RED HAT SUMMIT

BOSTON, MA
JUNE 23-26, 2015

OPEN HYPERCONVERGED INFRASTRUCTURE

Presenters
Sean Murphy – Storage Prod Mgr
Paul Cuzner – Storage Tech Mktg

Date 06.24.15

Your Presenters

Sean Murphy

- Product Mgr with the Storage BU

Paul Cuzner

- Technical Mktg lead with the Storage BU

INTRODUCTION

AGENDA

- The What & the Why
- Under the Hood
- Q&A

SETTING THE STAGE

- Red Hat is an enterprise infrastructure provider
  - Always looking thru a next-gen IT *solution* lens

- HCI space is H-O-T

- We are working upstream toward oVirt / Gluster HCI integration

The What & Why

The What

Hyper-convergence

- Collapse compute, storage into small footprint
  - ...scalable resource pool, with redundancy / high availability
  - ...eliminate the need for discrete components
- Value Prop is centered around simplicity, TCO
- User profile – mid-market to large enterprise

oVirt-GlusterFS

- Toward an Open Source Hyperconverged platform
  - ...Linux, oVirt, Gluster – integrated.

The What

A proven, general purpose scale-out distributed storage

- Unified namespace, supporting thousands of clients
- Gluster runs completely in user space

oVirt-Gluster Integration

- Native Gluster Storage Domain type
- Enable Volume Management from oVirt WebAdmin and REST-API

Why?

“why it just makes sense”

In a word: Simplify!

- Single team managing infrastructure
- Simplify ITIL flows, improve project delivery turnaround
- Simplify hardware planning, procurement
- Simplify hardware deployment, mgmt
- A ‘level playing field’ for capex
- Single budget provides compute and storage
- Hardware flexibility = no lock-in

Why?

The market & the goods

User / Market Demand

– From SMB to Enterprise - planning, deploying

– Hot market, with growing demand for an open HCI value prop

– HCI market grew 162.3% in 2014 (to a market value of $373 million)

– Forecast: up 116% in 2015 (to reach $807 million)

...*within two years, over 50 percent of enterprises across all sectors will use some form of HCI to run their VMs*

Best of Breed Components (greater than the sum of the parts)

– oVirt a robust Linux virt platform

– Gluster technology an industry-proven distributed storage system

UNDER THE HOOD

SOLUTION STRATEGY

- Systemd resource management
- Data security
- Multiple layers of cache
- QoS features (disk profiles)
- Libgfapi support in qemu-kvm optimizes the i/o path
- Offload data reconstruction to hardware
- Hardware freedom/choice
- Auto data rebalance
- Mixed SATA/SAS and SSD

COMPONENT OVERVIEW

- oVirt 3.6.x
- RHEL 7.x
- LVM
- SSD managed by dmcache
- Hardware RAID
- Glusterfs 3.7.x provides the data layer
- libgfapi
- Synchronous 3-way data replication
- OpenSSL
- Commodity x86-64 servers

GLUSTERFS FEATURES

GLUSTERFS IN A NUTSHELL

Top 5 features;

✔ elastic volumes - grow and shrink, non-disruptively

✔ automatic self healing

✔ modular, userspace architecture based on *translators*

✔ synchronous replication

✔ no central meta-data server = no single point of failure

DATA PLACEMENT

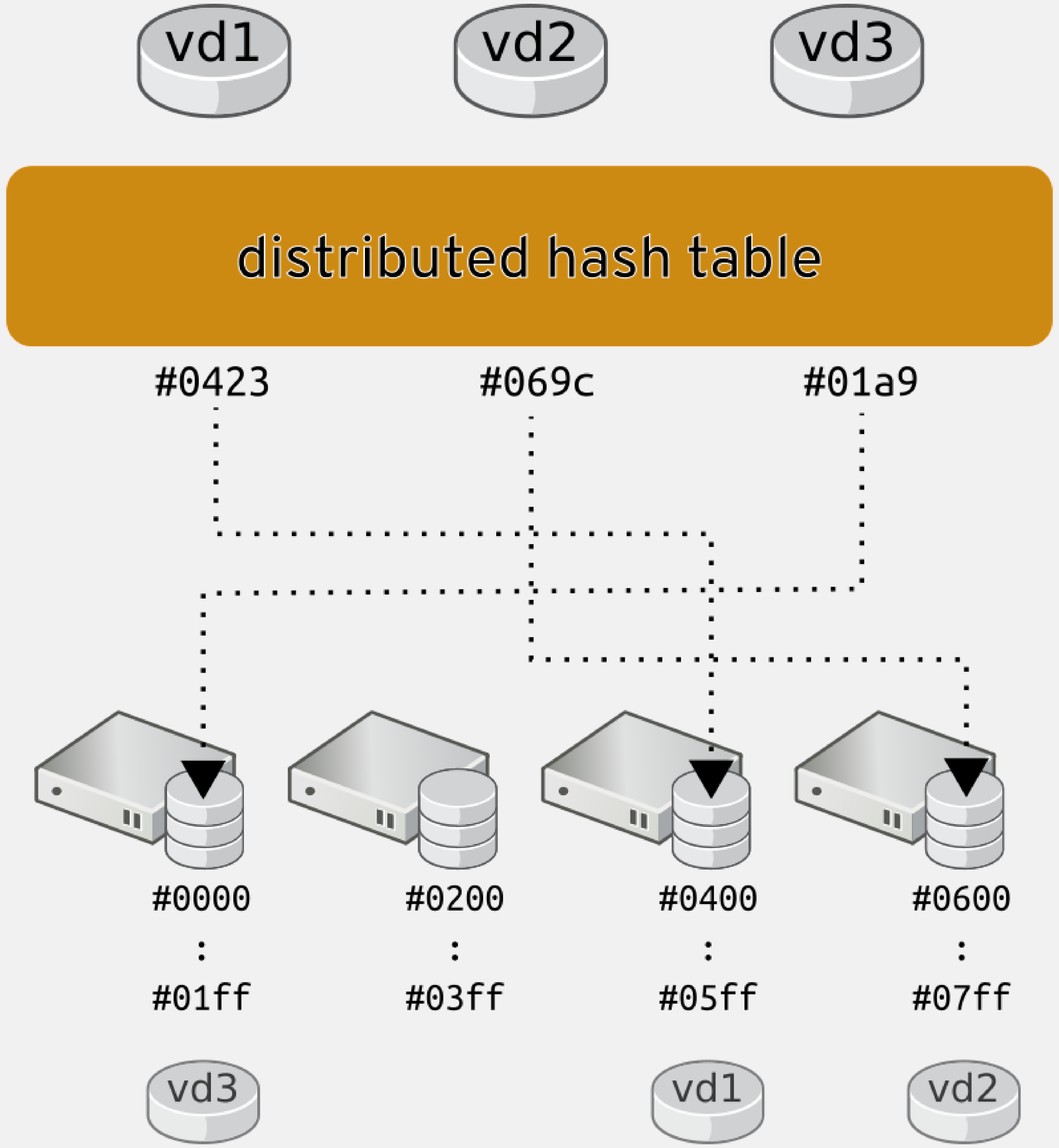
elastic hash algorithm generates a hash value from a file path

each brick within the cluster is assigned a hash range

the client is aware of the hash ranges from each brick

direct data path

each file holds metadata in xattr



DATA LOCALITY

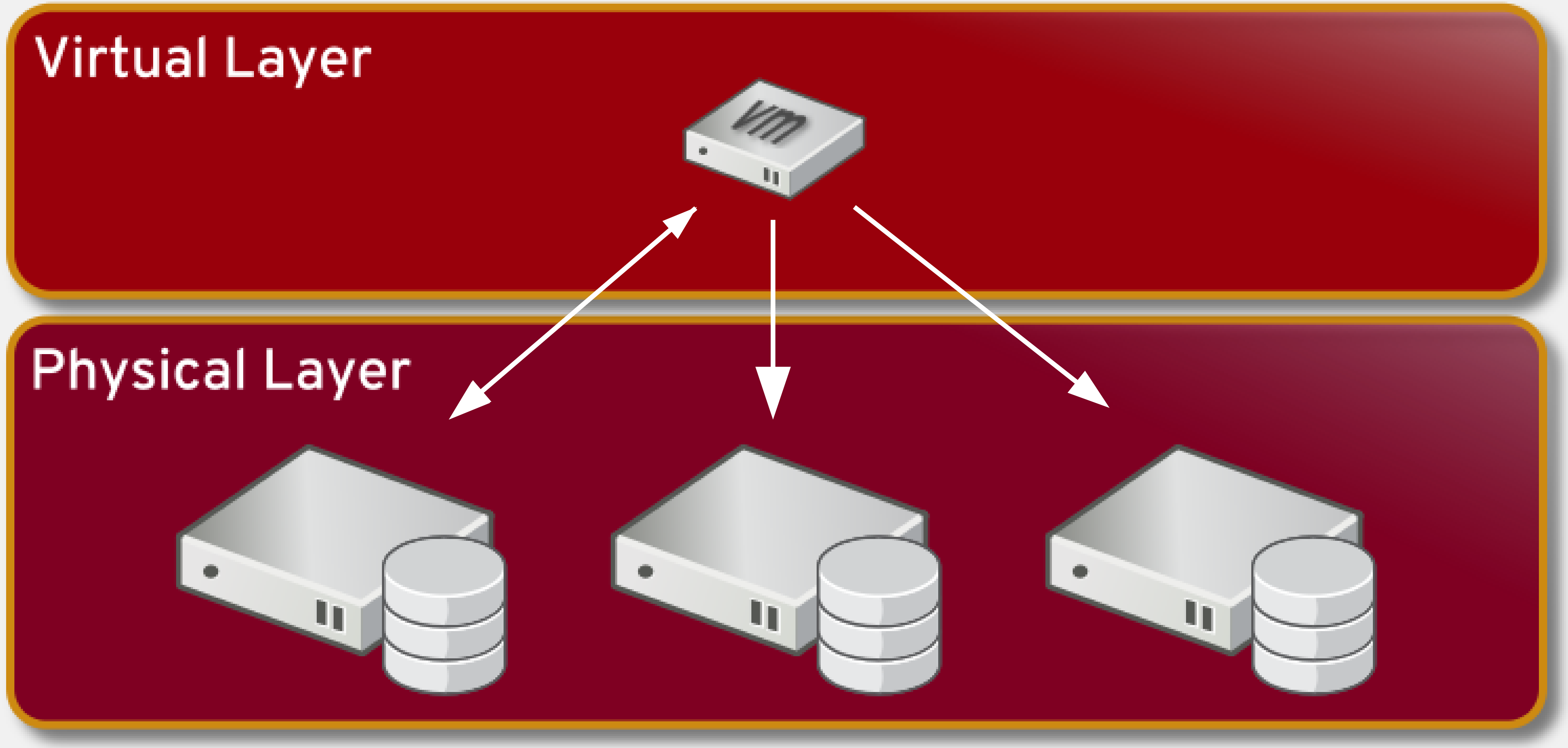
First read generates a pre-op request to each brick that owns the vdisk

first to respond is the winner!

local brick on the executing node is chosen

reads focus on the local brick

in the event the local brick is lost, failover to one of the other bricks is automatic



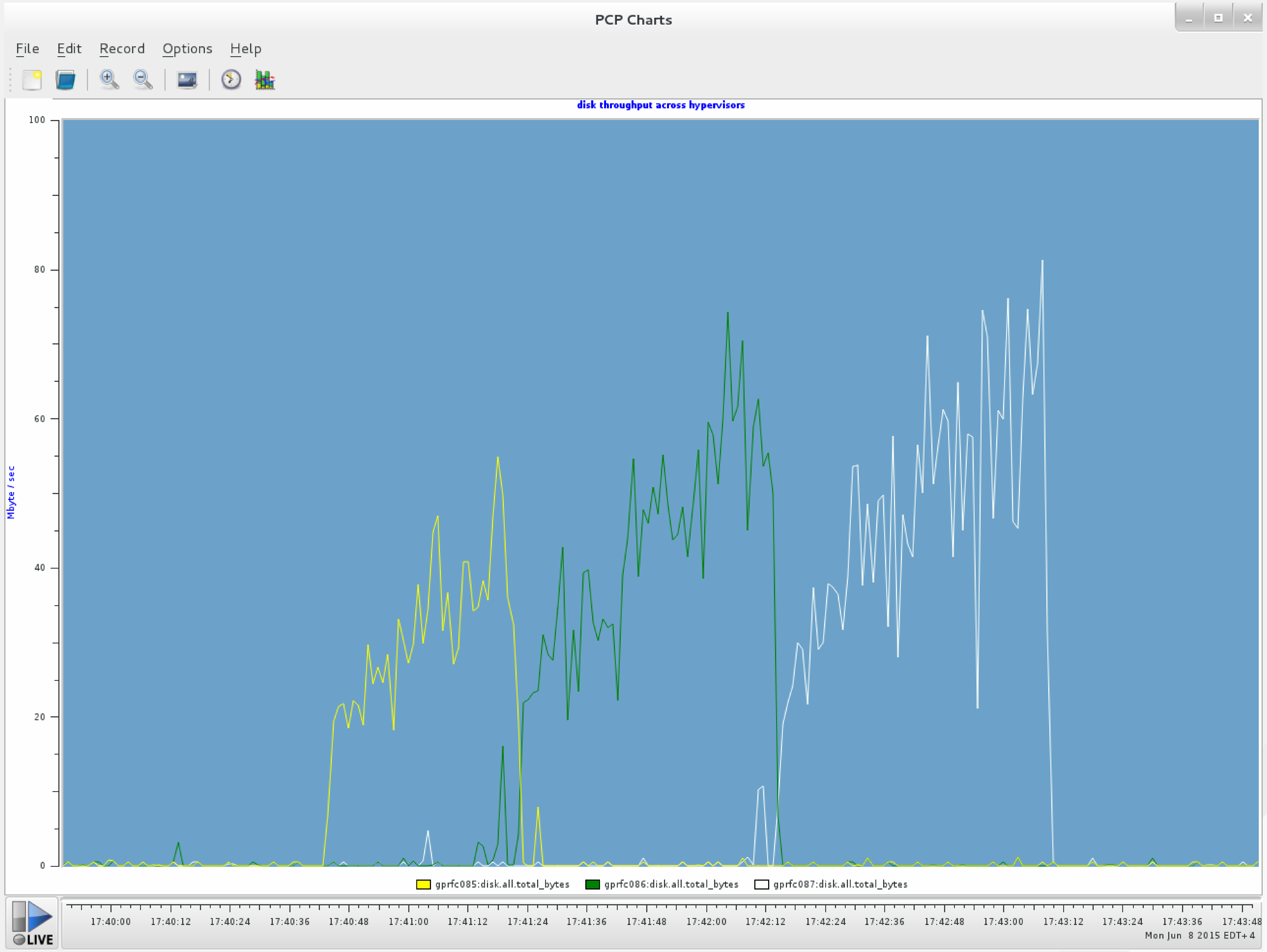
DATA LOCALITY... IN ACTION

RHEL7 vm running fio

- 128k block
- Synchronous
- directio
- 10g file test file

vm live migrated during benchmark run

transition of the brick servicing I/O matches the vm's host

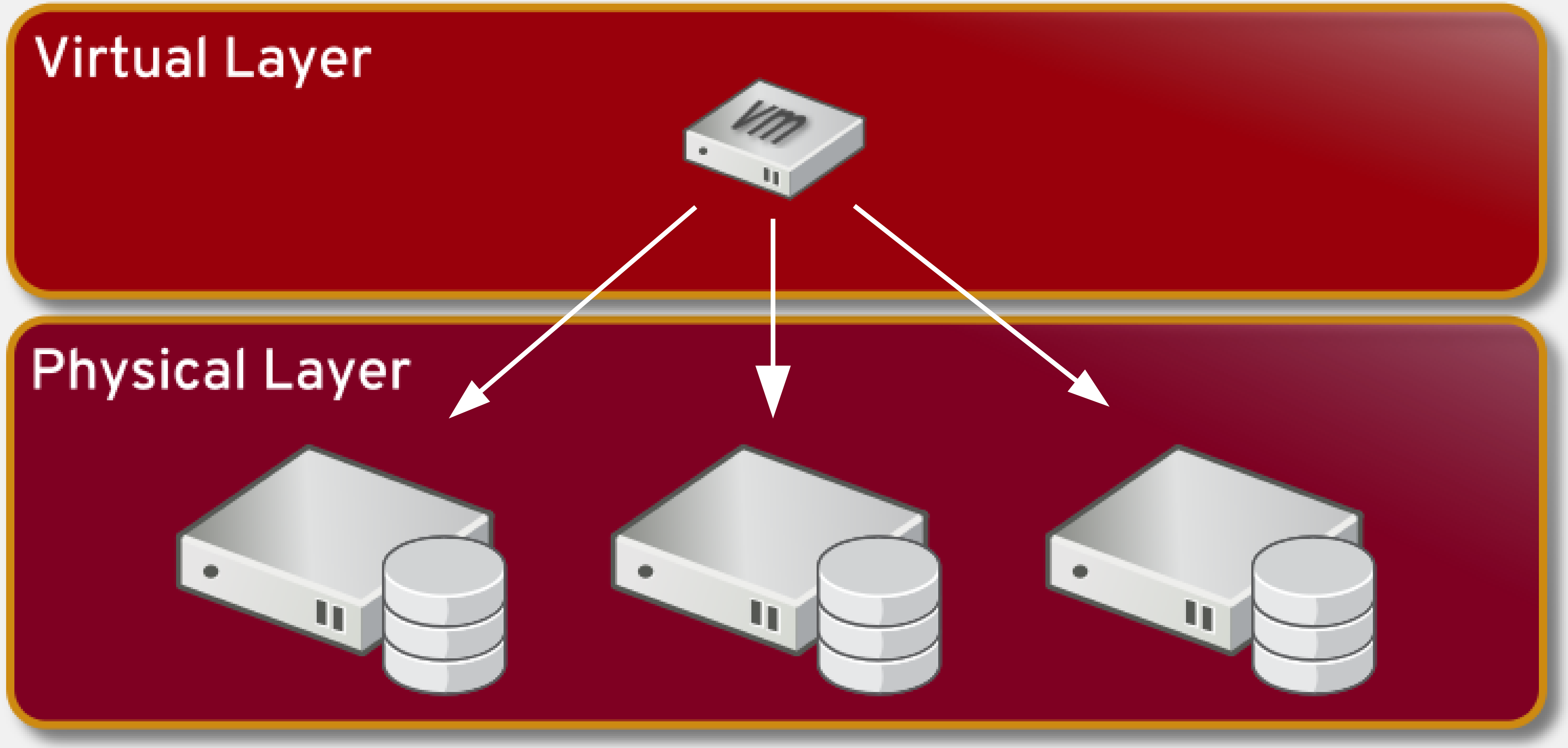


DATA INTEGRITY

Data must be protected across nodes at all times

All writes MUST be written to multiple nodes

glusterfs does this using synchronous replication

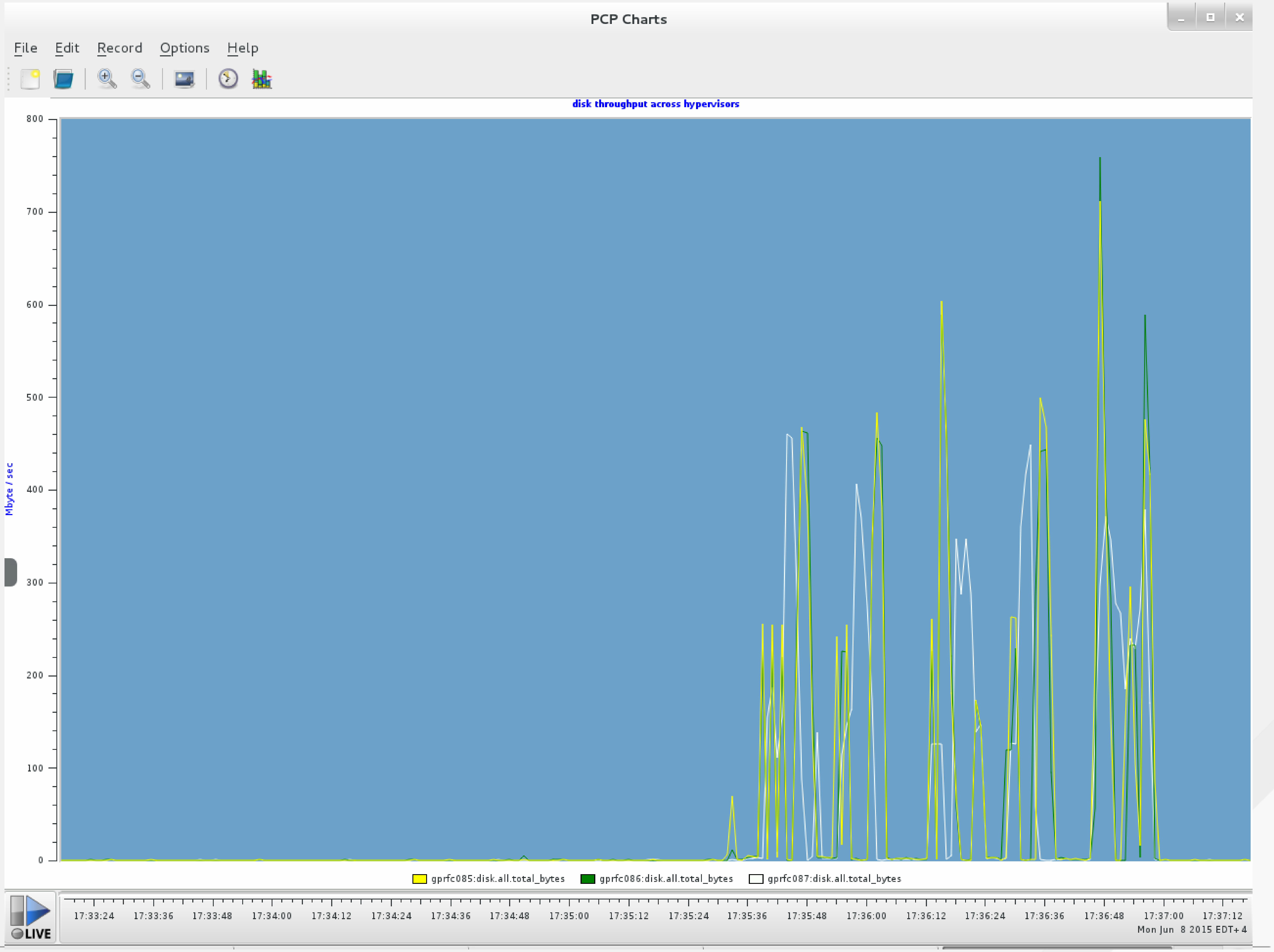


DATA INTEGRITY... IN ACTION

100% write workload

fsync across systems are well aligned

ensures 3 copies of the data are consistent and available



DATA RECOVERY – AUTOMATED SELF HEAL

Maintain Data Redundancy

- grace period or timeout
- data re-replicated

Considerations

- available capacity is reduced
- re-replicate may represent additional workload
- problematic for small clusters

Limit impact to active workload

- track changes
- apply changes when node/disk is available again
- more usable capacity

Considerations

- during the outage, data redundancy goal is not met
- self heal only starts once the node or replacement is brought back into the cluster

PERFORMANCE FEATURES

EXPLOITING CACHE

physical I/O = latency

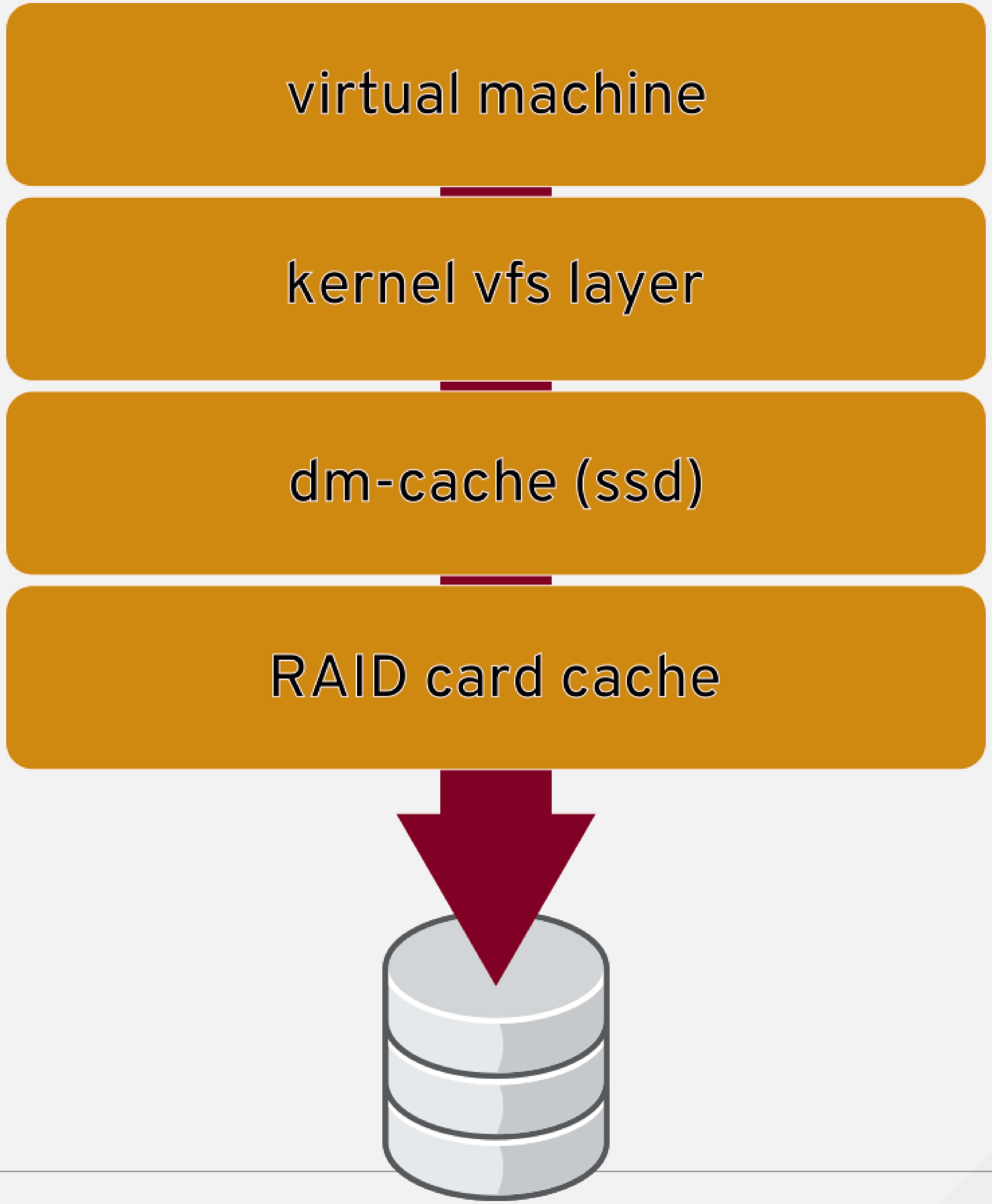
write latency can’t be avoided

cache reduces read latency

all vdisks are files...vfs

cloned disks may fit in page cache

dmcache supports random read



virtual machine

kernel vfs layer

dm-cache (ssd)

RAID card cache

CACHE EFFECTS

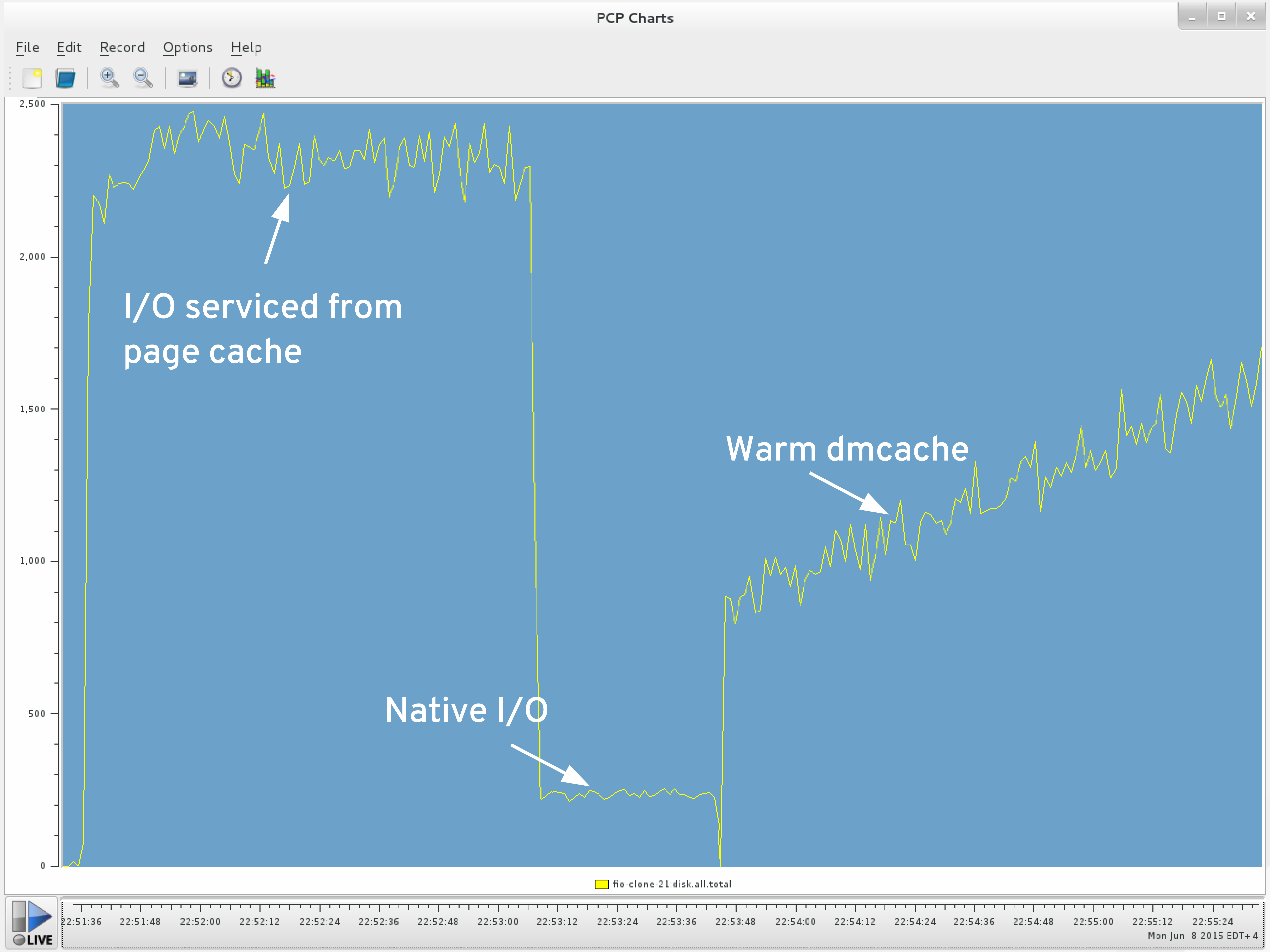
Test Conditions

- Single vm
- fio random read workload
- 8k blocksize
- sync / directIO
- single thread
- 10g dataset

buffer hits : < 0.5ms

disk I/O : ~ 5-6ms

warm dmcache : <=1ms



I/O serviced from page cache

Warm dmcache

Native I/O

MEASURING CACHE EFFECTIVENESS

```
[root@hypervisor1 ~]# pcp -h localhost dmcache
@ Mon Jun  8 20:02:45 2015 (host hypervisor1.lab.redhat.com)

---device--- ---%used--- ---------reads--------- --------writes---------
meta  cache    hit   miss  ratio            hit    miss ratio
Bricks-raid5  0.7%  7.3%  425.49   10.86   97.5%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  399.95   16.79   95.7%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  372.50   8.89    97.9%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  324.19   15.81   97.6%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  409.76   7.90    97.2%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  417.95   11.86   99.1%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  356.69   12.84   94.8%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  354.49   16.79   95.0%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  332.97   6.92    98.0%  0.00 0.00   0%
Bricks-raid5  0.7%  7.3%  326.73   7.90    96.8%  0.00 0.00   0%
```

SECURITY CONSIDERATIONS

PROTECTING VIRTUAL DISKS

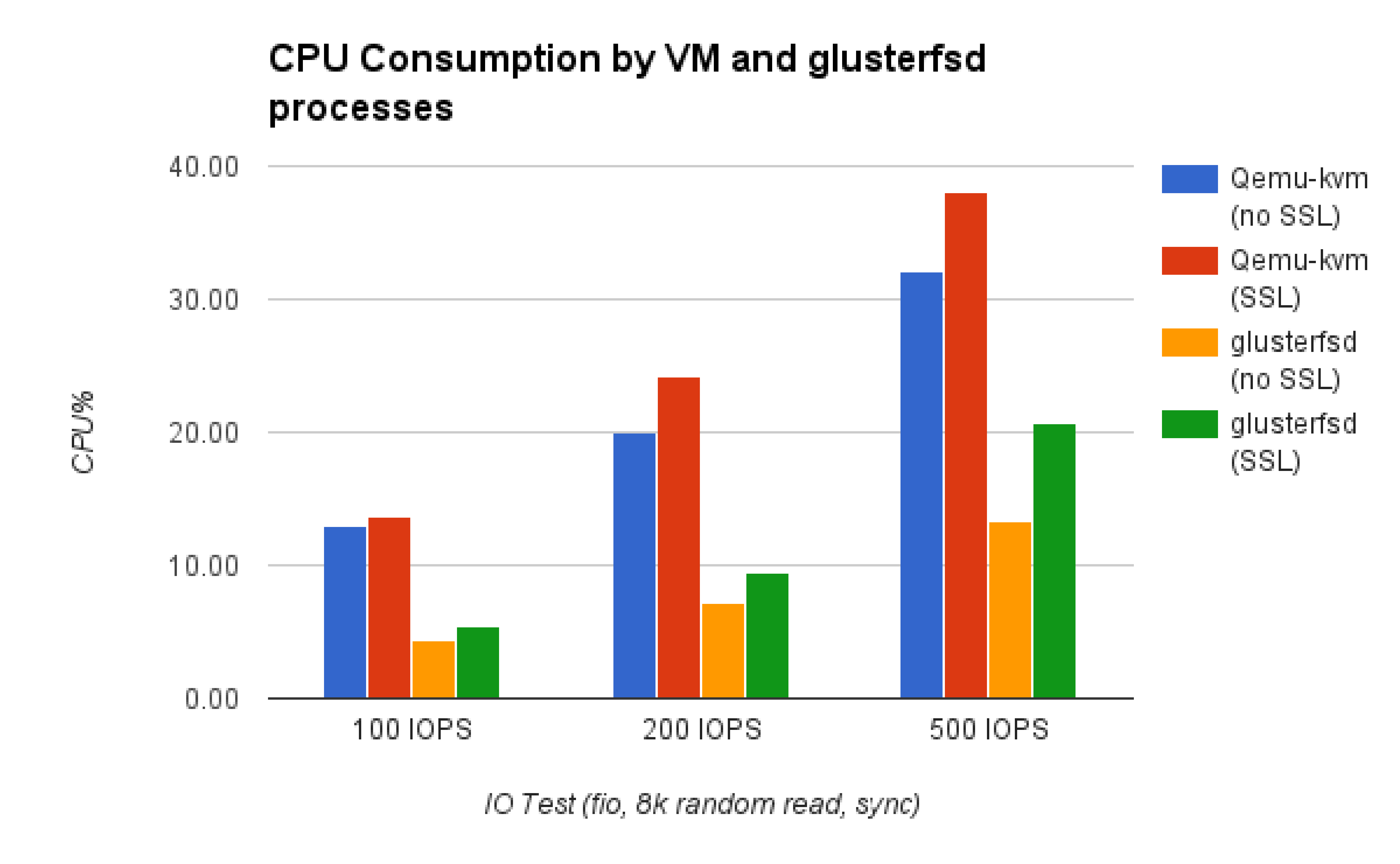
Vdisks accessed via a network

- Non-routable VLAN
- auth.allow

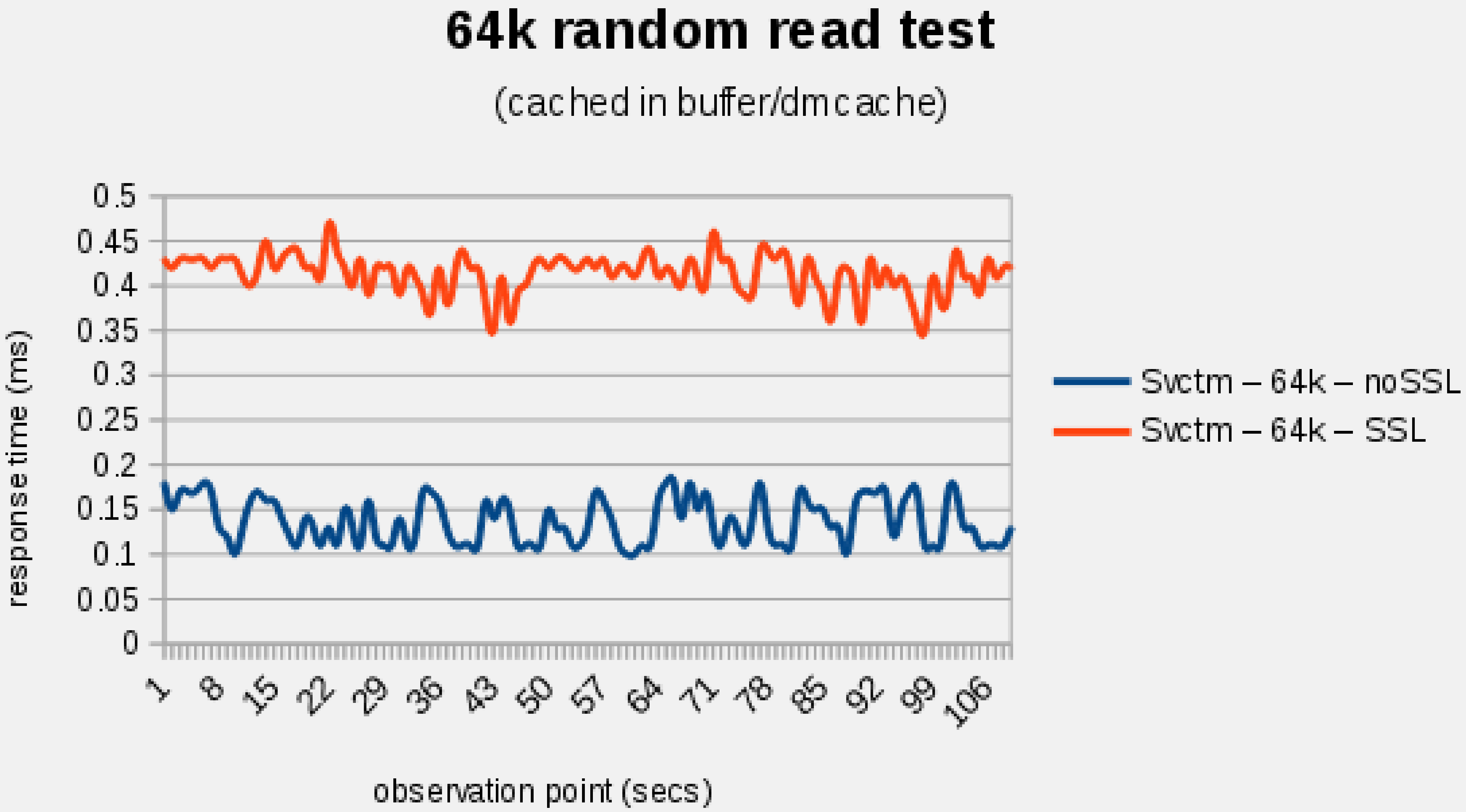
Data path encryption

The cpu cost of encryption

Security headache becomes a planning exercise



I/O LATENCY IMPACT*



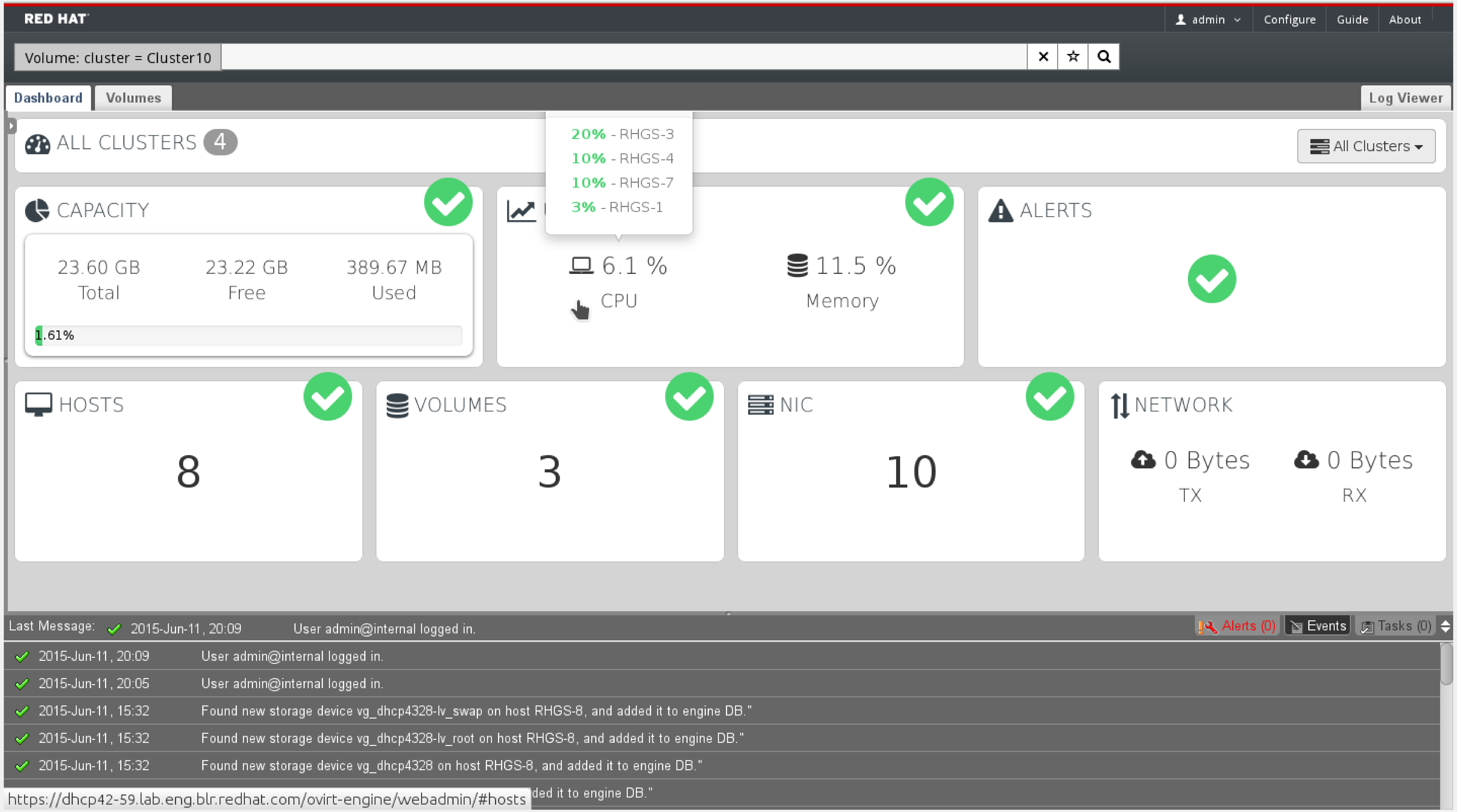
* More validation to come!

INTEGRATION

MANAGEMENT CONSIDERATIONS

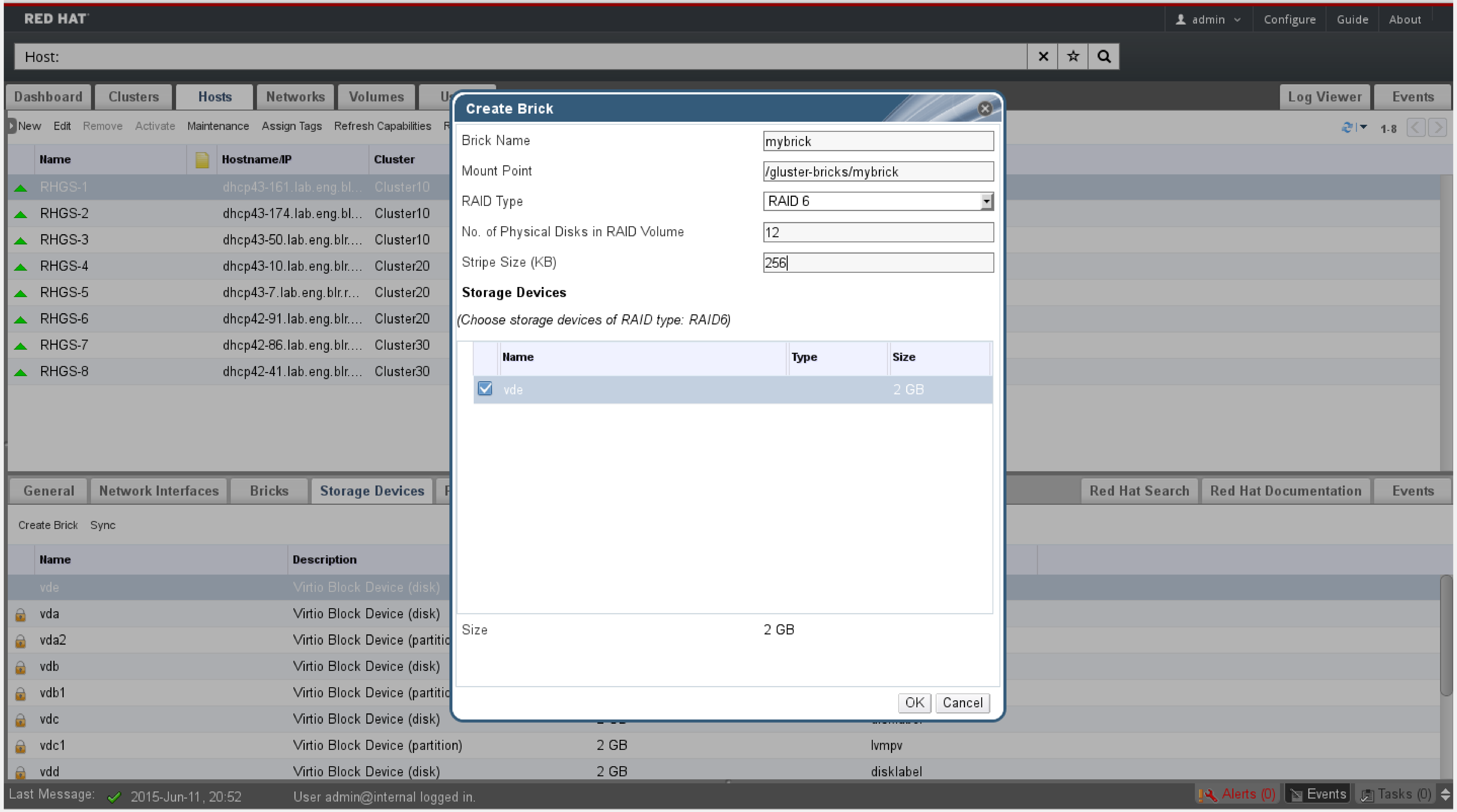
- Management engine options;
  - self-hosted running natively on glusterfs
  - remote options
- Dashboard provides an “at a glance” view
- Storage and compute managed within a single interface
- New disks added through the UI
- SSD integration is read only

THE DASHBOARD VIEW



```
Last Message:  ✓ 2015-Jun-11, 20:09    User admin@internal logged in.
✓ 2015-Jun-11, 20:09    User admin@internal logged in.
✓ 2015-Jun-11, 20:05    User admin@internal logged in.
✓ 2015-Jun-11, 15:32    Found new storage device vg_dhcp4328-lv_swap on host RHGS-8, and added it to engine DB."
✓ 2015-Jun-11, 15:32    Found new storage device vg_dhcp4328-lv_root on host RHGS-8, and added it to engine DB."
✓ 2015-Jun-11, 15:32    Found new storage device vg_dhcp4328 on host RHGS-8, and added it to engine DB."
https://dhcp42-59.lab.eng.blr.redhat.com/ovirt-engine/webadmin/#hosts ded it to engine DB."
```

DISK MANAGEMENT



Name	Hostname/IP	Cluster
RHGS-1	dhcp43-161.lab.eng.bl...	Cluster10
RHGS-2	dhcp43-174.lab.eng.bl...	Cluster10
RHGS-3	dhcp43-50.lab.eng.blr....	Cluster10
RHGS-4	dhcp43-10.lab.eng.blr....	Cluster20
RHGS-5	dhcp43-7.lab.eng.blr.r...	Cluster20
RHGS-6	dhcp42-91.lab.eng.blr....	Cluster20
RHGS-7	dhcp42-86.lab.eng.blr....	Cluster30
RHGS-8	dhcp42-41.lab.eng.blr....	Cluster30

Create Brick

Brick Name: mybrick

Mount Point: /gluster-bricks/mybrick

RAID Type: RAID 6

No. of Physical Disks in RAID Volume: 12

Stripe Size (KB): 256

Storage Devices

(Choose storage devices of RAID type: RAID6)

	Name	Type	Size
☑	vde		2 GB

Size: 2 GB

OK Cancel

Name	Description		
vde	Virtio Block Device (disk)		
vda	Virtio Block Device (disk)		
vda2	Virtio Block Device (partitio		
vdb	Virtio Block Device (disk)		
vdb1	Virtio Block Device (partitio		
vdc	Virtio Block Device (disk)		
vdc1	Virtio Block Device (partition)	2 GB	lvmpv
vdd	Virtio Block Device (disk)	2 GB	disklabel

Last Message: ✓ 2015-Jun-11, 20:52 User admin@internal logged in.

Alerts (0) Events Tasks (0)

ADMINISTRATION

- Web GUI (oVirt)
- REST / API
- oVirt python SDK + gluster bindings for libgfapi
- Integration example - vm2brick tool
- Support Tools
  - Performance co-pilot
  - dmcache CLI reports
  - Common sysadmin perf tools (e.g., iostat, vmstat, iotop)
  - Ovirt data warehouse for reporting, trending, analysis

[1] https://github.com/pcuzner/vm2brick

NEW GLUSTERFS FEATURES

FEATURES JUST LANDED IN v3.7

Glusterfs 3.7.x introduces

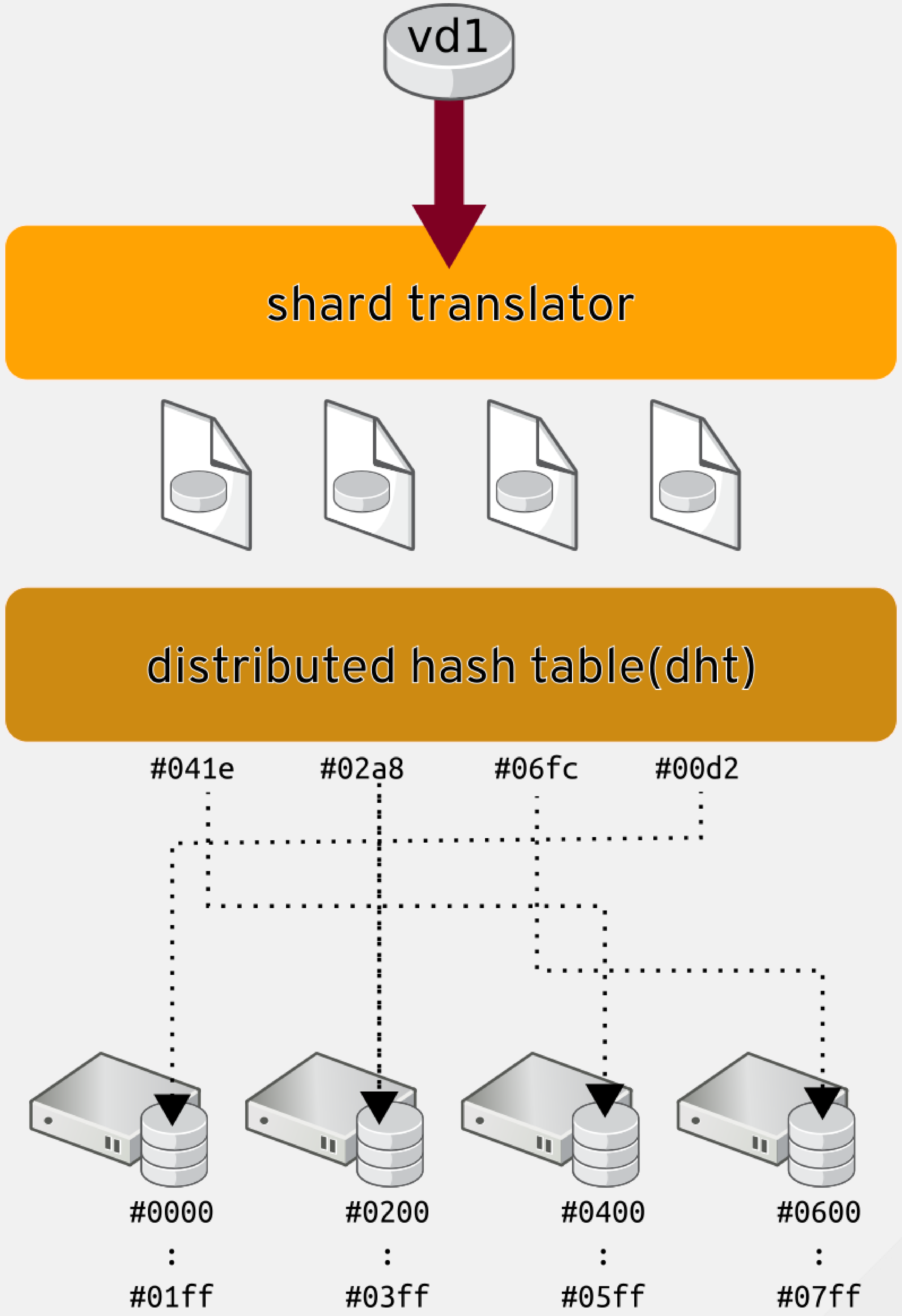
- Sharding - enhanced granularity for
  - self heal
  - rebalance
  - geo-replication
- arbiter volumes
- rebalance performance enhancements
- multi-threaded epoll

...and more! [1]

[1] http://blog.gluster.org/2015/05/glusterfs-3-7-0-has-been-released-introducing-many-new-features-and-improvements-2/

FEATURE FOCUS - SHARDING

- shard is a translator that sits client-side
- configurable shard size (default 4MB)
- larger files = more shards = wide striping
- shards get distributed across bricks like normal files



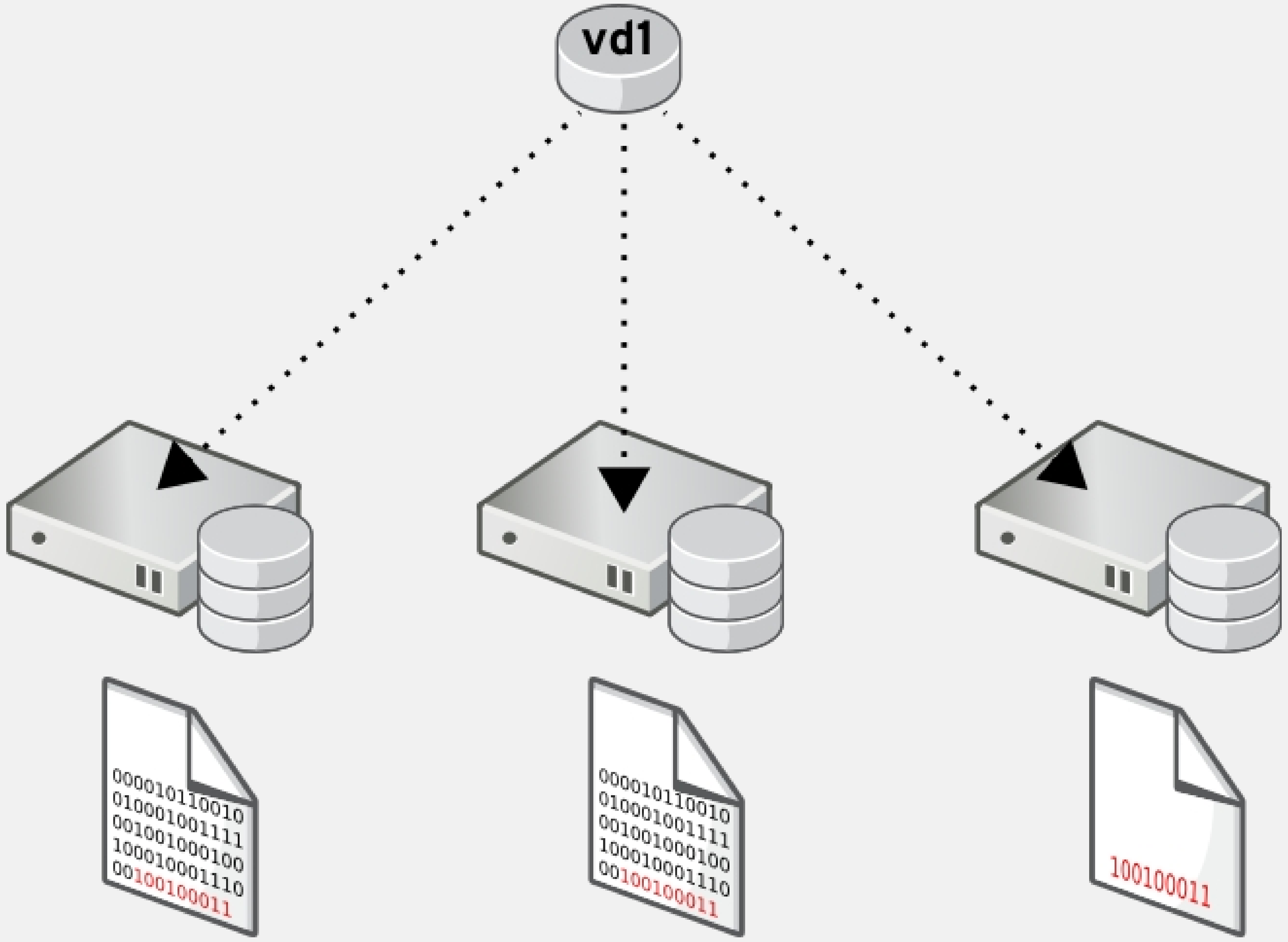
FEATURE FOCUS – ARBITER VOLUMES

The challenge of distributed storage;

- with only 2 copies - split brain is possible
- with 3 copies - costs go up!

Rather than consume more space, let's address the problem

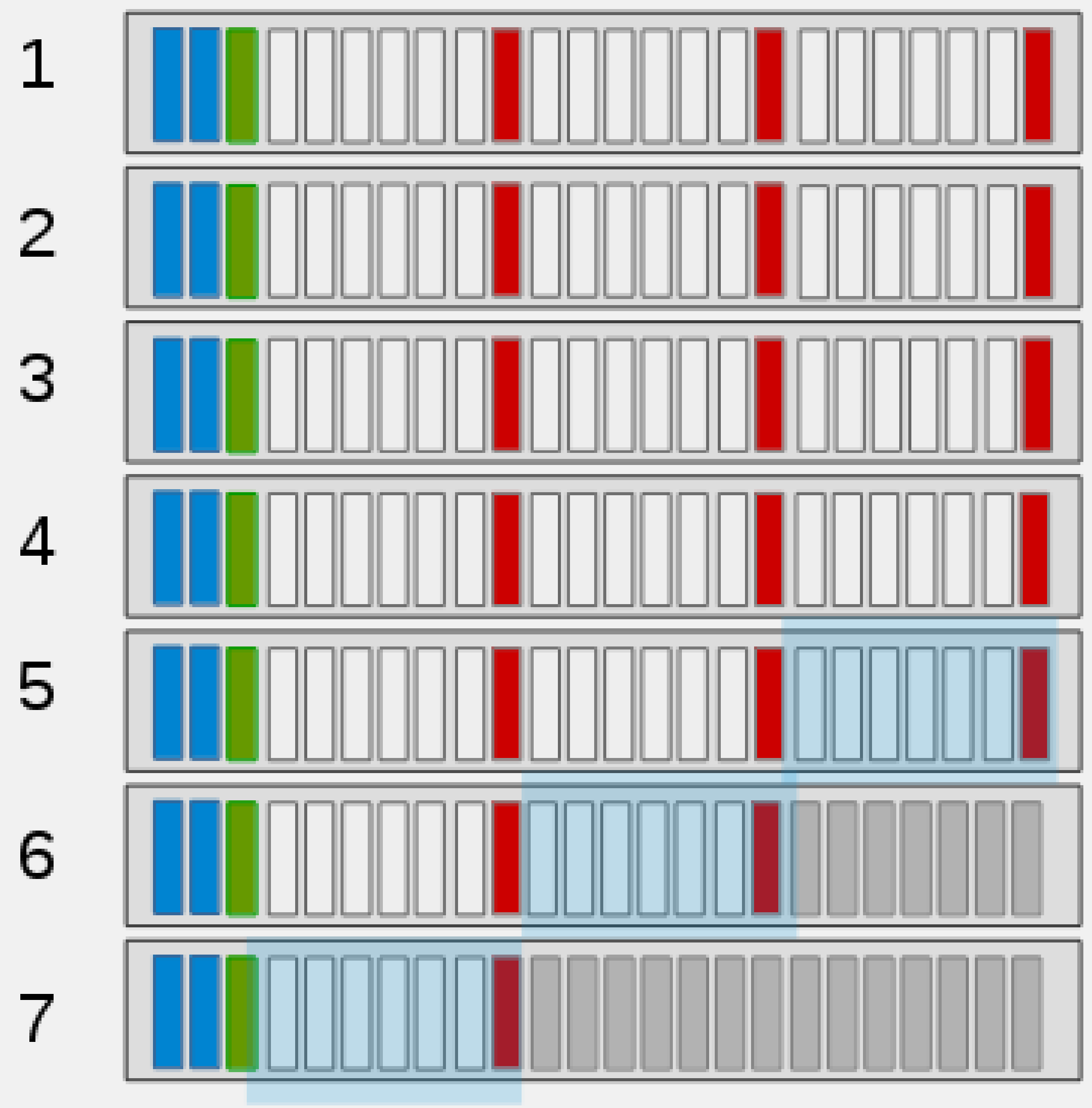
- 2 copies of the data is a must!
- tie-breaker is needed to avoid split brain



DEPLOYMENT SCENARIO

POTENTIAL GROWTH MODEL

Configuration Phase : Growth Increment #4



Hypervisors
7

Raw Capacity (VMs)
108TB

Usable Capacity (VMs)
36TB

Replica sets
6

Bricks
18

- 🟦 OS
- 🟩 Hot Spare
- ▨ Unused HDD Slot
- □ 10k HDD
- 🟥 SSD
- □□□□□□🟥 Capacity Increment
- ▨▨▨▨▨▨▨ empty drive bays

Further Info...

- http://www.ovirt.org/Features/Self_Hosted_Engine_Gluster_Support
- http://www.ovirt.org/Features/Self_Hosted_Engine_Hyper_Converged_Gluster_Support
- http://www.ovirt.org/Features/GlusterFS-Hyperconvergence
- https://fosdem.org/2015/schedule/event/hyperconvergence/
- http://www.ovirt.org/images/6/6c/2015-ovirt-glusterfs-hyperconvergence.pdf

Q&A

RED HAT
SUMMIT

LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.