Storage

Maxs

Maximums (per host): Virt disks = 2048   LUNs/Volumes = 256   Paths = 1024   NAS mounts = 256
FC   HBAs = 8 (ports = 16)   Targets per HBA = 256 Paths to each LUN = 32   LUN size = 64TB
FCoE   Adapters = 4
iSCSI HW   HBAs = 4   Targets per HBA = 62-128 (depends on card)   Paths to each LUN = 8
iSCSI SW   NICs = 8   Targets = 256   Paths to each LUN = 8
Maximums (per volume): Powered-on VMs = 2048 Hosts = 64 (DRS limit, 2048 for linked clones)
VMFS5 = 64TB (less 16KB)   vRDMs = 2TB (less 512B)   pRDMs = 64TB (less 512B)
File size = 2TB (less 512B)   NFS =16TB   Extents = 32   Extent size = 2TB (less 512B)
Maximums (per datastore cluster): Virt disks = 9000   Datastores = 32   (datastore clusters per vCenter = 256)
Storage vMotions: 8 per datastore   2 per host   128 vMotions per datastore

Firewall Ports: iSCSI – 3260 TCP, NFS – 111 TCP/UDP, 2049 TCP/UDP

Shell Commands

`--help` for esxcli namespaces & commands relative to location. `localcli` bypasses hostd

Rescan SCSI HBAs (devices, paths, claimrules, FS):	`esxcli storage core adapter rescan`
List all SCSI paths:	`esxcli storage core path list`
Map VMFS volumes to devices/partitions:	`esxcli storage filesystem list`
List unresolved snapshot/replicas of volumes:	`esxcli storage vmfs snapshot list`
SATP claiming rules:	`esxcli storage nmp satp rule list`
List nmp devices with SATP & PSP:	`esxcli storage nmp device list`
List all claim rules:	`esxcli storage core claimrule list`
List storage devices with properties/filters:	`esxcli storage core device list`
Lists HBA drivers & information:	`esxcli storage core adapter list`
Show each device’s VAAI support:	`esxcli storage coredevice vaai status get`
List FCoE HBA adapters:	`esxcli fcoe adapter list`
List FCoE CNAs:	`esxcli fcoe nic list`
List iSCSI adapters:	`esxcli iscsi adapter list`
Show current iSCSI session:	`esxcli iscsi session list`
Discover iSCSI devices:	`esxcli iscsi adapter discovery rediscover`
Check if software iSCSI is enabled:	`esxcli iscsi software get`
List the NFS filesystems & mounts:	`esxcli storage nfs list`
Test VMkernel connectivity:	`vmkping [-s 9000] <ipaddress>`
SCSI performance statistic tool:	`vscsiStats`
Create/Delete/Modify VMDKs, RDMs, VMFS volumes & storage devices:	`vmkfstools`

	Access	Boot VM	Datastore	RDM	VM MSCS	HA, DRS, vMotion, FT, SRM
Local	Block (LUNs)	Yes	VMFS	No	No	No
FC	Block (LUNs)	Yes	VMFS	Yes	Yes	Yes
iSCSI	Block (LUNs)	Yes	VMFS	Yes	No	Yes
NFS	File	Yes	NFS	No	No	Yes

Array types: Active-Active - IO to all LUNs simultaneously through all ports/SPs, without performance degradation. Active-Passive - one port actively provides access, others are backup for that LUN (but can be active for other LUNs). Path thrashing can occur. ALUA (Asymmetric Logical Unit Access) - on non Active-Active arrays, paths are not available or not optimized to every port on every SAN SP. ALUA support on a SAN helps the hosts find/manage the best paths for failover & load balancing.
Virtual Port (iSCSI only) - SANs are Active-Active but mask all connections, ESXi accesses all LUNs via single virtual port. SANs handle failover & load balancing.
Multipathing: path failover (redundancy) & load balancing.
Claim rules: specifies which MPP (MultiPathing Plugin), native or 3rd party, manages physical path. Rules based on SAN discovered, listed in /etc/vmware/esx.conf. Path evaluation every 5 mins.
NMP (Native MPP): Secondary rules applied for: • SATPs (Storage Array Type Plugins) - handles failovers for array type. Rules search order: drivers, then vendor/model, lastly transport, *VMW_SATP_DEFAULT_AA* default if array not recognized. • PSPs (Path Selection Plugins) - handles load-balancing for each device.
PSPs policies: *VMW_PSP_FIXED* - default for active/active, uses preferred path (marked with *) when available, default policy if device not recognized. *VMW_PSP_MRU* (Most Recently Used) - default for active/passive (& iSCSI), if SATP is VMW_SATP_ALUA then active/optimized path used, if not ALUA then uses first working path found at boot. *VMW_PSP_RR* (Round Robin) - safe for all arrays, rotates through paths (not MSCS LUNs).
Resignaturing: VMFS datastores have unique UUID in file system metadata. Replicated or snapshotted disks keep same UUID. Can mount with existing signature or assign new signature. Resignaturing assigns new UUID & label, then mounts. If not resignaturing, original must be offline.
Rescans: datastore rescans are automatic for many operations. Manual rescans may be required: zoning changes, new SAN LUNs, path masking changes on host, cable reconnected, changed CHAP settings, adding/removing iSCSI addresses, re-adding hosts. Rescans LUN 0 to LUN 255. *Disk.MaxLUN* reduces number of LUNs scanned to increase boot times & rescans. If LUN IDs are sequential disable sparse LUN support.
Zoning: at the switch. **LUN masking**: mask certain LUN IDs at Array’s SP or ESXi host using claim rules.
PDL (Permanent Device Loss): if LUN is being removed, detach it so volume is unmounted. Host may detect SCSI sense codes to determine LUN is offline & it is not a APD.
APD (All Paths Down): No active paths to storage device. Unexpected so host continually retries paths.
LUN queue depth: SCSI device driver parameter that limits number of commands a LUN can accept. Excess commands are queued in VMkernel. Increase queue depth if VMs’ commands consistently exceeds queue depth. Procedure depends on host adapter. Setting higher than default can decrease number of LUNs supported. Change *Disk.SchedNumReqOutstanding* to match - it limits requests each VM can issue.
LUN Device IDs: • SCSI inquiry – returned by device , unique across hosts, persistent, T10 standard e.g. naa.#, t10.# or eui.# • Path-based – not unique, not persistent, e.g. mpx.path • Legacy – created in addition to SCSI inquiry or Path-based, e.g. vml.# • Runtime name – host specific, not persistent, first path to device, adapter:channel:target:LUN, e.g. vmhba#:C#:T#:L#
FCoE – CNA (Converged Network Adapter) or NIC with partial offload & SW initiator. Disable STP (Spanning Tree Protocol) - might delay FIP (FCoE Initialization Protocol). Enable PFC (Priority-based Flow Control) & set to AUTO.
iSCSI: Interfaces: • iSCSI HBA (independent HW) • NIC with iBFT/iSCSI offload (dependent HW) & SW initiator • Regular NIC & SW initiator. Only 1 SW initiator per host.
Independent HW configured in Storage configuration. Non-independent configuration - 1 VMkernel interface to 1 active NIC, others unused, bind adapters. Set “MAC Address Changes” PG policy to *Accept*. SW initiator enabled by default.
Boot from iSCSI SAN: only Independent HW LUN installs get diagnostic partition.
iSCSI Nodes: • IP address • iSCSI name (IQN e.g. iqn.yyyy-mm.reversed_domain_name:string or EUI e.g. eui.string) • iSCSI alias – not unique, friendly name.
iSCSI Discovery methods: • Dynamic - uses *SendTargets*, target responds with list. Targets added to Static list, removed targets can return after HBA rescan/reset or host reboot • Static - can manually add/remove items, only with hardware initiators
iSCSI SAN access control: • Initiator name • IP addresses • CHAP
CHAP authentication: • 1-way (unidirectional) target authenticates initiator (set on SAN) • Mutual (bidirectional) target & initiator can authenticate each other – only SW iSCSI or dependent HW cards.
CHAP character limits: • only alphanumeric (no symbols) • CHAP name ≤ 511, CHAP secret ≤ 255
NFS: Not supported: Access via non-root credentials (delegate user).
SIOC (Storage IO Control): shares VM's disk IO across datastore's hosts. Monitors latency, adjusts VM's host queue access. Can also enforce VM IOPS limits. Enable on datastore, set shares/limit on VM. Limit must be set on all VM’s disks. Datastores must be managed by single vCenter. Not supported – RDMs, datastores with multiple extents. Auto-tiering arrays must be certified for SIOC. Enabled by default on Storage DRS enabled datastore clusters. Congestion Threshold is upper limit for datastore before SIOC allocates based on shares. Low threshold = lower device latency & strong VM IO isolation. High threshold = high throughput & weak isolation. Default 30ms, max 100ms, min 10ms.
Datastore Cluster: Manage multiple datastores as one logical resource. Can be VMFS or NFS, but not both in same cluster. Can’t combine SRM replicated & non-replicated datastores in same cluster. Recommendations: datastore of similar size and IO capabilities in cluster, don’t mix HW accelerated & non accelerated backed datastores.
Datastore Maintenance Mode: automatically evacuates all VMs from a datastore. Affinity rules can prevent entering maintenance mode, to override use setting *IgnoreAffinityRulesForMaintenance* = 1.
Storage DRS: Load balancing & Initial placement of VMDKs to balance I/O & spread free space. By default, IO evaluated every 8 hrs, IO latency threshold is 15ms. Datastores used by multiple datacenters ignored by IO balancing (space balancing still used). All datastore connected hosts must be ESXi 5. Mandatory recommendations when: • Datastore out of space • Affinity/Anti-Affinity rules violated • Datastore enters maintenance mode
DRS settings preserved if feature is disabled. Storage DRS & SRM not aware of each other.
Storage DRS Automation Levels: • Manual (default) • Fully Automated • Disabled.
Aggressiveness thresholds: • Space Utilization • IP Latency. Advanced options: • Space Utilization difference – ensures minimum difference between source & destination • IO Load Balancing Invocation Interval • IO Imbalance Threshold. Create scheduled task to change automation or aggressiveness level so migrations more likely off-peak.
Storage DRS affinity rules: • VMDK Anti-Affinity (Intra-VM) - VM with multiple disks are split across datastores • VMDK Affinity - VM’s disks kept together • VM Anti-Affinity (Inter-VM) - disks from specified VMs (only 2) are kept on apart. By default VM’s disks kept together. Anti-affinity not enforced for user initiated migrations & not applicable to ISOs or swap files.
Storage Profiles: Set the requirements of the VM home files & disks.
HW acceleration: (enabled by default) offloads operations to supported arrays. Faster, consumes less host resources, reduces fabric bandwidth usage.
Block: *Full copy* (or clone blocks or copy offload) - array copies data without host read/write. *Block zeroing* (or write same) – array zeros out blocks of newly allocated storage. *Hardware assisted locking* (or Atomic Test & Set (ATS)) - discrete VM locking avoiding SCSI reservations, allows disk locking per sector, instead of entire LUN.
NFS: *File clone* - similar to VMFS File Copy except entire files are cloned instead of file segments. *Reserve space* - arrays can allocate space for thick format disks. *Extended file statistics* - accurate reporting of space usage of VMs.
Storage Capability: system or user defined. Outine of datastore’s capacity, performance, availability, redundancy, etc.
Storage vMotion: supports vSphere snapshots and linked clones. Uses Mirror Mode driver.
SSD: Can be used for host’s VM swapping, very high IO for increased consolidation, guest OS can recognize it as SSD. Use SATP claim rule to tag SSD if not recognized automatically.
NPIV (N-Port ID Virt): FC HBA port assigns dedicated VPORT to VM as WWN pair - see VM section
Links: Resolution Path – Troubleshooting Storage http://communities.vmware.com/docs/DOC-16708
NetApp vSphere Storage best practices whitepaper http://media.netapp.com/documents/tr-3749.pdf
File System alignment whitepaper (NetApp) http://media.netapp.com/documents/tr-3747.pdf
vSphere handling of LUNs detected as snapshot http://kb.vmware.com/kb/1011387