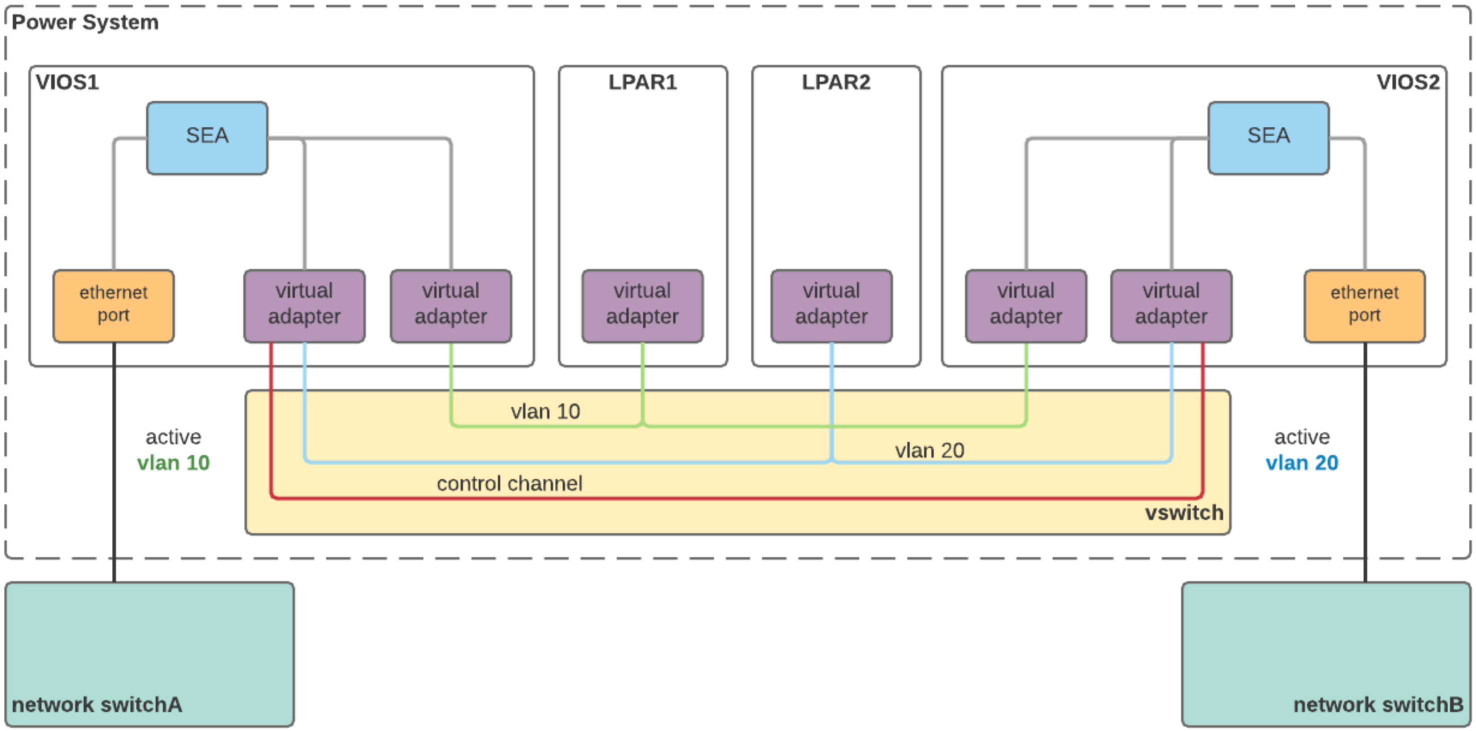
- Disk virtualization
  - VSCSI
  - NPIV
  - Shared Storage Pool
  - Tuning
- Network virtualization
  - **Shared Ethernet Adapter**
  - SR-IOV
  - VNIC

SEA FAILOVER



Great improvement in shared ethernet adapter configuration in firmware 780:
The control channel is automatically managed.

SEA LOAD SHARING



Most popular SEA configuration when multiple vlans are used.

virtual adapter network buffer tuning

Change default virtual buffers values :

chdev -dev <VENT> -attr max_buf_huge=128 -perm
chdev -dev <VENT> -attr min_buf_huge=128 -perm

chdev –l <VENT> -a max_buf_large=128 -perm
chdev -dev <VENT> -attr min_buf_large=128 -perm

chdev -dev <VENT> -attr max_buf_medium=512 -perm
chdev -dev <VENT> -attr min_buf_medium=512 -perm

chdev -dev <VENT> -attr max_buf_small=4096 -perm
chdev -dev <VENT> -attr min_buf_small=4096 -perm

chdev -dev <VENT> -attr max_buf_tiny=4096 -perm
chdev -dev <VENT> -attr min_buf_tiny=4096 –perm

In a high network activity environment, you can set buffer values to maximum.
This operation cannot be performed adapter online.
Need to be done on vio servers and client partitions.

Enabling TCP Segmentation offload and aggregation

Virtual I/O Server configuration

- **largesend** on SEA: enable TCP segmentation offload on emitted packets
  chdev -dev <SEA> -attr largesend=1
- **large_receive** on SEA : enable TCP receive segment aggregation
  chdev -dev <SEA> -attr large_receive=yes
- Already enabled by default on physical adapter

LPAR configuration

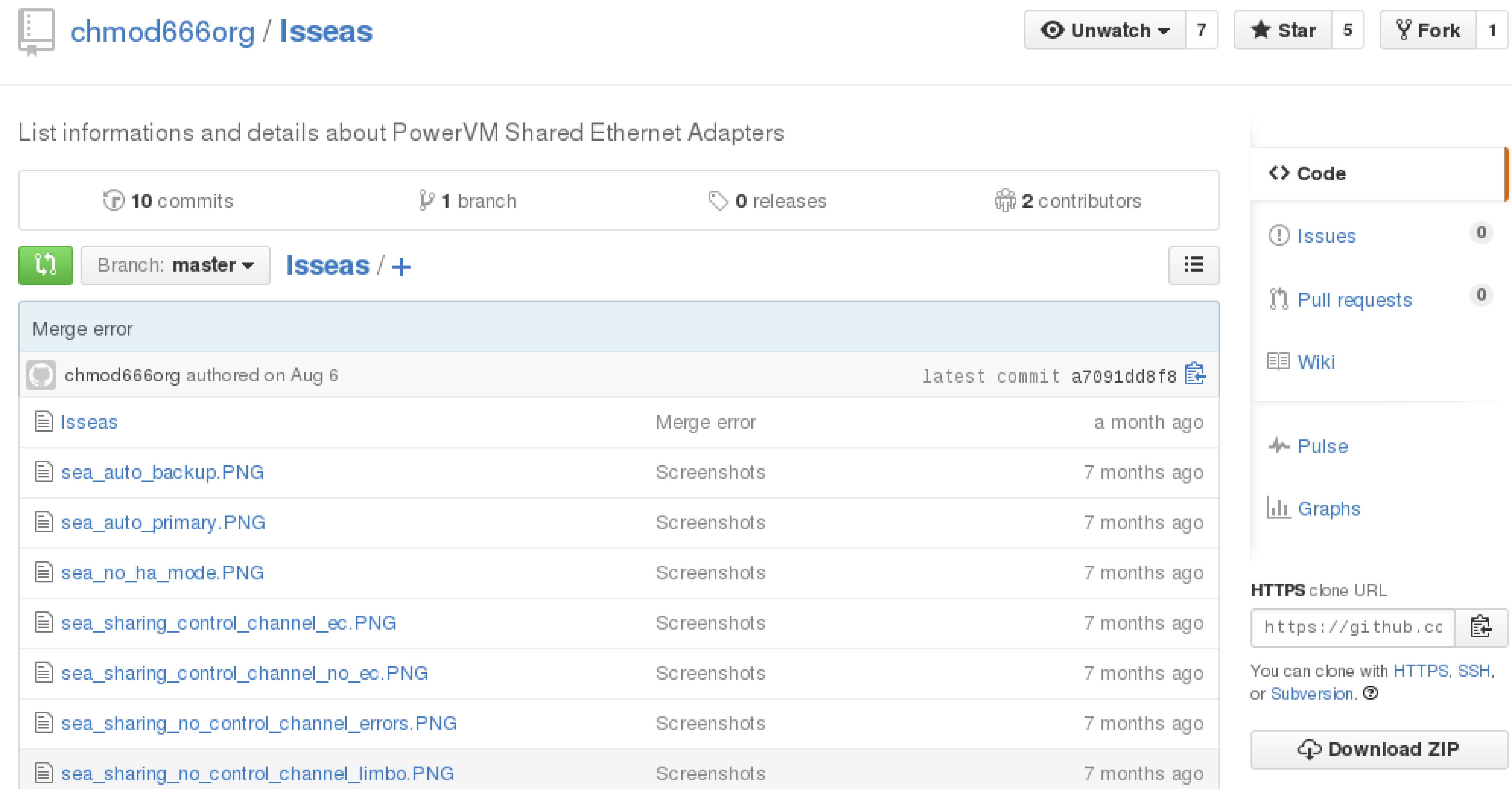
- **mtu_bypass** on VETH : enable TCP segmentation offload
  chdev –l <enX> -a mtu_bypass=on

lsseas

```
+--------------------------------------------------------+
SEA :  ent18
ha_mode                : Sharing
state                  : BACKUP_SH
number of adapters     : 5
become backup/primary: 1/0
priority               : 2
vlans                  :  13 1662 1666 1670 1674 1678 12 1663 1667 1671 1675 1679 11 1664 1668 1672 1676 1680 10 1665 1669 1673 1677 1684
flags                  :  THREAD  LARGESEND  LARGE_RECEIVE  ACCOUNTING
+--------------------------------------------------------+
 ETHERCHANNEL
adapter phys_adapters                    mode      hash_mode       jumbo
------- -------------                    ----      ---------       -----
ent16   ent2,ent3,ent6,ent7              8023ad    src_dst_port    no
 REAL ADAPTERS
adapter slot  hardware_path                   link selected_speed        running_speed         actor_system      actor_sync  partner_system    partner_port partner_sync
------- ----  -------------                   ---- --------------        -------------         ------------      ----------  --------------    ------------ ------------
ent2    C4    U2C4E.001.DBJN914-P2-C4-T3      Up   1000_Mbps_Full_Duplex 1000_Mbps_Full_Duplex 6C-AE-8B-69-5E-2A IN_SYNC     00-23-04-EE-BF-90 0x231A       IN_SYNC
ent3    C4    U2C4E.001.DBJN914-P2-C4-T4      Up   1000_Mbps_Full_Duplex 1000_Mbps_Full_Duplex 6C-AE-8B-69-5E-2A IN_SYNC     00-23-04-EE-BF-90 0x231B       IN_SYNC
ent6    C4    U2C4E.001.DBJD038-P2-C4-T3      Up   1000_Mbps_Full_Duplex 1000_Mbps_Full_Duplex 6C-AE-8B-69-5E-2A IN_SYNC     00-23-04-EE-BF-90 0x231C       IN_SYNC
ent7    C4    U2C4E.001.DBJD038-P2-C4-T4      Up   1000_Mbps_Full_Duplex 1000_Mbps_Full_Duplex 6C-AE-8B-69-5E-2A IN_SYNC     00-23-04-EE-BF-90 0x231F       OUT_OF_SYNC
 VIRTUAL ADAPTERS
adapter slot  hardware_path                   priority active port_vlan_id vswitch        mode   vlan_tags_ids
------- ----  -------------                   -------- ------ ------------ -------        ----   -------------
ent8    C10   U9117.MMD.65ED82C-V2-C10-T1     2        False  10           vdcb           VEB    1665,1669,1673,1677,1684
ent9    C11   U9117.MMD.65ED82C-V2-C11-T1     2        False  11           vdcb           VEB    1664,1668,1672,1676,1680
ent10   C12   U9117.MMD.65ED82C-V2-C12-T1     2        True   12           vdcb           VEB    1663,1667,1671,1675,1679
ent11   C13   U9117.MMD.65ED82C-V2-C13-T1     2        True   13           vdcb           VEB    1662,1666,1670,1674,1678
 CONTROL CHANNEL
adapter slot  hardware_path                   port_vlan_id  vswitch
------- ----  -------------                   ------------  -------
ent12   C14   U9117.MMD.65ED82C-V2-C14-T1     99            vdcb
```

lsseas

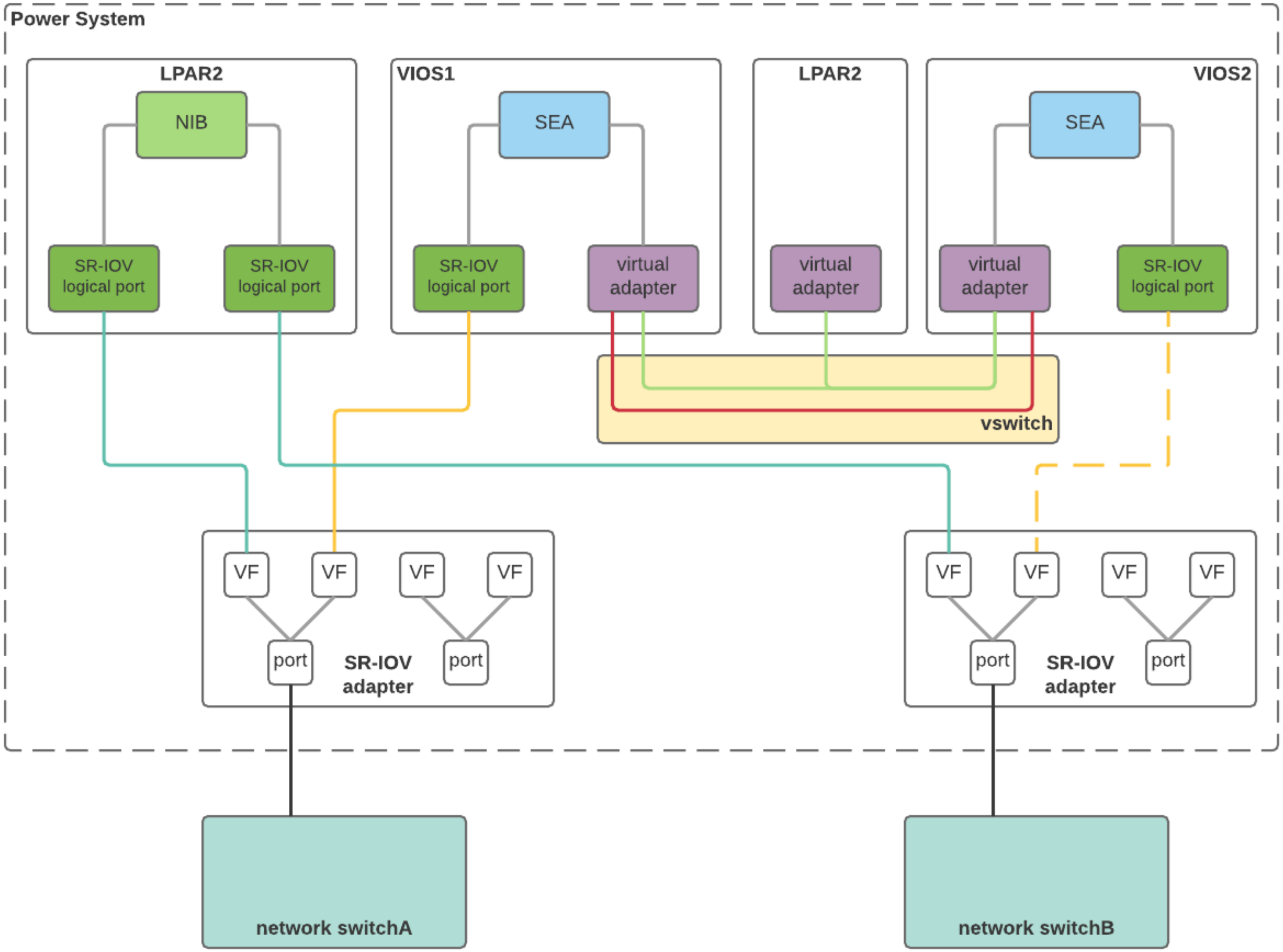
download site : https://github.com/chmod666org/lsseas



Agenda

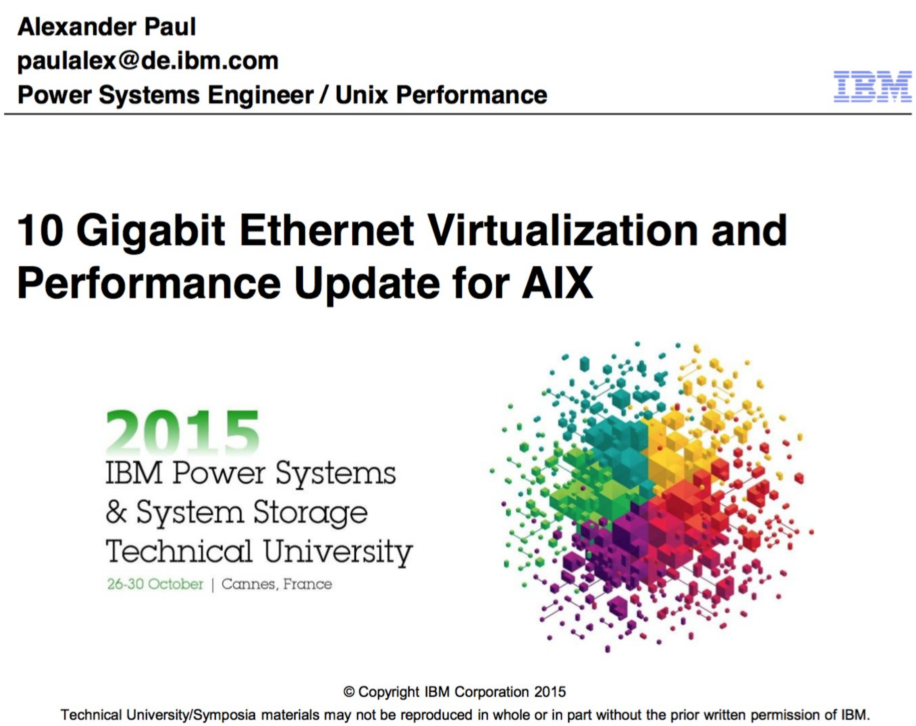
- Disk virtualization
  - VSCSI
  - NPIV
  - Shared Storage Pool
  - Tuning
- Network virtualization
  - Shared Ethernet Adapter
  - **SR-IOV**
  - VNIC
- DPO

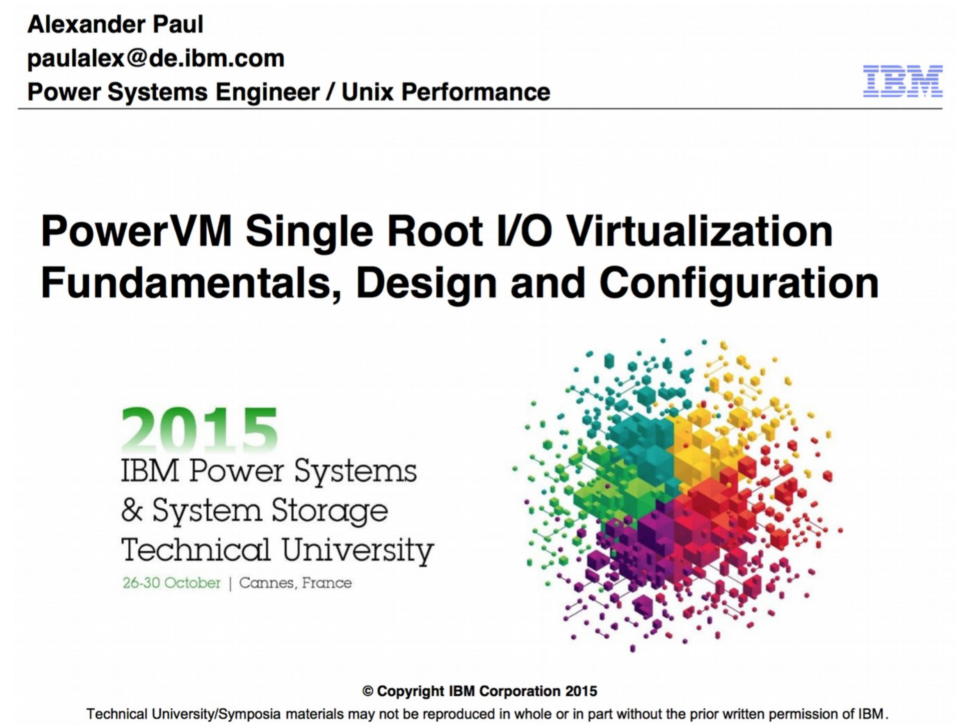
SR-IOV overview



Single Root I/O Virtualization. PCIe standard.
Share physical port between multiple partitions.
Live Partition Mobility not supported.

SR-IOV and virtual ethernet great presentations

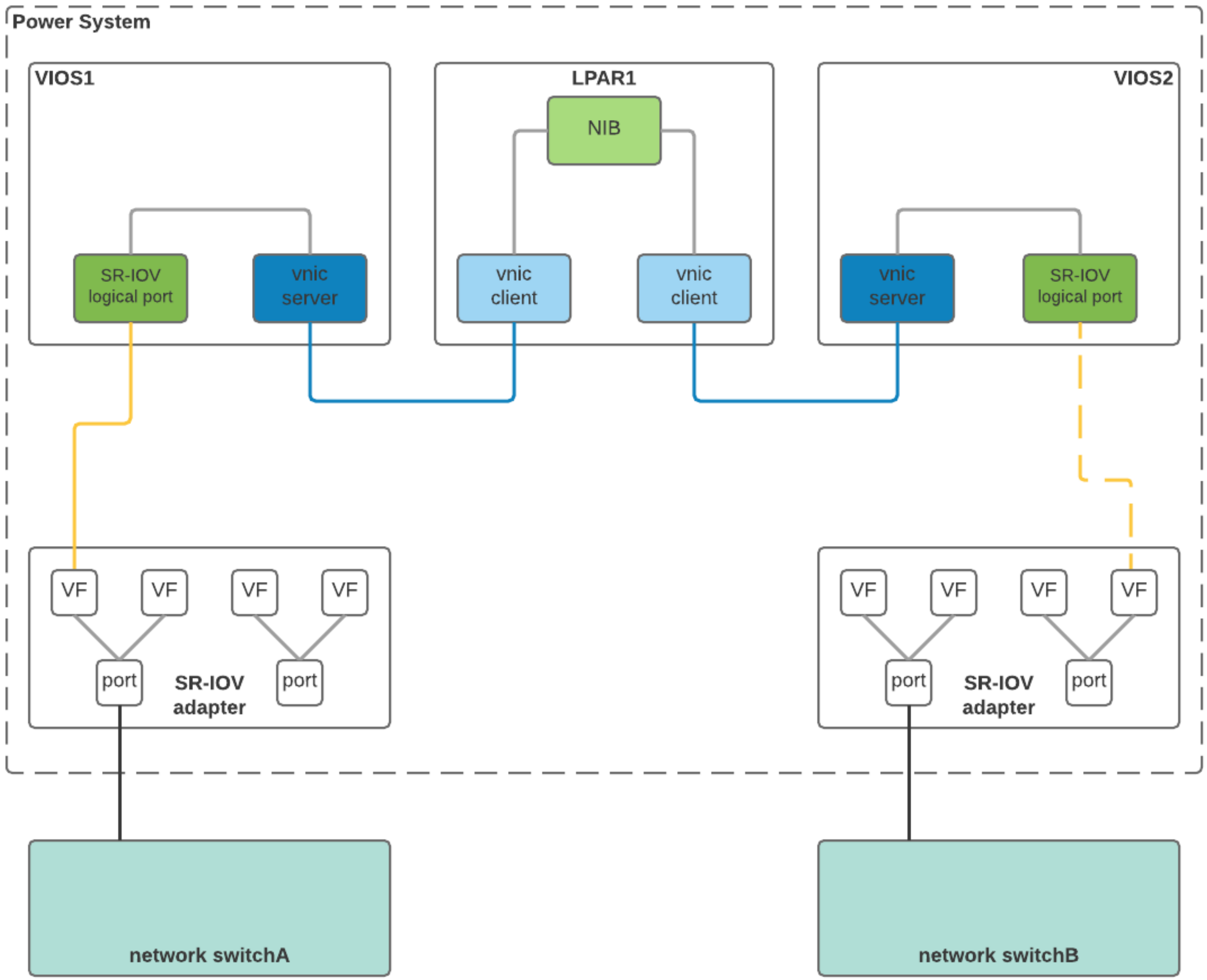




Agenda

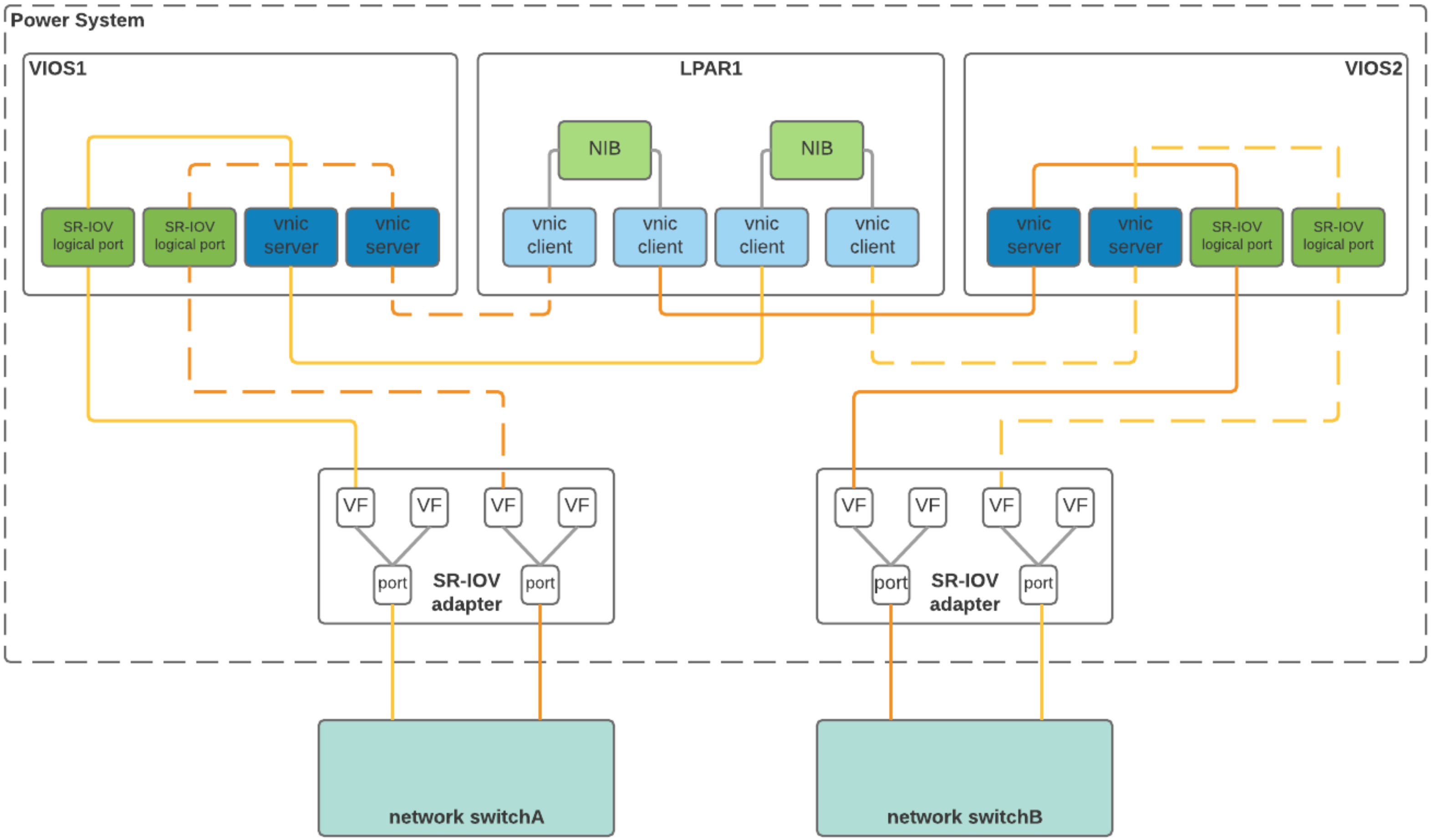
- Disk virtualization
  - VSCSI
  - NPIV
  - Shared Storage Pool
  - Tuning
- Network virtualization
  - Shared Ethernet Adapter
  - SR-IOV
  - **VNIC**
- DPO

VNIC overview



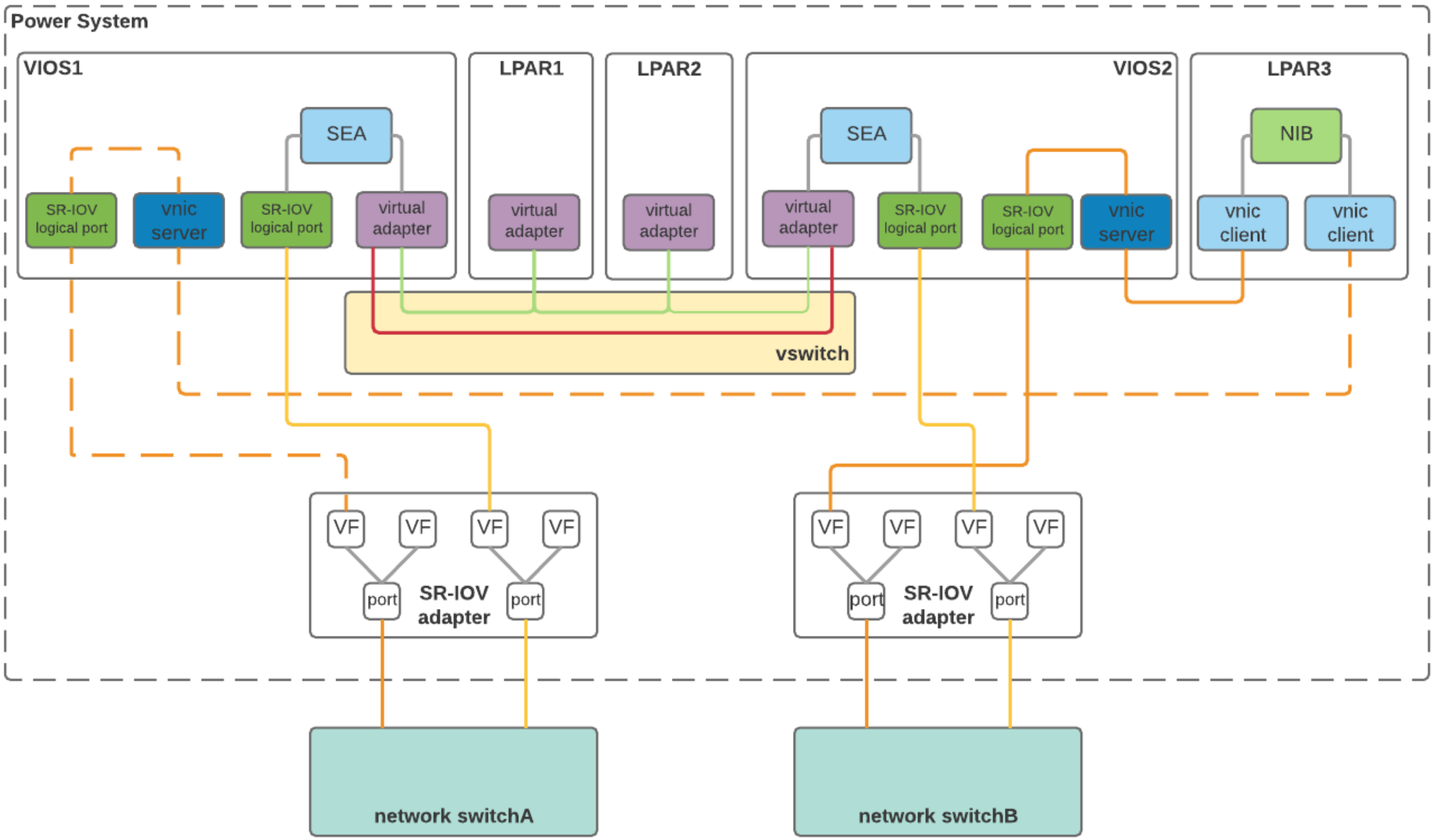
Dedicated VNIC allows Live Partition Mobility.
It's a new kind of virtual adapters.

VNIC



VNIC configuration can become complex when each partition needs multiple vlan.
In this example, for 2 vlans, you have 6 network adapters on your client partition.

VNIC + SEA configuration(customer implementation)



Here VNIC adapters are only used for a partition requesting a high performance level.
SR-IOV is used to create SEA used by other partitions.
Note: only one logical port with promiscuous mode by physical port. Mandatory for SEA.

Agenda

- Disk virtualization
  - VSCSI
  - NPIV
  - Shared Storage Pool
  - Tuning
- Network virtualization
  - Shared Ethernet Adapter
  - SR-IOV
  - VNIC
- **DPO**

Dynamic Platform Optimizer (DPO)

- **Optimizer is launched via HMC command-line interface**

- **DPO re-assigns memory and cores to partitions in order to attain better placement affinity**

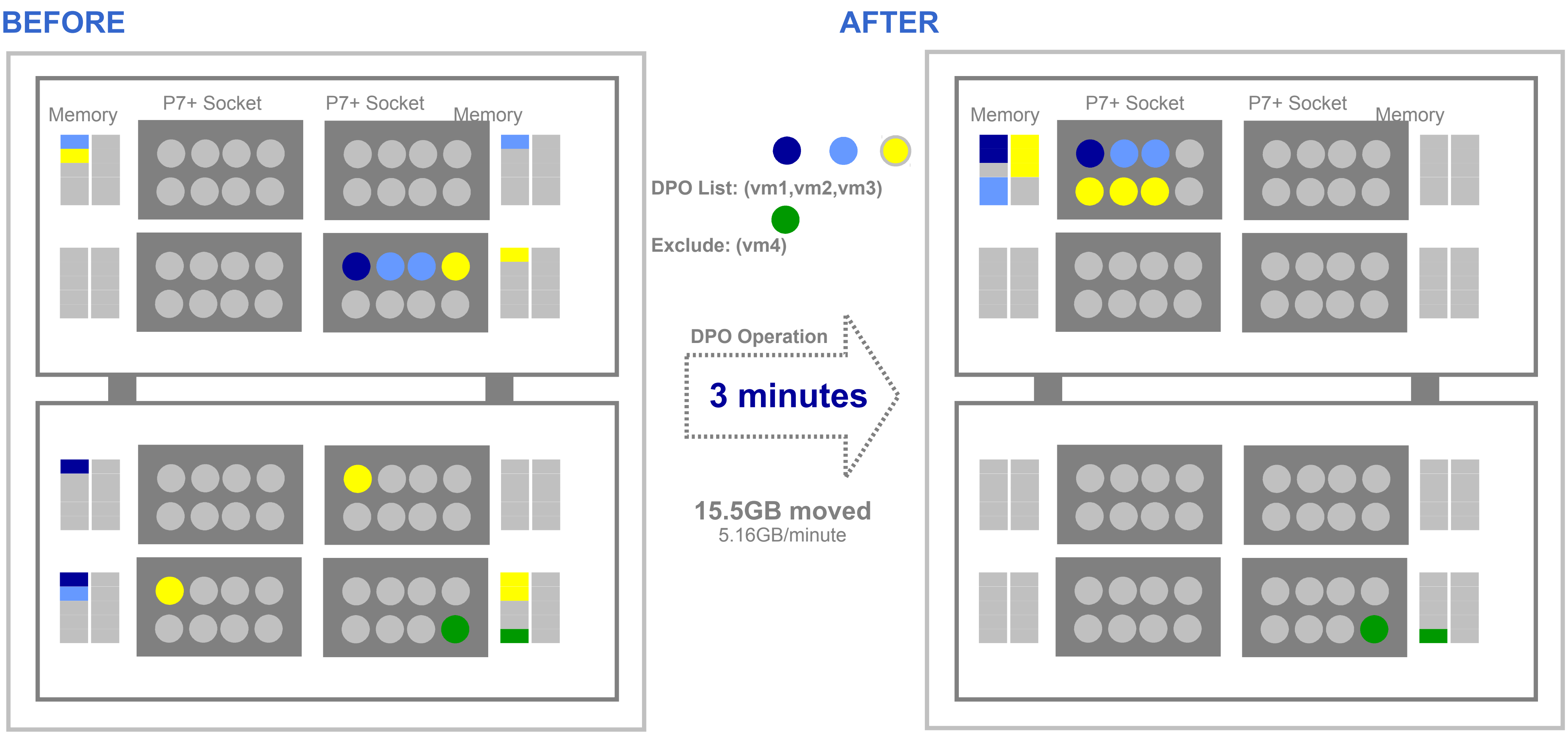
- **Requested/protected partition lists**
  - **Sets of partitions can be prioritized or protected (untouched) by the DPO operation**
  - **DPO should NEVER be used to fix running LPARs that are not "DPO aware"**
  - **Use –xid to exclude running LPARs that are not "DPO aware" (AIX < 6.1.8 or 7.1.2)**
  - **use –id to list "DPO aware" LPARs**

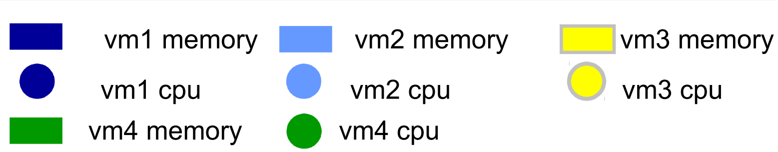
- **Notion of current and potential "affinity score"**
  - **Enables system administrator to make decisions about value of running optimizer**

DPO Effectiveness (P7+ 2Node 9117-MMD)

3 IDLE Partitions Affinitized in 3 minutes

- 15.5 GBs moved
- Total Memory for all 3 partitions involved in DPO operation = 28GB





Configuration Details

Saturn IOC+ 64Core / 1TB Memory
LMB Size=256MB

VM1:1CPU, 4GB **VM2:**2CPU, 8GB **VM3:**3CPU, 16GB

DPO Command Reference

*Executed From HMC Command Line

//show current system score

lsmemopt -m <sysname> -o currscore

curr_sys_score=76

//project score if DPO is executed

lsmemopt -m <sysname> -o calcscore

curr_sys_score=76,predicted_sys_score=86,requested_lpar_ids=none, protected_lpar_ids=none

//execute DPO on all partitions

optmem -m <sysname> -t affinity -o start

//execute DPO on all partitions, lpar ID3 has highest priority

optmem -m <sysname> -t affinity -o start –id 3

//execute DPO on all partitions, except lpar IDs 3 4 and 5

optmem -m <sysname> -t affinity -o start –xid 3,4,5

//show progress

lsmemopt -m <sysname>

opt_id=3,in_progress=1,status=In progress,type=affinity, progress=47, requested_lpar_ids=none,protected_lpar_ids=none,impacted_lpar_ids=none

//abort DPO

optmem -m <sysname> -o stop

IBM Systems Lab Services and Training