STO1279

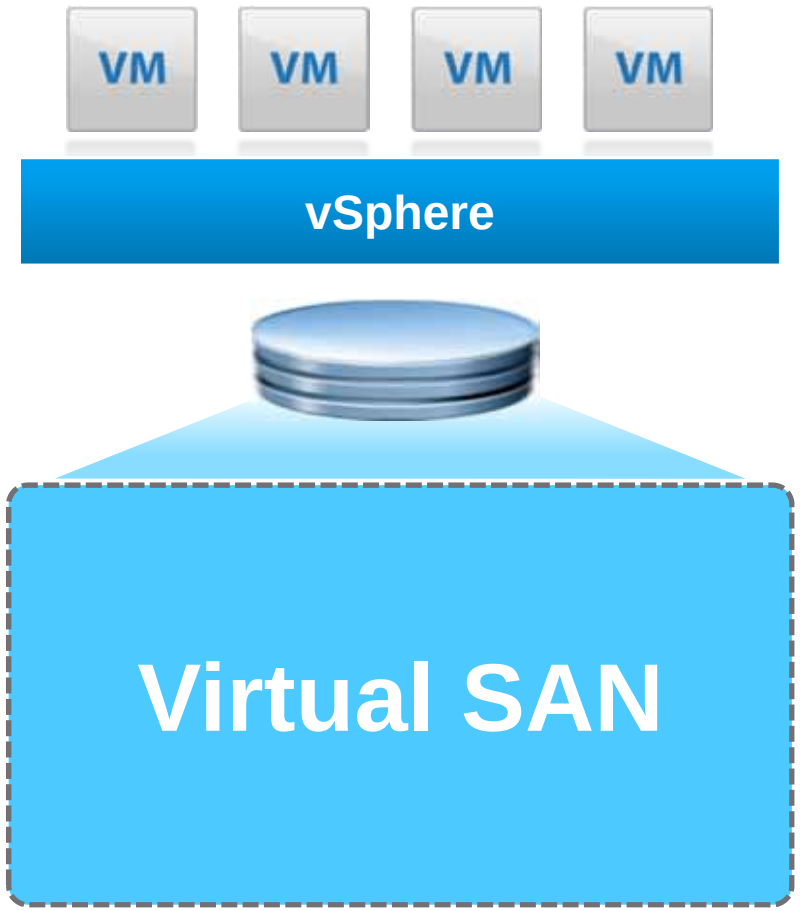
Virtual SAN Architecture Deep Dive

Christos Karamanolis, VMware, Inc
Christian Dickmann, VMware, Inc

Disclaimer

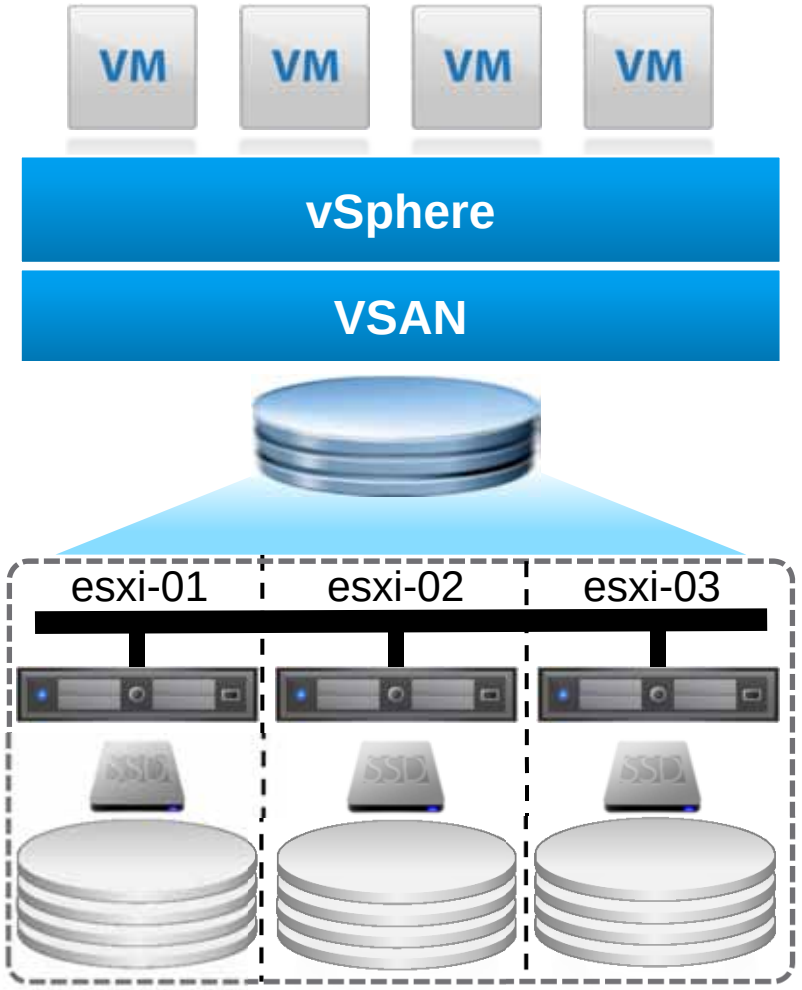
- This presentation may contain product features that are currently under development.
- This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.
- Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.
- Technical feasibility and market demand will affect final delivery.
- Pricing and packaging for any new technologies or features discussed or presented have not been determined.

Virtual SAN: Product goals



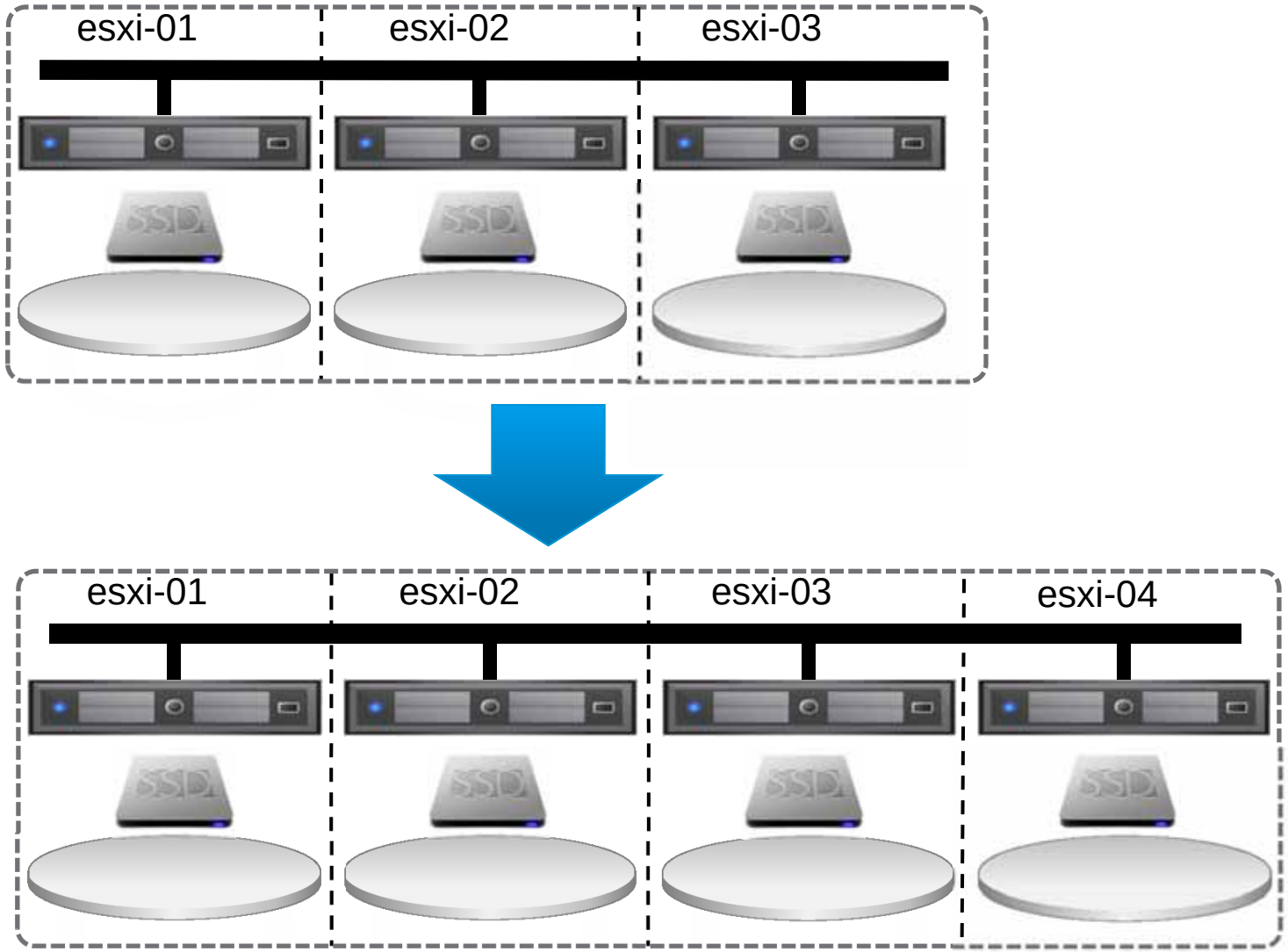
1. Targeted customer: **vSphere admin**
2. Compelling Total Cost of Ownership (TCO)
   - CAPEX: capacity, performance
   - OPEX: ease of management
3. The Software-Defined Storage for VMware
   - Strong integration with all VMware products and features

What is Virtual SAN?

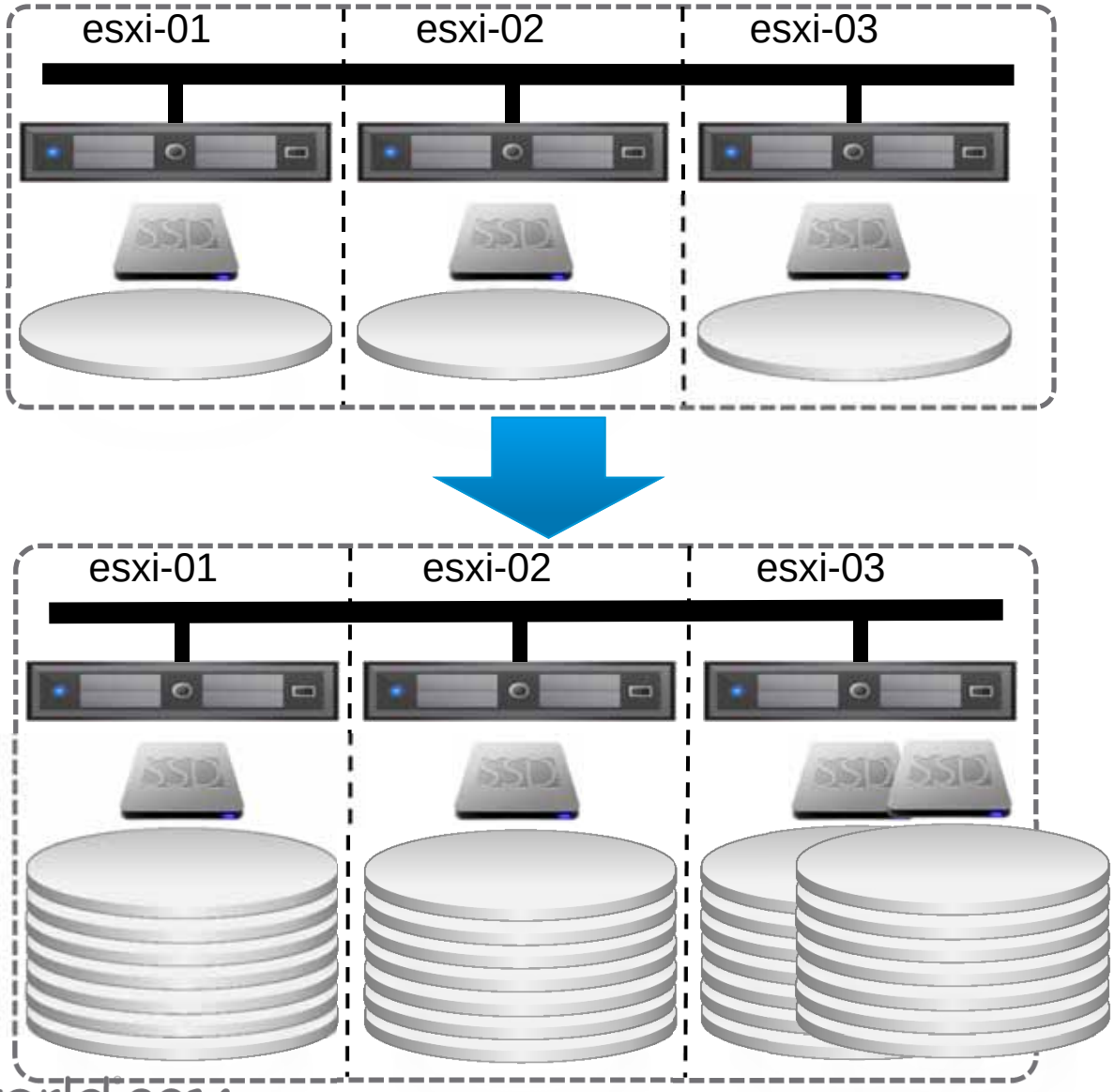


- Software-based storage built in ESXi
- Aggregates local Flash and HDDs
- Shared datastore for VM consumption
- Converged compute + storage
- Distributed architecture, no single point of failure
- Deeply integrated with VMware stack

Virtual SAN Scale Out



Virtual SAN Scale Up



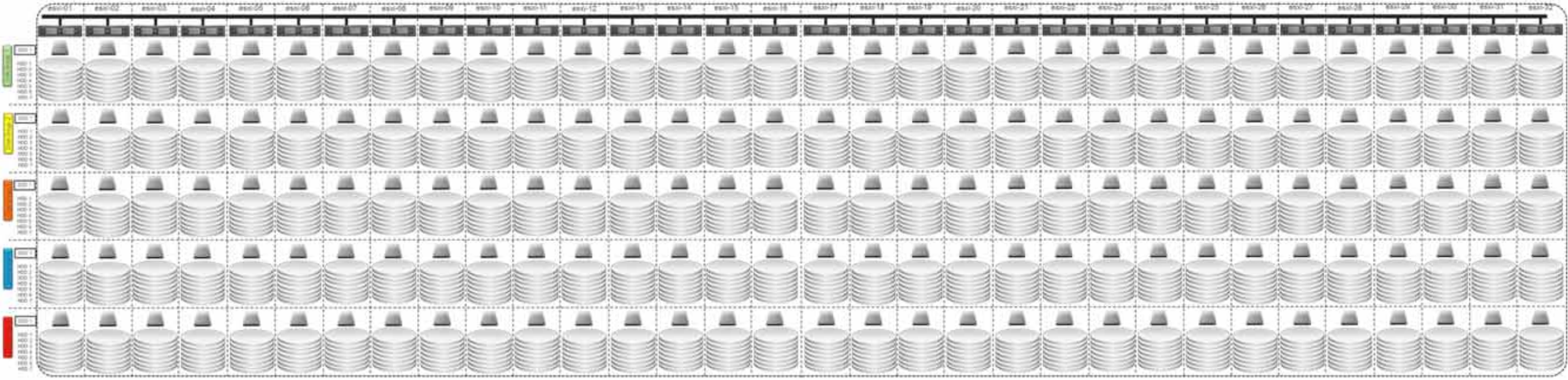
Single Virtual SAN datastore scalability

Cluster: **3 - 32 nodes**; up to 5 SSDs, 35 HDDs per host

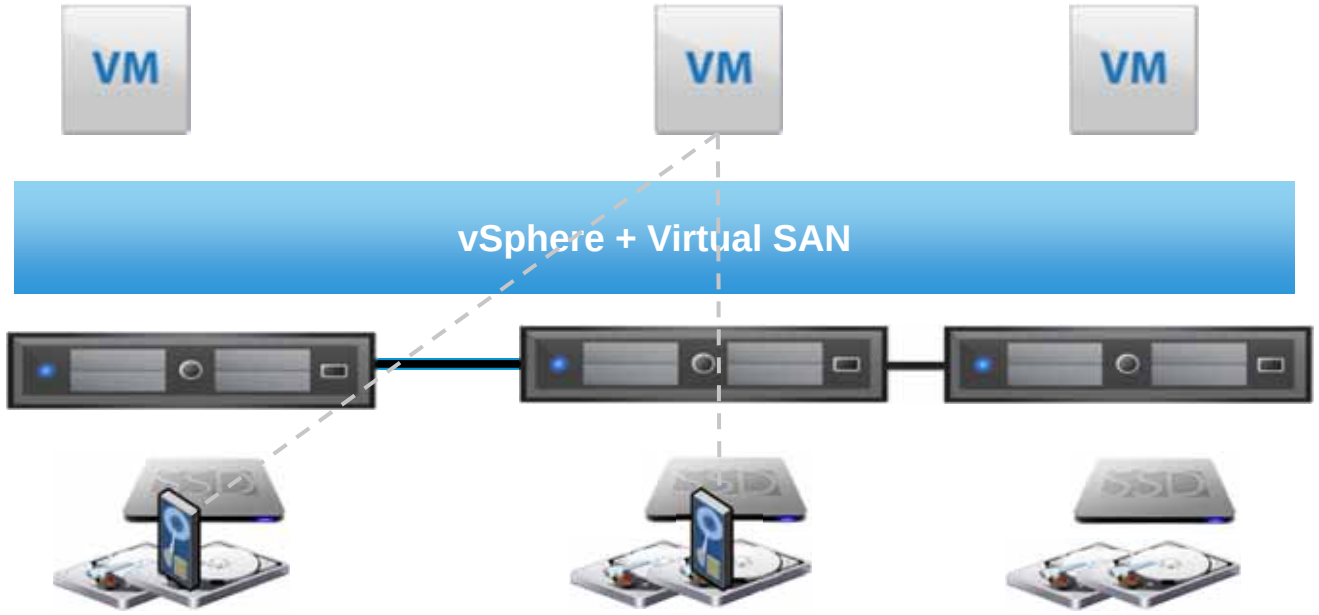
Capacity: **4.4 Petabytes**

Performance: **2M IOPS** – 100% reads

640K IOPS – 70% reads

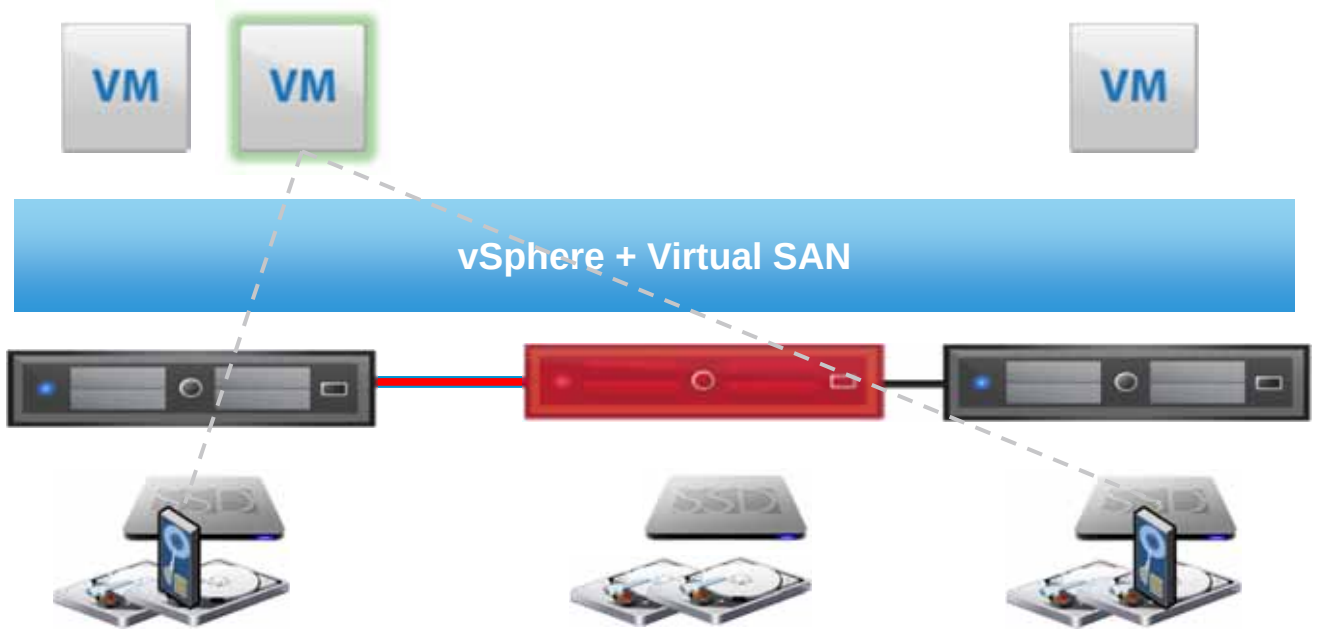


Virtual SAN Is Highly Resilient Against Hardware Failures



- ✓ **Simple** to set resiliency goals via policy
- ✓ Enforced **per VM** and **per vmdk**
- ✓ **Zero data loss** in case of disk, network or host failures
- ✓ **High availability** even during network partitions
- ✓ Automatic, distributed **data reconstruction** after failures
- ✓ Interoperable with vSphere HA and Maintenance Mode

Virtual SAN Is Highly Resilient Against Hardware Failures



- ✓ **Simple** to set resiliency goals via policy
- ✓ Enforced **per VM** and **per vmdk**
- ✓ **Zero data loss** in case of disk, network or host failures
- ✓ **High availability** even during network partitions
- ✓ Automatic, distributed **data reconstruction** after failures
- ✓ Interoperable with vSphere HA and Maintenance Mode

Virtual SAN (VSAN) is NOT a Virtual Storage Appliance (VSA)

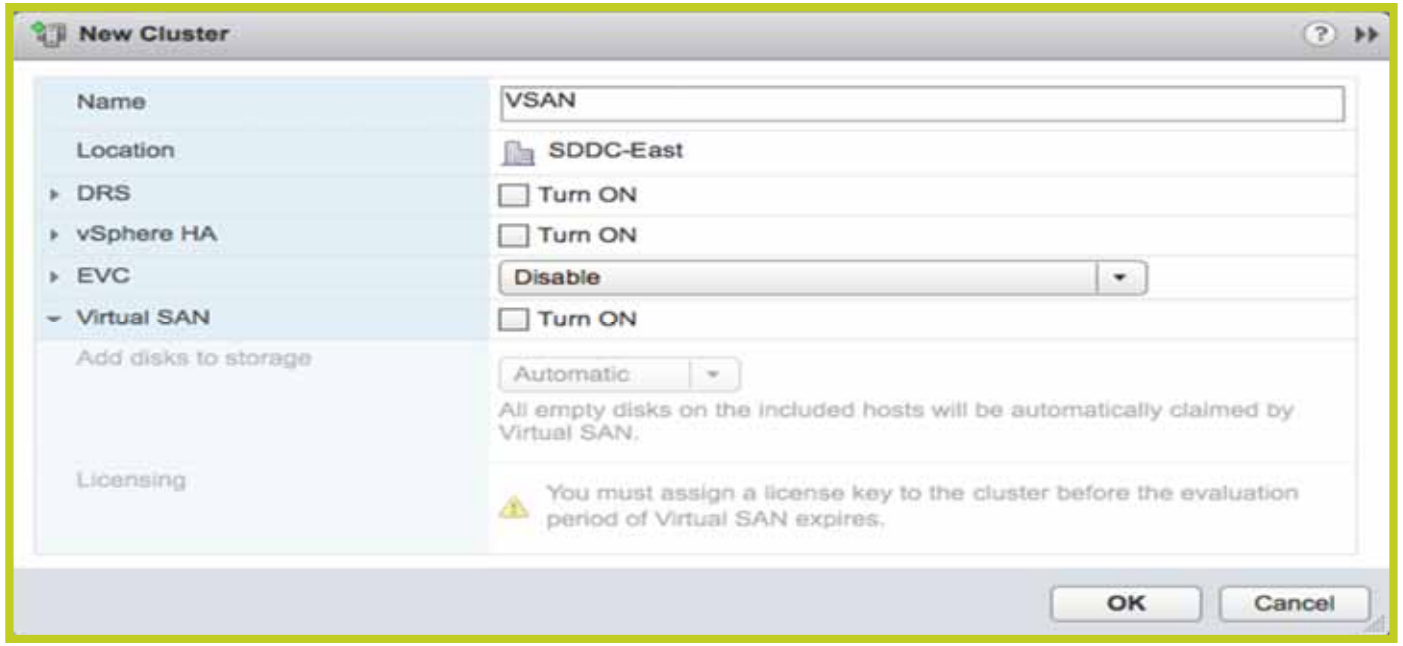
- Virtual SAN is fully integrated with vSphere (ESXi & vCenter)
- Drivers embedded in ESXi 5.5 contain the Virtual SAN smarts
- Kernel modules: **most efficient I/O path**
  - Minimal consumption of CPU and memory
  - Specialized I/O scheduling
  - Minimal network hops, just one storage and network stack
- Eliminate unnecessary **management** complexity (appliances)

Virtual SAN – **Not a VSA**

Virtual SAN – **Embedded** into vSphere

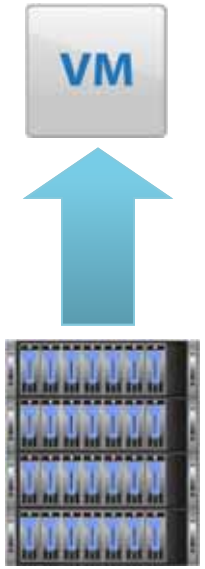
Simple cluster configuration & management

One click away!!!



- Virtual SAN configured in **Automatic mode**, all empty local disks are claimed by Virtual SAN for the creation of the distributed vsanDatastore.
- Virtual SAN configured in **Manual mode**, the administrator must manually select disks to add the the distributed vsanDatastore by creating Disk Groups.

Simplified Provisioning For Applications



Legacy

5. Consume from pre-allocated bin
4. Select appropriate bin
3. Expose pre-allocated bins
2. Pre-allocate static bins
1. Pre-define storage configurations

- ✖ Overprovisioning (better safe than sorry!)
- ✖ Wasted resources, wasted time
- ✖ Frequent Data Migrations

VSAN

1. Define storage policy
2. Apply policy at VM creation

Resource and data services are automatically provisioned and maintained

- ✓ No overprovisioning
- ✓ Less resources, less time
- ✓ Easy to change

Virtual SAN Storage Policies

Storage Policy	Use Case	Value
Object space reservation	**Capacity**	Default 0 Max 100%
Number of failures to tolerate (RAID 1 – Mirror)	**Availability**	Default 1 Max 3
Number of disk stripes per object (RAID 0 – Stripe)	**Performance**	Default 1 Max 12
Flash read cache reservation	**Performance**	Default 0 Max 100%
Force provisioning		Disabled

How To Deploy A Virtual SAN Cluster

Software + Hardware

Component Based

Choose individual components …

- ***Any*** Server on vSphere Hardware Compatibility List

- SSD or PCIe

- SAS/NL-SAS/ SATA HDDs

- HBA/RAID Controller

…using the VMware Virtual SAN Compatibility Guide (VCG) (1)

Virtual SAN Ready Node

40 OEM validated server configurations ready for Virtual SAN deployment (2)

CISCO · DELL · FUJITSU · hp · HITACHI Inspire the Next · HUAWEI · IBM · SUPERMICRO

VMware EVO:RAIL

Hyper-Converged Infrastructure

A Hyper-Converged Infrastructure Appliance (HCIA) for the SDDC

Each EVO:RAIL HCIA is pre-built on a qualified and optimized 2U/4 Node server platform.

Sold via a single SKU by qualified EVO:RAIL partners (3)

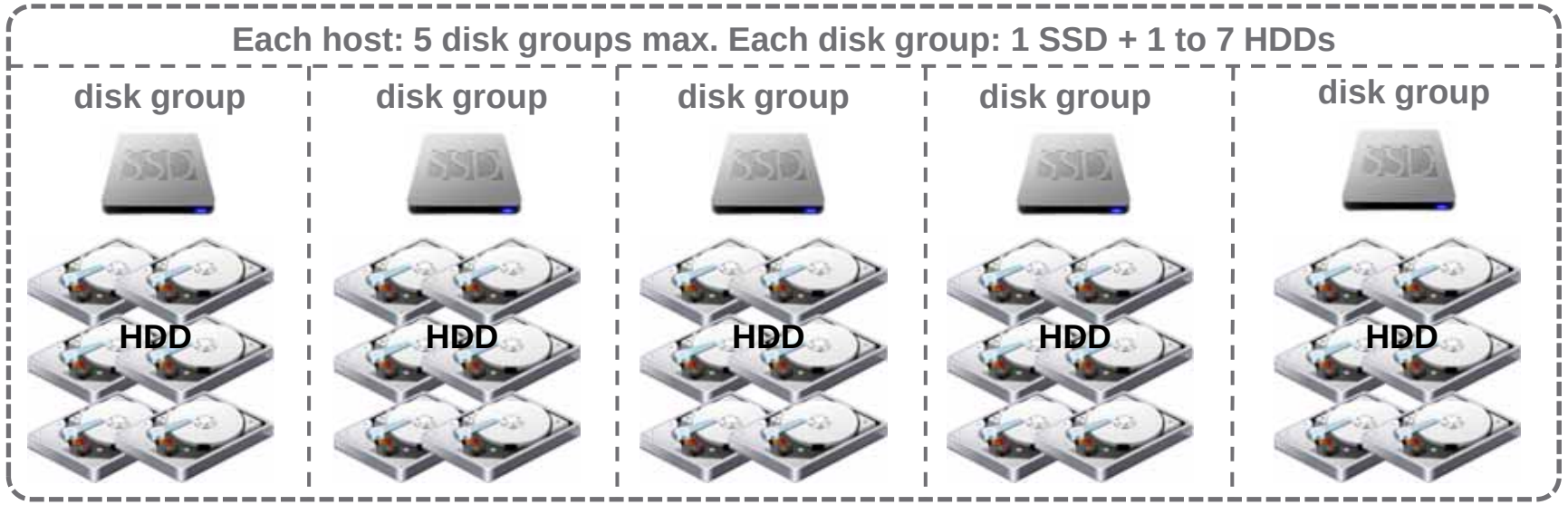
Maximum Flexibility ⟷ ***Maximum Ease of Use***

Note: 1) Components must be chosen from Virtual SAN HCL, using any other components is unsupported – see Virtual SAN VMware Compatibility Guide Page
2) VMware continues to update/add list of the available Ready Nodes, please refer to Virtual SAN VMware Compatibility Guide Page for latest list
3) EVO:RAIL availability in 2H 2014. Exact dates will vary depending on the specific EVO:RAIL partner

VSAN Hardware

Virtual SAN Disk Groups

- Virtual SAN organizes storage devices in disk groups
- A host may have up to 5 disk groups
- A disk group is composed of 1 flash device and 1-7 magnetic disks
- Compelling cost model:
  - **HDD – Cheap capacity**: persist data, redundancy for resiliency
  - **Flash – Cheap IOPS**: read caching and write buffering



Flash Devices

All writes and the **vast majority** of **reads** are served by flash storage

1. **Write-back Buffer (30%)**
   - Writes acknowledged as soon as they are persisted on flash (on all replicas)
2. **Read Cache (70%)**
   - Active data set always in flash, hot data replace cold data
   - Cache miss – read data from HDD and put in cache

A performance tier tuned for virtualized workloads

- High IOPS, low $/IOPS
- Low, predictable latency

Achieved with modest capacity: **~10% of HDD**

.

Magnetic Disks (HDD)

Capacity tier: low $/GB, work best for sequential access

- Asynchronously retire data from Write Buffer in flash
- Occasionally read data to populate Read Cache in flash

Number and type of spindles still matter for performance when…

- Very large data set does not fit in flash Read Cache
- High sustained write workload needs to be destaged from flash to HDD

SAS/**NL-SAS**/SATA HDDs supported

- Different configurations per capacity vs. performance requirements

Storage Controllers

SAS/SATA Storage Controllers

- Pass-through or “RAID0” mode supported

Performance using RAID0 mode is controller dependent

- Check with your vendor for SSD performance behind a RAID-controller
- Management headaches for “volume” creation

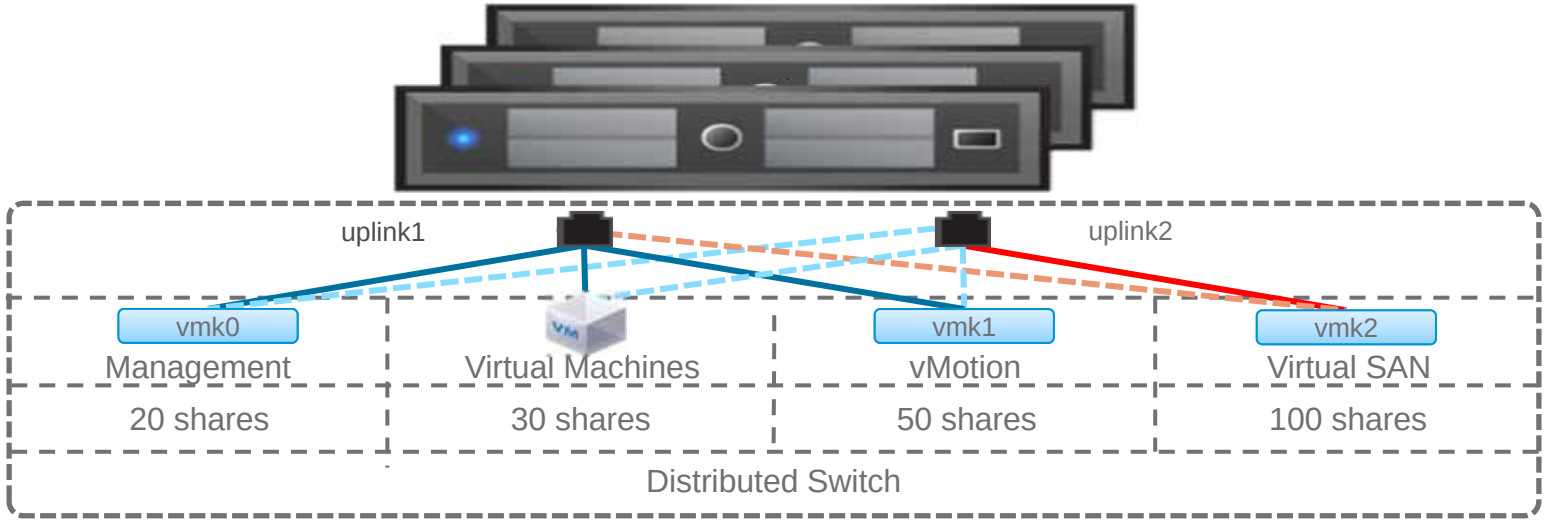
Storage Controller **Queue Depth** matters

- Higher storage controller queue depth will increase performance

Validate number of drives supported for each controller

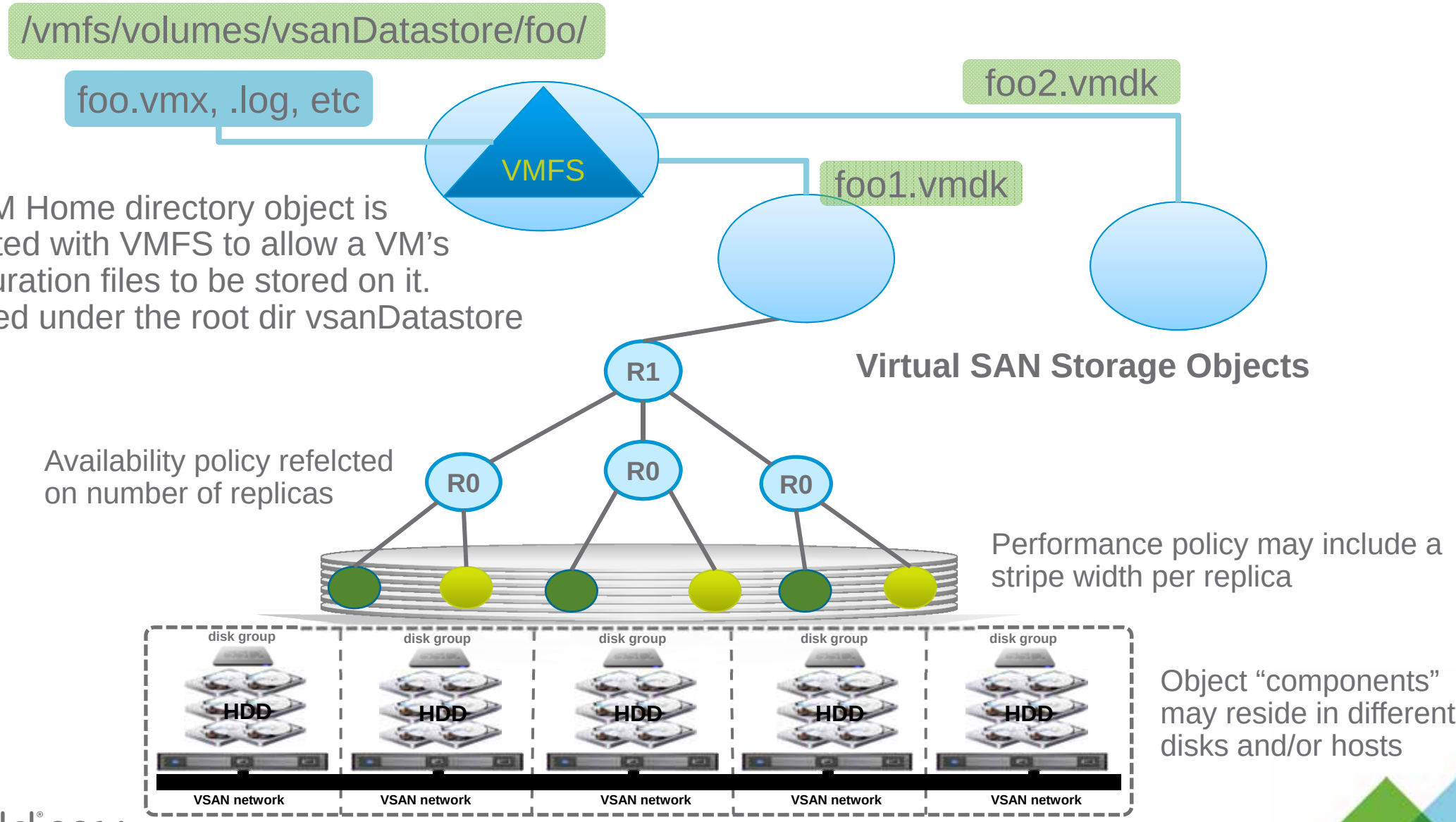
Virtual SAN Network

- New Virtual SAN traffic VMkernel interface.
  - Dedicated for Virtual SAN intra-cluster communication and data replication.
- Supports both Standard and Distributes vSwitches
  - Leverage NIOC for QoS in shared scenarios
- NIC teaming – used for availability and not for bandwidth aggregation.
- **Layer 2** Multicast must be enabled on physical switches.
  - Much easier to manage and implement than Layer 3 Multicast

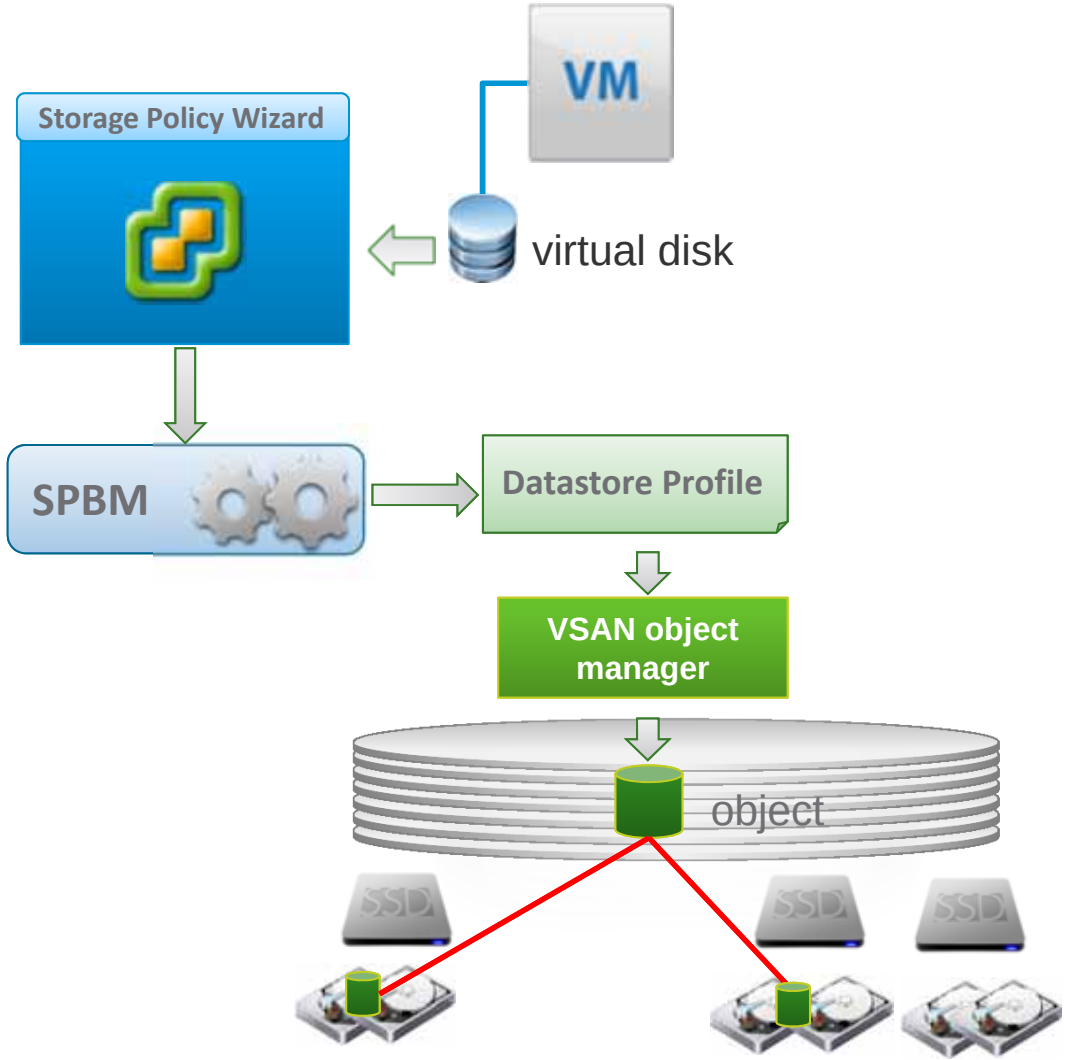


Data storage

Object and Components Layout



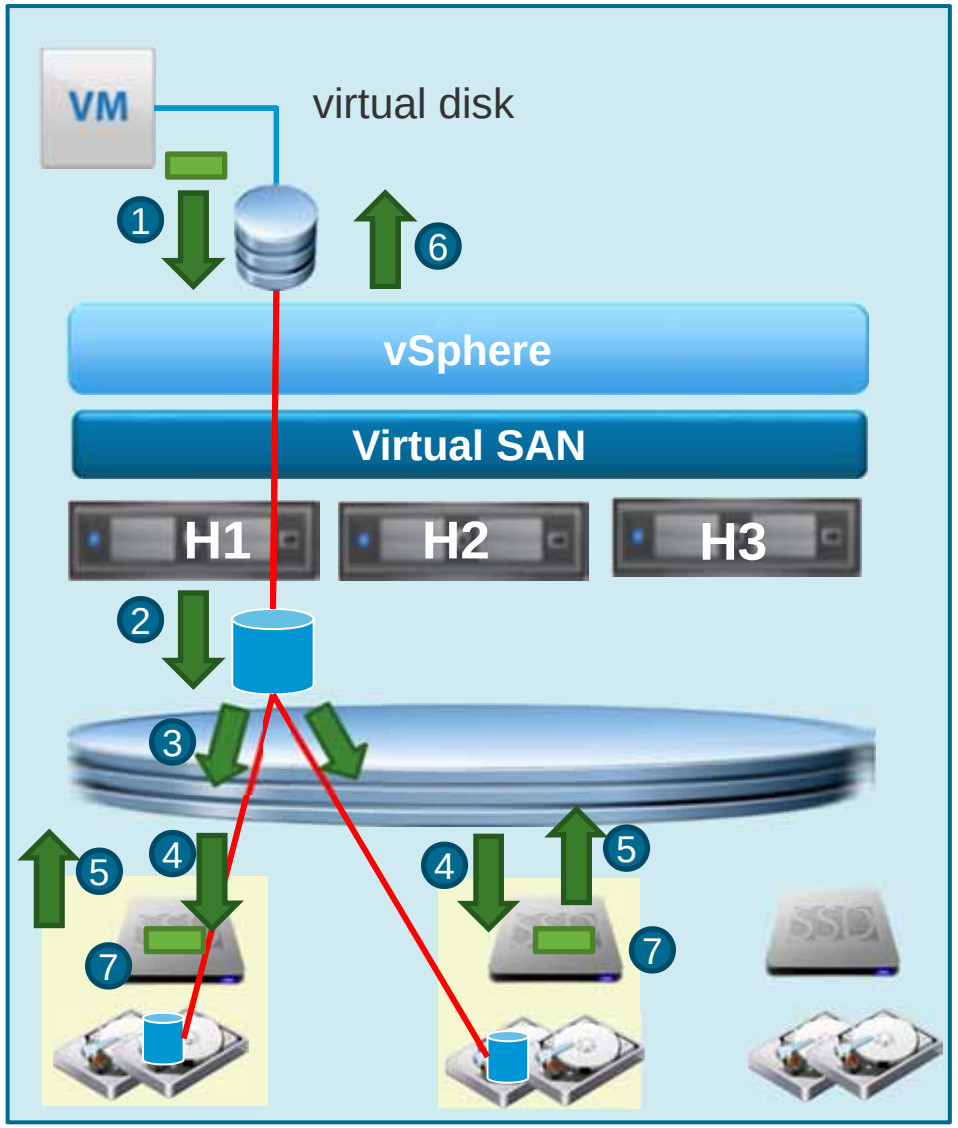
Advantages of objects



- A storage platform **designed for SPBM**
  - Per VM, per VMDK level of service
  - Application gets exactly what it needs
- Higher **availability**
  - Per object quorum
- Better **scalability**
  - Per VM locking, no issues as #VMs grows
  - No global namespace transactions

Deep breath…

Anatomy of a Write



VM running on host **H1**

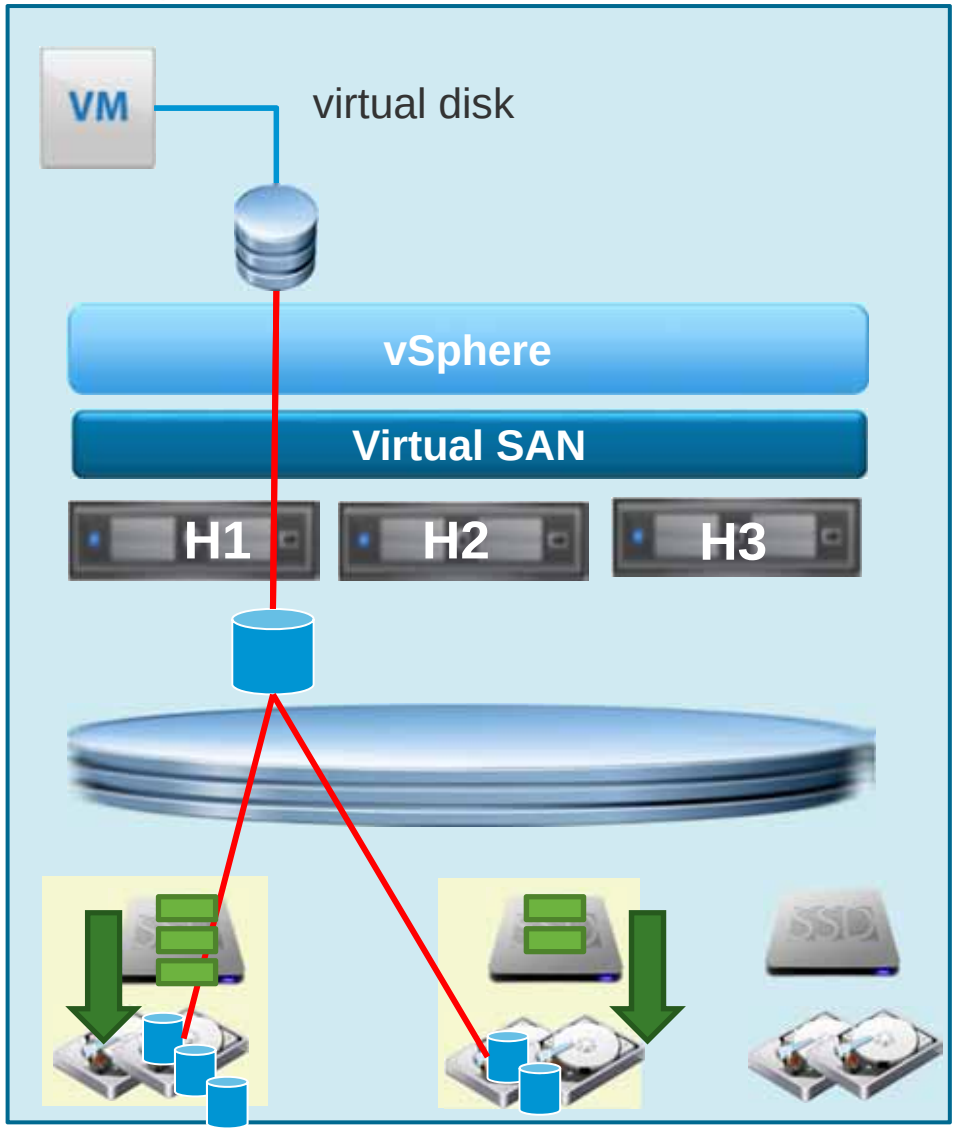
H1 is **owner** of virtual disk object

Number Of Failures To Tolerate = 1

Object has **2 replicas** on H1 and H2

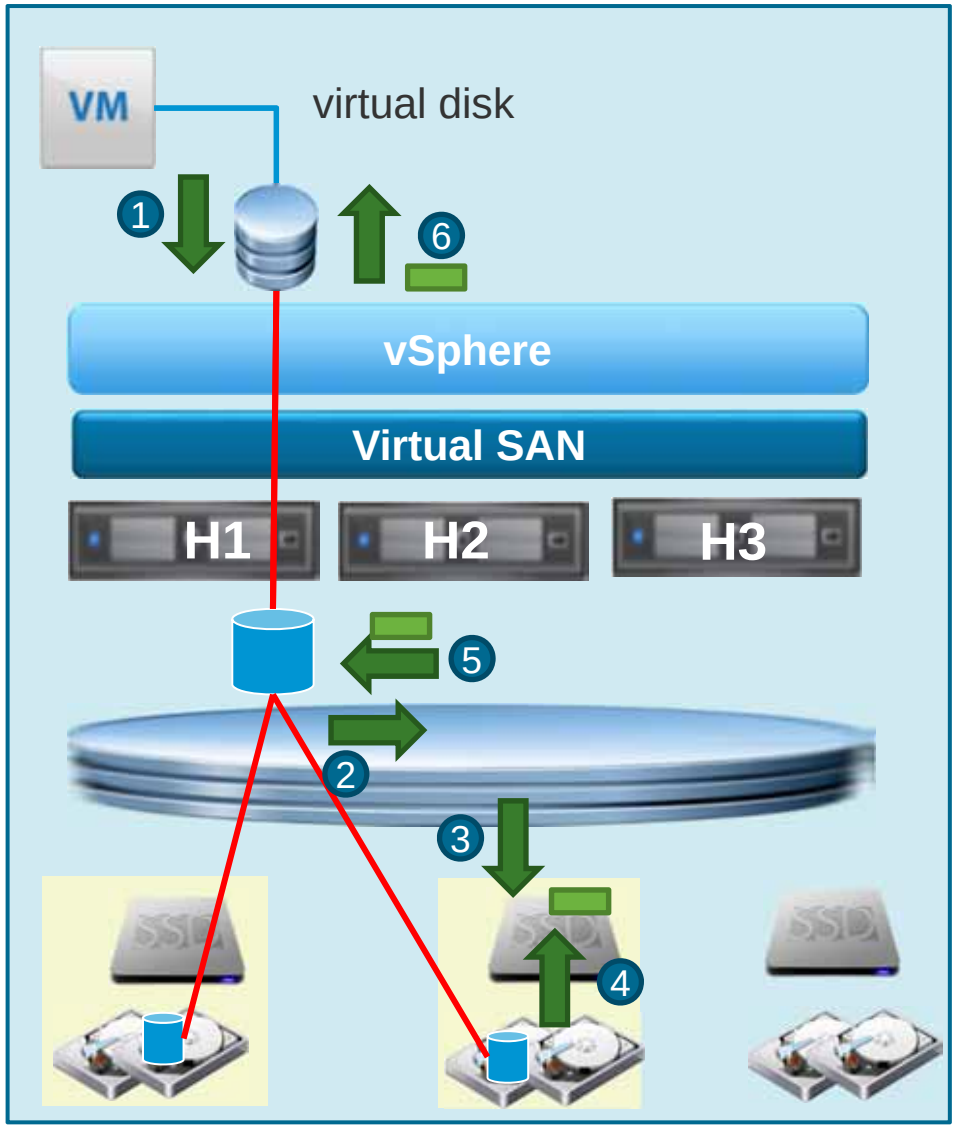
1. Guest OS **issues** write op to virtual disk
2. Owner **clones** write op
3. **In parallel**: sends "**prepare**" op to H1 (locally) and H2
4. H1, H2 **persist** op to Flash (log)
5. H1, H2 **ACK** prepare op to owner
6. Owner waits for ACK from both 'prepares' and **completes** I/O
7. Later, owner **commits** batch of writes

Destaging Writes from Flash to HDD



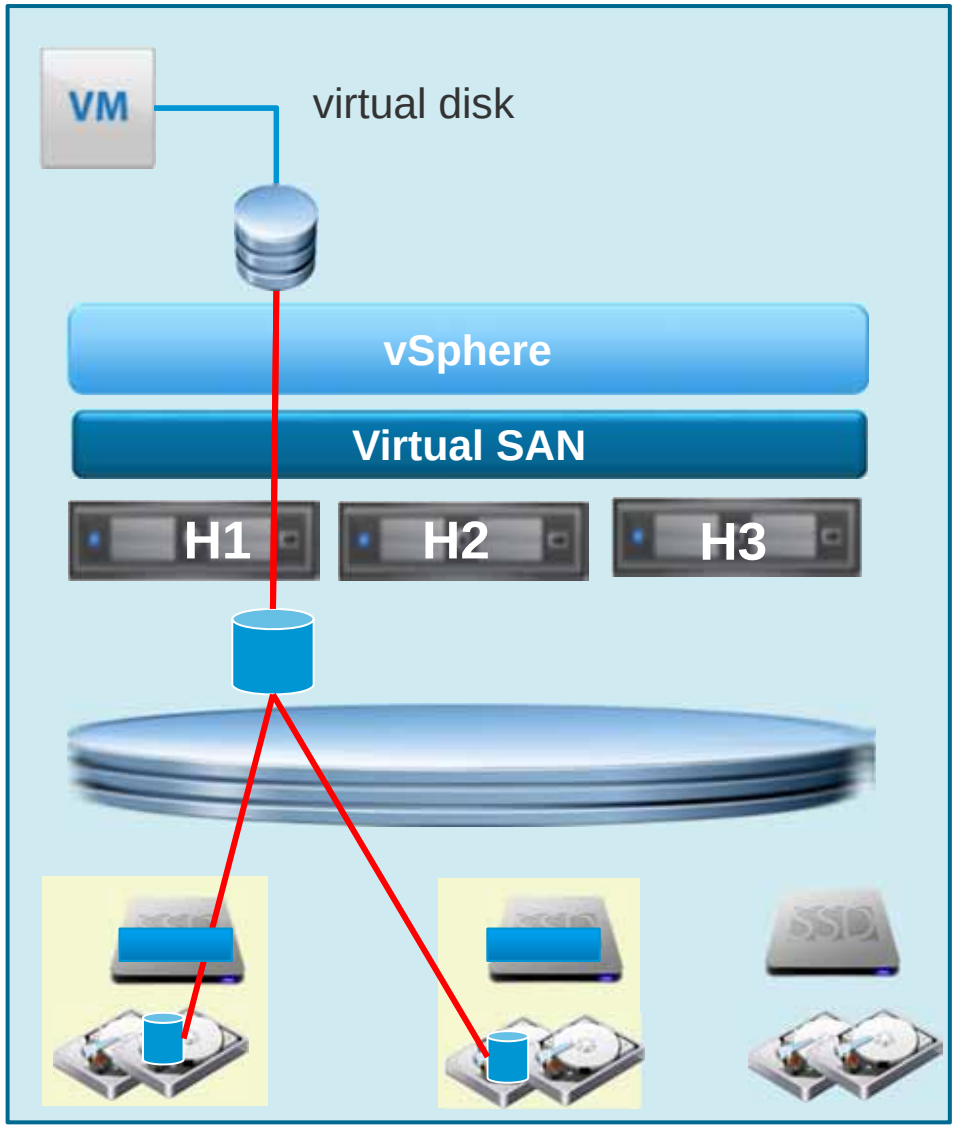
- **Data from committed writes** accumulate on Flash (Write Buffer)
  - From different VMs / virtual disks
- **Elevator algorithm** flushes written data to HDD asynchronously
  - Physically **proximal** batches of data per HDD for improved performance
  - **Conservative**: overwrites are good; conserve HDD I/O
  - **HDD write buffers** are flushed, before discarding writes from SSD

Anatomy of a Read



1. Guest OS **issues** a read on virtual disk
2. **Owner chooses** replica to read from
   - **Load balance** across replicas
   - Not necessarily local replica (if one)
   - A block always read from same replica; data cached on at most 1 SSD; **maximize effectiveness**
3. At chosen replica (H2): read data from SSD **Read Cache,** if there
4. Otherwise, read from HDD and place data in SSD Read Cache
   - Replace ‘cold’ data
5. Return data to owner
6. Complete read and return data to VM

Virtual SAN Caching Algorithms



- VSAN exploits temporal and spatial locality for caching
- **Persistent cache** by the replica (Flash)
  - **Not** by the client! Why?

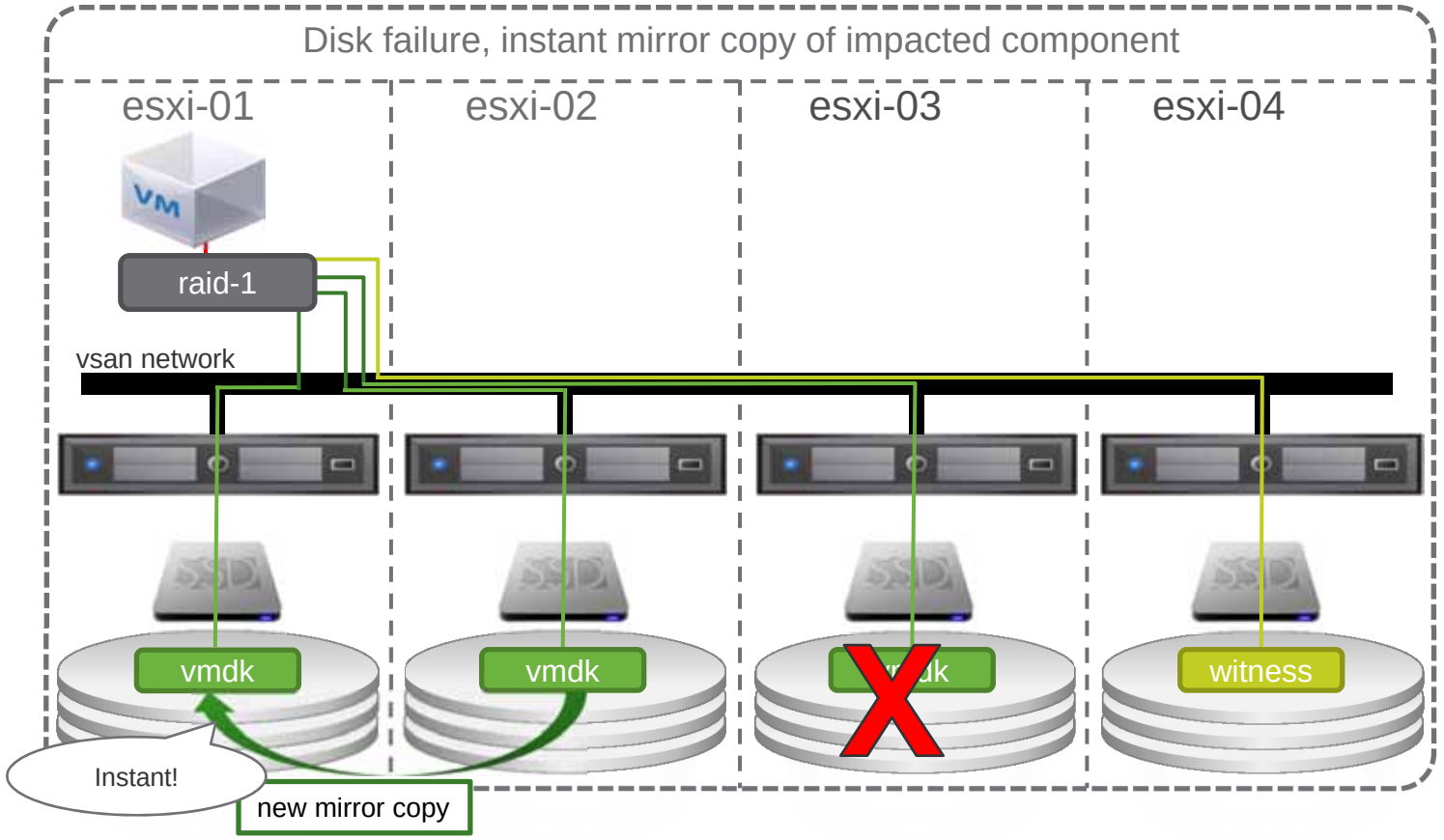
- Improved **flash utilization** in cluster
- **Avoid data migration** with VM migration
  - DRS: 10s of migrations per day
- **No latency penalty**
  - Network latencies: 5 – 50 usec (10GbE)
  - Flash latencies with real load: ~1 msec

- VSAN supports **in-memory local cache**
  - Memory: very low latecy
  - View Accelerator (**CBRC**)

Fault tolerance

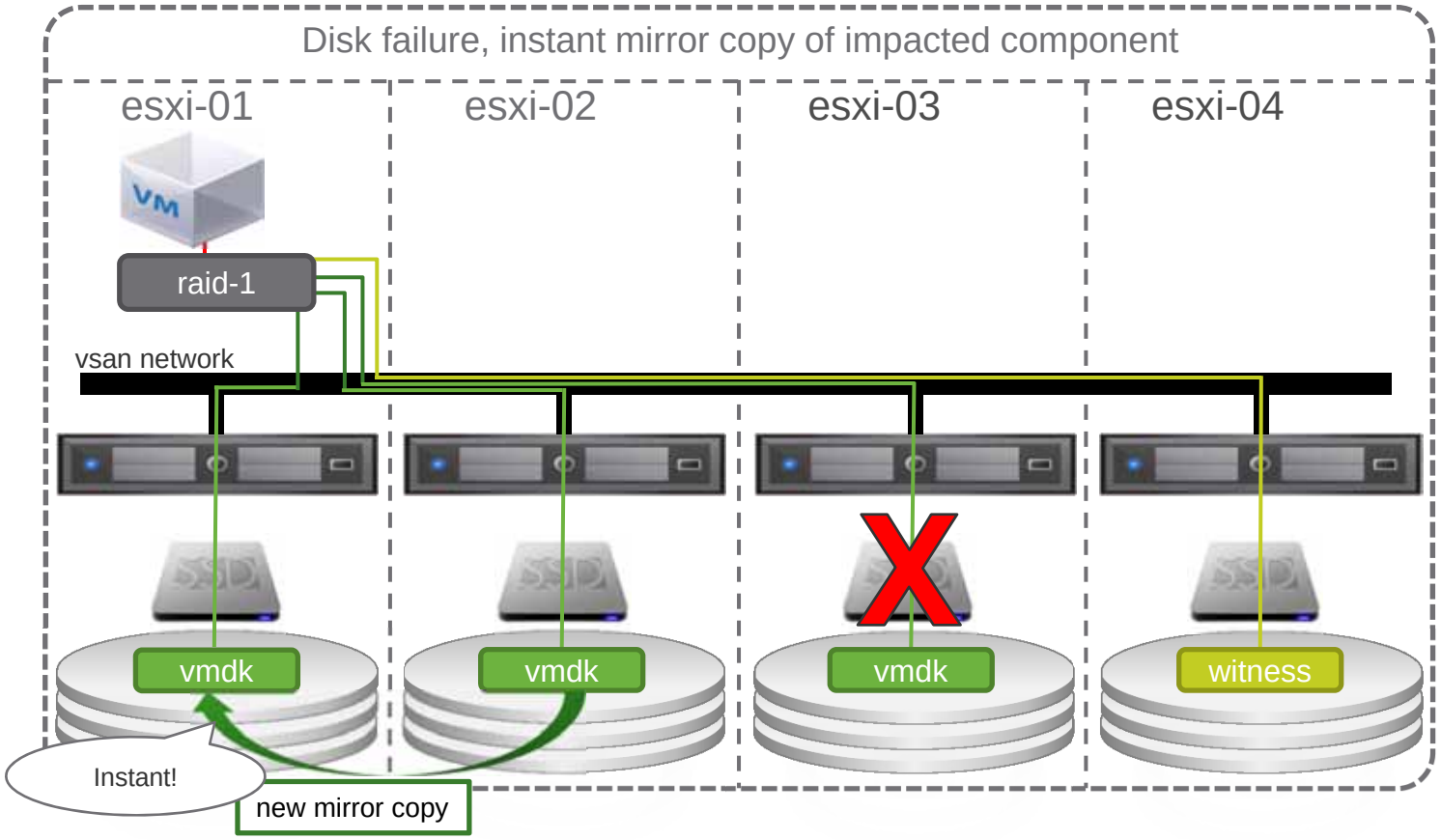
Magnetic Disk Failure: Instant mirror copy

- **Degraded** - All impacted components on the failed HDD instantaneously re-created on other disks, disk groups, or hosts.



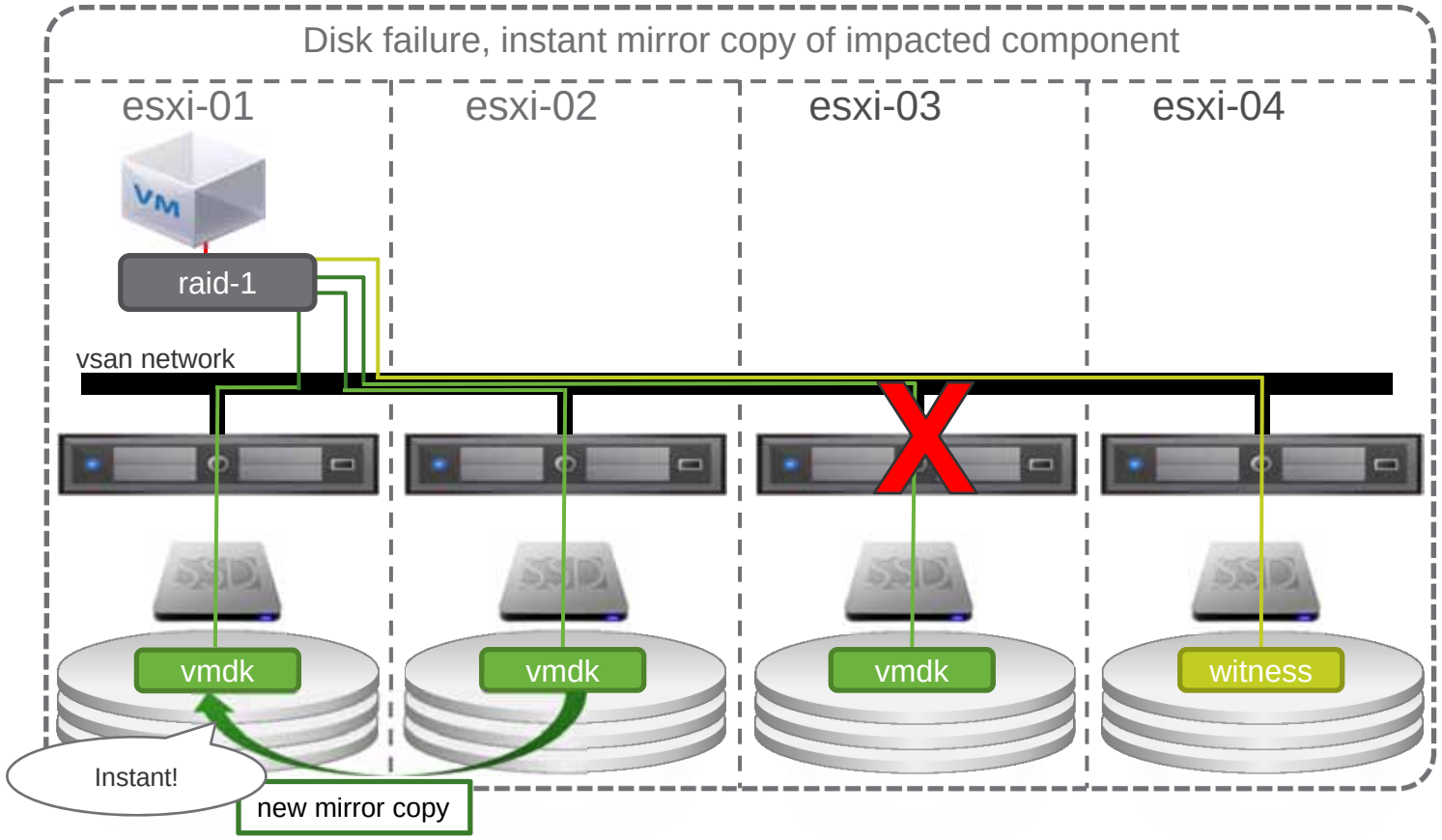
Flash Device Failure: Instant mirror copy

- **Degraded** – Entire disk group failure. Higher reconstruction impact. All impacted components on the disk group instantaneously re-created on other disks, disk groups, or hosts.

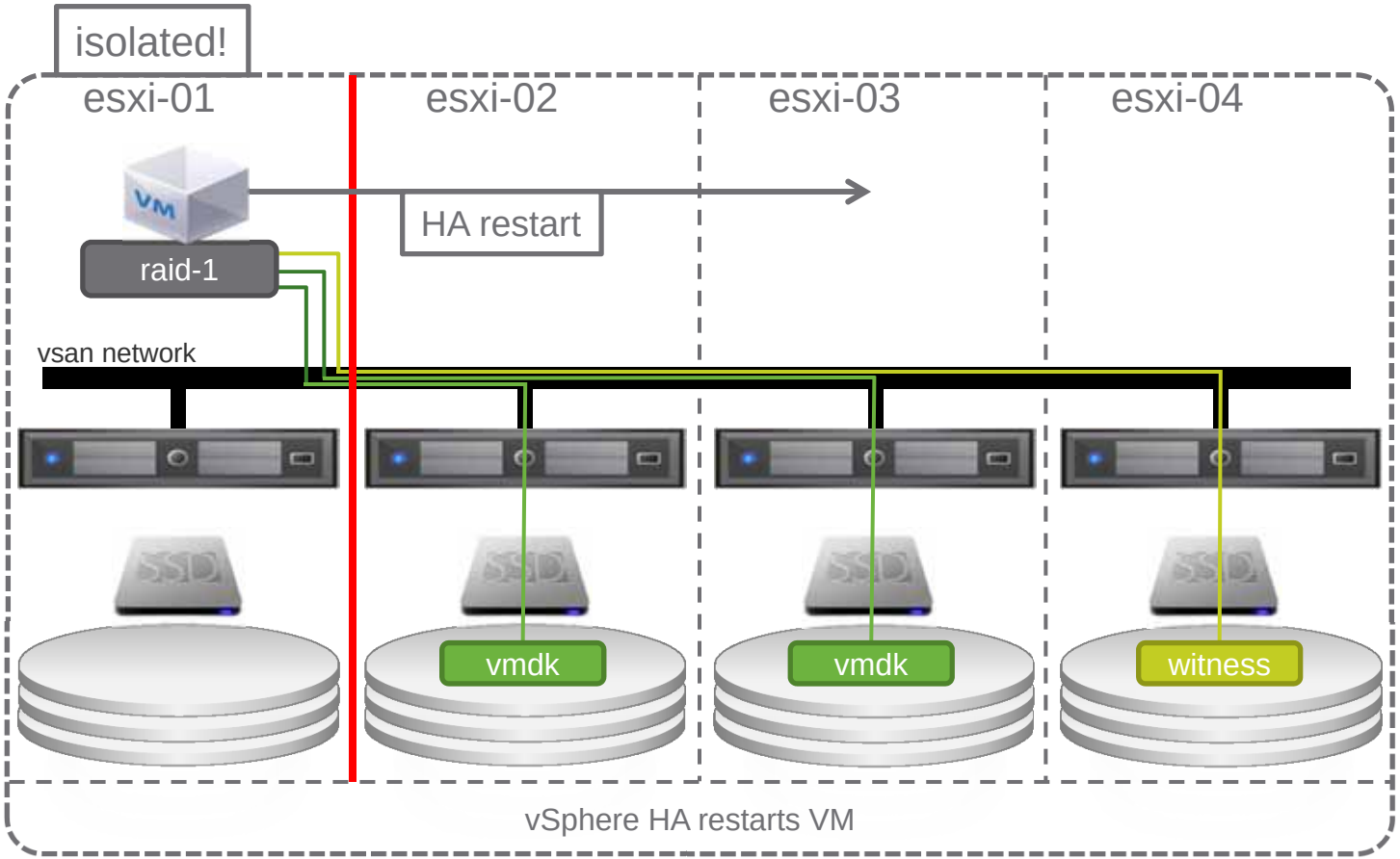


Host Failure: 60 Minute Delay

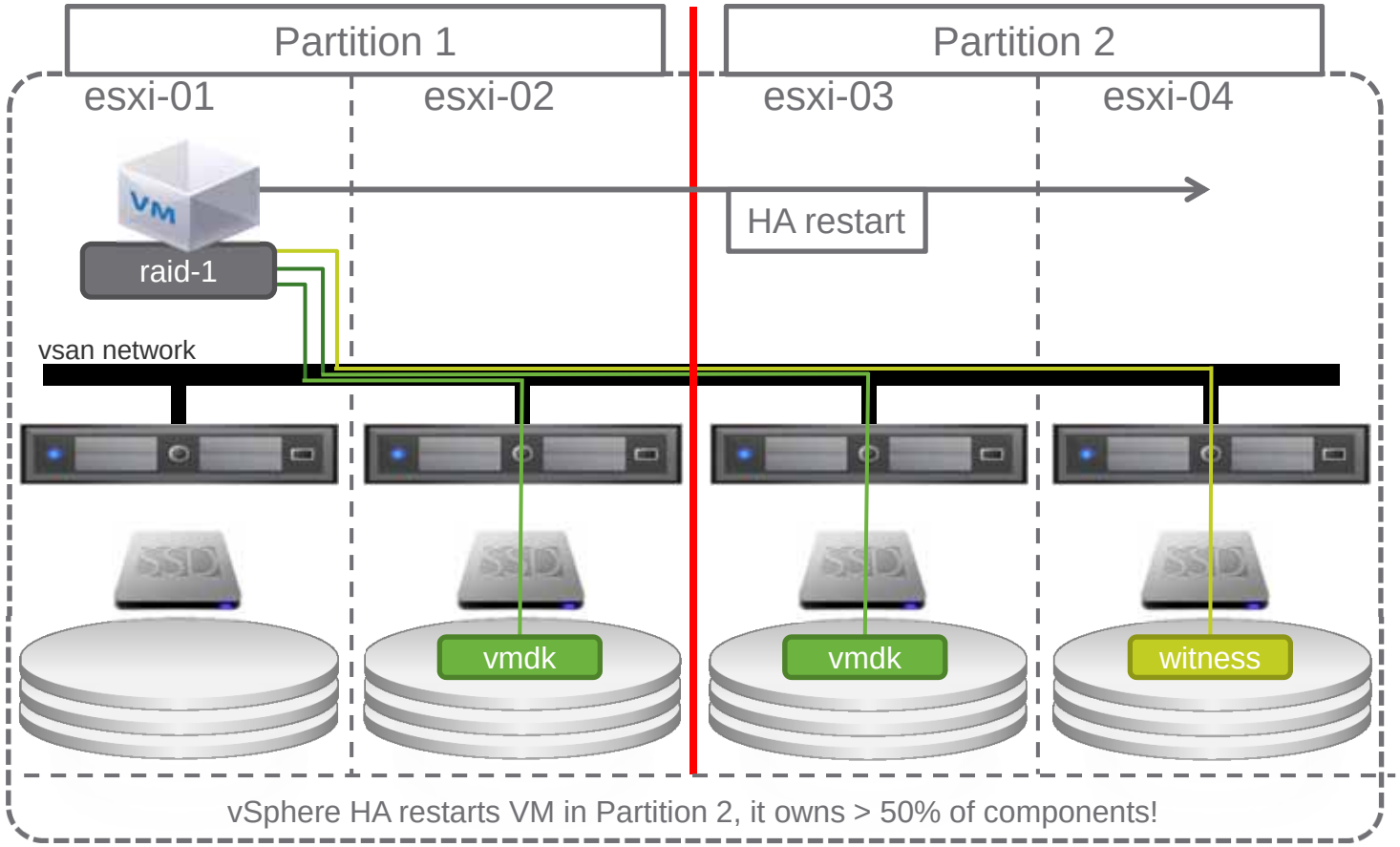
- **Absent** – Host failed or disconnected. Highest reconstruction impact. Wait to ensure not transient failure. Default delay of 60 min. After that, start reconstructing objects and components onto other disk, disk groups, or hosts.



Virtual SAN 1 host isolated – HA restart



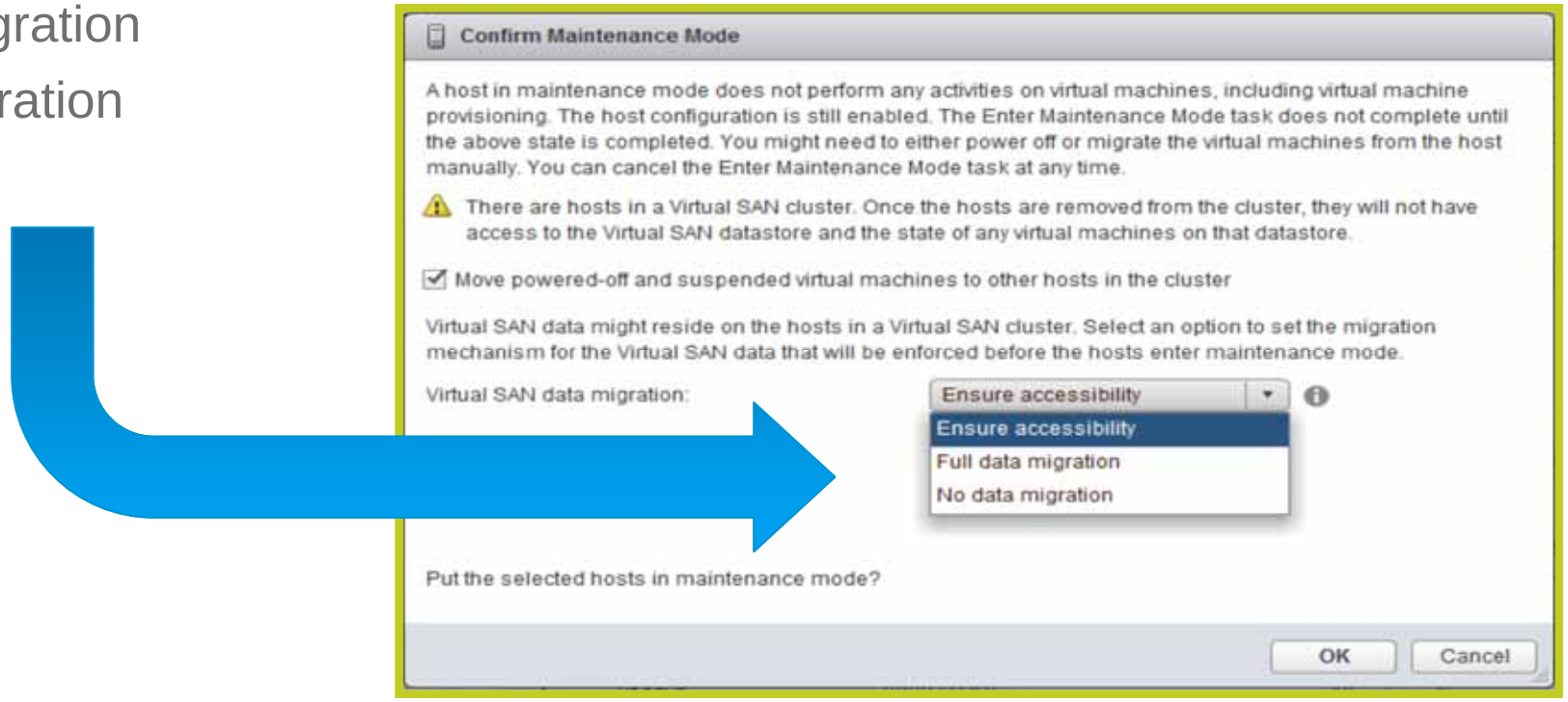
Virtual SAN partition – With HA restart



Maintenance Mode – planned downtime

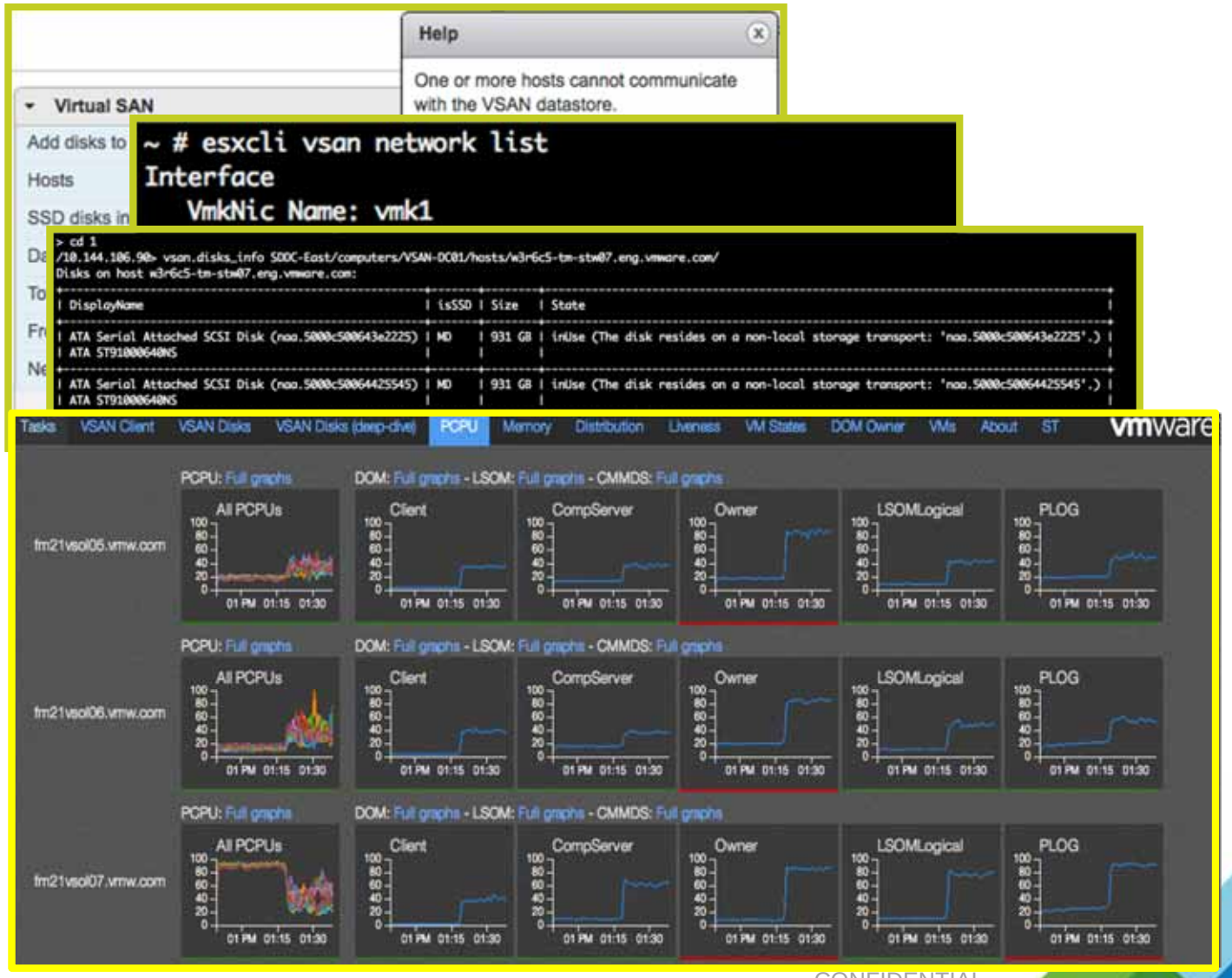
3 Maintenance mode options:

- Ensure accessibility
- Full data migration
- No data migration



Virtual SAN Monitoring and Troubleshooting

- vSphere UI
- Command line tools
- Ruby vSphere Console
- VSAN Observer



Virtual SAN Key Benefits

Radically Simple

- Enabled/configured in two clicks
- Policy-based management
- Self-tuning and elastic
- Deep integration with VMware stack
- VM-centric tools for monitoring & troubleshooting

High Performance

- Flash acceleration
- Up to 2M IOPS from 32 nodes
- Low, predictable latencies
- Minimal CPU, RAM consumption
- Matches the VDI density of all flash array

Lower TCO

- Eliminates large upfront investments (CAPEX)
- Grow-as-you-go (OPEX)
- Flexible choice of industry standard hardware
- Does not require specialized skills


Thank You
vmworld 2014


Fill out a survey
Every completed survey is entered into a drawing for a $25 VMware company store gift certificate
vmworld 2014

STO1279

Virtual SAN Architecture Deep Dive

Christos Karamanolis, VMware, Inc
Christian Dickmann, VMware, Inc