Revision Notes for VMware Certified Professional on Virtual Infrastructure 3

Mark Wilson, December 2006

These notes were made whilst revising for my VCP3 examination – they are nothing more than a collection of notes and do not represent any insight gained from taking the test itself (I find that writing things out is a good way to prepare for an exam). Feel free to use them as an aide memoire, or even as a reference when working with VMware Infrastructure 3, but take note of the following:

Acronyms/abbreviations

Many acronyms used within this document are standard IT terms (e.g. SCSI, SAN) – in general, any new terms will be expanded at first use.

Some abbreviations used throughout the document include:

- ESX: VMware ESX Server (also ESX2 or ESX3 to indicate major version)
- VC: VMware VirtualCenter (also VC1 or VC2 to indicate major version)
- HA: VMware High Availability
- DRS: VMware Distributed Resource Scheduling
- VI: Virtual Infrastructure (e.g. VI3)
- VM: Virtual machine (also VM2 or VM3 to indicate version)
- VMtools: VMware Tools (e.g. VMtools3)

In addition, the letters l, p and v may be prefixed to common terms to indicate logical, physical or virtual instances, e.g.:

- lCPU: logical CPU – either HyperThreading or dual-core.
- pCPU: physical CPU; pNIC: physical NIC; pswitch: physical switch.
- vCPU: virtual CPU; vNIC: virtual NIC; vswitch: virtual switch; vdisk: virtual disk.

These letters are also used to indicate migration states, e.g. P2V (physical to virtual) and V2V (virtual to virtual).

Online resources

VMware Technical Support – http://www.vmware.com/support/.

ESX Server documentation online – http://www.vmware.com/support/pubs/esx_pubs.html.

VirtualCenter documentation online – http://www.vmware.com/support/pubs/vc_pubs.html.

Virtual Infrastructure documentation online – http://www.vmware.com/support/pubs/vi_pubs.html.

VMware knowledge base – http://www.vmware.com/kb/.

VMware technology network (VMTN) – http://www.vmware.com/vmtn/.

VMware community forums – http://www.vmware.com/community/.

VCP portal http://mylearn1.vmware.com/portals/certification/.

VI3 exam – example questions – http://mylearn1.vmware.com/portals/certification/questions.cfm.

VI3 exam blueprint – http://mylearn1.vmware.com/lcms/mL_faq/528/VCP%20on%20VI3%20Blueprint.pdf.

TestKing practice exam – http://www.testking.com/certification-training-VMware.htm (not tested).

Virtual infrastructure

"Virtual Infrastructure allows dynamic mapping of compute, storage and network resources to business applications" [source: VMware].

- Multiple VMs running on the same physical computer. Each has its own operating system, storage and network identity.
  - Virtualisation uses underlying physical hardware wherever possible (unlike emulation).
  - Not emulated or simulated – each VM has its own set of virtual hardware.
  - Not sessions (e.g. terminal server) – each VM has its own operating system.
- Benefits:
  - Isolation – if one VM crashes, others are unaffected – host multiple applications on a single physical server and increase resource utilisation.
  - Encapsulation – a VM is a small collection of files – easy to image and clone.
  - Compatibility – a standard x86 hardware set is used.
  - Hardware independence – physical hardware changes do not affect VMs – can change/move hardware more freely (some limitations around processor).
- CPU virtualisation modes:
  - Direct execution – VMM runs the VM directly on the pCPU – near native speed – generally used for user-level application code because doesn't access privileged state data.
  - Virtualisation mode – used if direct execution is not possible – adds a varying amount of virtualisation overhead – used for operating system code that modifies privileged state data – consequently applications that make a lot of system calls will run more slowly in a VM.

VMware products

- Categorised as:
  - Datacenter: VI3 (application lifecycle management and data centre operations); P2V assistant (server consolidation).
  - Developer: VMware Workstation (application lifecycle management and field operations); VMTN subscription (application lifecycle management).
  - Enterprise desktop: VMware ACE (desktop security).
  - Free virtualisation: VMware Player (run, share and evaluate pre-built VMs); VMware Server (test and development, software evaluation and server provisioning).
- ESX:
  - VM platform (hypervisor) installed on "bare metal" – Intel or AMD hardware.
  - VMkernel entirely dedicated to execution of VMs – has complete control over physical resources – fine grained resource allocation and dynamic adjustment.
  - Service console for user logons and remote access.
  - VMFS – high performance file system.
  - Scale up to 16 CPUs and 64GB RAM, host up to 128 vCPUs as 1, 2 or 4-way VMs (with VSMP) – 4-8 VMs per core.
- VC:
  - Tool to manage VI with a single view – runs on Windows.
  - Create and manage inventory; provision VMs from templates; migrate running VMs (VMotion); balance workload (DRS); provide high availability and disaster recovery (HA).
- VDI:
  - Use VI to provide a virtual desktop infrastructure.
  - Hosted inside secure data centre; accessed remotely (RDP, VNC, etc.); lightweight client; data back into data centre; suited to branch or remote office scenarios; ease of recovery/redeployment; protect confidential information and intellectual property.
  - 16:1 consolidation on a 2-way server cf. 40:1 with Citrix but Citrix does not allow desktop interaction, some applications are not suitable and it's not recommended to mix workloads.
- P2V:
  - Migrate existing physical machines to VMs – Windows NT/2000/XP/2003.
  - 3 steps: Image; virtualise; reinstantiate (boot).
    - Image using third party tools (e.g. Symantec Ghost) or with boot CD and P2V imaging.
    - Virtualise with GUI-driven P2V Assistant – runs on Windows (physical or virtual).
    - Reinstantiate on any VMware product.
  - Fast and safe – original files preserved.

  - Will be replaced by VMware Converter (currently in beta).
  - Third party products also available, e.g. PlateSpin PowerConvert, LeoStream, HP Systems Insight Manager (SIM) virtual machine management (VMM) and server migration pack (SMP) which can also perform incremental upgrades, plus V2V, V2P, etc. – also some freeware, e.g. Bart PE with Ultimate P2V.
  - Illegal to P2V an OEM-licensed version of Windows.
- Workstation:
  - VM platform for the desktop – runs as an application (hosted architecture) on Windows or Linux host.
  - Guest OS can be Windows, Linux, NetWare or Solaris x86.
  - Access to wide range of hardware via the host OS (e.g. wireless LAN) – virtualisation layer maps this from guest to host.
  - 2-way VSMP support is experimental.
  - Support for 64-bit OSs (even on 32-bit hardware).
  - Snapshot capability (requires a flat file).
- ACE:
  - Allows packaging of OS, applications, data and security settings into an "assured computing environment".
  - Lock down PC endpoints and protect company resources.
  - ACE Manager creates a pre-packaged desktop environment for deployment to a user's unmanaged PC.
  - Virtual rights management (VRM) used to control VM lifecycles, secure data and ensure compliance with IT policies – takes over network stack and enforces policy (patching, USB devices, expiry dates, encrypted disks, network access, copy protection).
  - Burn image to CD/DVD or place on a network share for installation or provisioning using a third-party tool.
  - ACE runs as a Windows application (Linux soon).
- Player:
  - Free software – run any preconfigured 32- or 64-bit VM on any PC.
  - Installed as a standard application on Windows or Linux.
  - VMs available for download from VMTN – e.g. browser appliance.
  - Can configure for performance/host PC devices but cannot create VMs.
- Server:
  - Free software – partition physical server to run multiple VMs – introduction to virtualisation.
  - Hosted architecture – runs as a standard Windows or Linux application but license invalidated if run on a workstation OS.
  - Use for test and development, evaluation, re-hosting legacy OSs, fast provisioning, or virtual appliances.
  - Can run any VM created by GSX, ESX or Workstation.
  - Support for Intel VT, VSMP and 64-bit guests.
  - Scale up to 16 CPUs and 64GB RAM (subject to host OS limits) – 2-4 VMs per core.
  - Can be managed with VC 1.4 – VC and agents are not free.
- VMware Technology Network (VMTN):
  - Online resources and subscription.
  - Annual subscription allows access to products for development and testing, including workstation, server, VI3 standard (not VC) and P2V Assistant.

Pricing, packaging and licensing

- VI3 sold in increments of 2-processors.
- Three editions – started, standard and enterprise:
  - Starter = ESX3, limited to 4 CPUs and 8GB RAM, no FC or iSCSI storage (NAS or local only), VMFS for local storage only, VC agent.
  - Standard = ESX3, VMFS, VSMP (up to 4-way), VC agent.
  - Enterprise = ESX3, VMFS, VSMP (up to 4-way), VC agent, VMotion, HA, DRS, VCB.
- 1 year of support is mandatory with all purchases and most upgrades (exception is purchase of a la carte upgrades).
  - Support and limited subscription (SLS) – minor upgrades only (i.e. 3.0.x-3.1.x).
  - Support and subscription (SnS) – major and minor upgrades (i.e. 3.0-4.0 and 3.0.x-3.1.x).
  - Cannot switch from SLS to SnS.
  - Subscription option pricing based on the list price of the related products (or bundle).
  - Gold support is business hours x 5 days – Platinum is 24 x 7.
- Licenses can be managed per-host or using a license server.
- VMware Infrastructure Acceleration Kit = VI3 Enterprise for 8 processors, VCMS, P2V Assistant Starter Edition, Gold SnS.
- A la carte products – VC, VMotion, HA, DRS, VCB.

- VC licensed per server/installation.
- VCB is not available for Starter Edition.
- VI3 will become VI4 when either ESX or VC have a major release.
- VMFS and VSMP inherit product number from ESX.
- VC agent, VCMS, VMotion, HA, DRS, VCB inherit product number from VC.
- Entitlement – a free of charge upgrade under the terms of SnS (e.g. ESX 2.x to 3.x).
- Upgrade – a paid-for change of product version (e.g. Standard to Enterprise).
- VMware GSX customers can upgrade to VI3 starter at no cost (not automatic entitlement – must be ordered). 1 year's support for the upgrade is also mandatory.
- Upon order, activation code issued via e-mail – redeemed on VMware website to obtain license file/key.
- Pre-VI3 licenses cannot be purchased but it is possible to obtain ESX 2.1/VC 1.x license keys. A VI3 replacement can be downloaded later at no charge.
- Centralised license management using FlexNet License Server (requires a Windows server). Existing FlexNet installations can be used, but default is to install with VC.
- Enables flexible license management for new hosts, upgrades and reallocating capacity.
- All licenses in a single file (.lic) – dynamically tracked and allocated in 2-processor increments.
- ESX Servers will continue to function, even if license server unavailable (14 day grace period, after which unable to power on VMs); however certain features (add host, enable features, upgrade, etc.) require a license server to be available.
- Decentralised licensing also available (e.g. where VC not in use).

Minimum hardware and software requirements

- VC:
  - 2GHz x86 CPU – more if database on same server.
  - 2GB RAM – more if database on same server.
  - 560MB free disk space (245MB for program and 315MB in %temp%) – 2GB recommended – more if database on same server (25 hosts, 8-16VMs each, 1 year, default settings will be 2.2GB SQL or 1.0GB Oracle).
  - 10/100Mbps Ethernet NIC (1000Mbps recommended).
  - Windows 2000 + SP4/XP/2003 (not 64-bit).
  - IE 5.5 or later.
  - Will scale to 20 concurrent clients, 50 ESX hosts, 1000 VMs – if 2xCPU and 3GB RAM then 50 concurrent clients, 100 ESX hosts and >2000 VMs.
- VC Database:
  - SQL Server 2000 SP4, Oracle 9iR2/10gR1 (10.1.0.3 or later)/10gR2, MSDE (not in production).
  - MSDE – 2GB to install but 1.5GB deleted post-installation.
- VI client:
  - 266MHz x86 CPU – 500MHz recommended.
  - 256MB RAM – 512MB recommended.
  - 150MB free disk space (55MB for program files and 100MB in %temp%).
  - 10/100Mbps Ethernet NIC (1000Mbps recommended).
  - Windows 2000 + SP4/XP/2003 (not 64-bit).
  - Microsoft .NET Framework v1.1 (included in installation).
- License Server:
  - 266MHz x86 CPU – 500MHz recommended.
  - 256MB RAM – 512MB recommended.
  - 25MB free disk space.
  - 10/100Mbps Ethernet NIC (1000Mbps recommended).
  - Windows 2000 + SP4/XP/2003 (not 64-bit).
  - Can run in a VM but not recommended.
- ESX:
  - 2x1.5GHz Intel Xeon/Viiv or AMD Opteron (32-bit mode)/Athlon 64x2 CPUs.
  - 1GB RAM.
  - Supported NIC (e.g. Broadcom 570x or Intel PRO/100).
  - SCSI, Internal RAID or FC HBA (IDE/PATA drives can be used for ESX but not for VM storage – SATA disks are unsupported).

Maximum hardware limitations

- ESX:

- 16 HBAs with 15 targets per HBA.
- 128 LUNs per storage array.
- 255 LUNs per ESX host (maximum LUN ID=255).
- 32 paths to a LUN.
- 128 VMFS volumes (each volume up to 32 extents, each extent up to 2TB).
- 128 vCPUs for all VMs running on the host.
- 64GB RAM.
- 64 adapters of all types (including storage and network).
- 20 NICs.
- 1024 ports per virtual switch.
- Up to 200 registered VMs.

ESX deployment

- ESX Server architecture:
  - VMs run on top of virtualisation layer (VMkernel).
  - Applications within VMs do not directly access hardware – all access is via the virtualisation layer.
  - Service console (based on Red Hat Enterprise Linux 3, update 6) supports administrative functions (firewall, SNMP, Apache Tomcat HTTP/Java server) and other services (systems management agents, etc.) – do not install standard Red Hat patches as they may break customisation.
  - VMkernel assumes that hardware is functioning – failures can cause ESX server to fail. One approach is clustering ESX Servers.
  - Supported on x86 hardware with Intel Xeon or AMD Opteron (in 32-bit mode) processors. Support for 64-bit guest OSs is experimental.
- 2 main considerations for installation – network connectivity and storage.
- ESX service console and VMkernel may be installed on either local or SAN-based disk.
- Boot options are local SCSI (or SAS), IDE/PATA (not SATA) or SAN.
- x86 disks limited to 4 primary partitions. Alternatively a primary partition can be replaced with an extended partition containing up to 63 logical partitions (IDE) or 15 logical partitions (SCSI).
- Linux files systems use mountpoints to build a hierarchy. The size of a directory is capped by the size of the corresponding partition and use of mount points effectively prevents the root partition from being filled.
- Default file systems:
  - /boot (boot disk) – ~100MB ext3 on boot disk (primary partition) – used by service console – most systems will need this to be at the beginning of the disk.
  - / (root) – ~5GB ext3 on boot disk (primary partition) – used by service console – minimum 2560MB.
  - not mounted (swap) – ~544MB swap on boot disk (logical in extended partition) – used by service console – create as 2x memory assigned to service console – if unsure then make 1.6GB (2x800MB – i.e. twice the maximum service console memory allocation).
  - /var/log (log files) – ~2GB ext3 on boot disk (logical in extended partition) – used by service console – minimum 500MB, 2GB recommended.
  - automatic (VM files, .iso images, etc.) – VMFS3 on any local disk (primary partition) – used by VMkernel.
  - not mounted (core dump) – ~100MB on any local or remote disk (logical in extended partition) – used by VMkernel – last partition on disk so can overrun.
- Others that may be useful:
  - /opt – >2GB – avoid HA filling root partition.
  - /tmp – ~2GB – avoid temporary files filling root partition (e.g. if run backup software within the service console).
- Install ESX in text mode or graphical mode. Text mode is useful for remote console access over a slow connection. Graphical is the default (1 minute timeout).
- Installer includes media test (validate downloaded .iso image).
- Mouse or tab past welcome screen, then set keyboard and mouse options (mouse irrelevant once installed as X Windows not supported in service console).
- Mandatory license acceptance.
- Will initialize all visible unpartitioned drives so beware with shared LUNs – zone and mask away SAN LUNs (remove cable as a precaution):
  - /dev/sda is SAN-based device.
  - /cciss/c0d0 is local SCSI device.
- Boot loader must be placed on first device in BIOS. Use LBA32 if exceed 1024 cylinder limit for boot partition.
- DHCP is an option but static addressing is recommended. Can also supply VLAN ID if required.

- Set time zone based on map, location (recommended) or UTC offset (daylight saving only works in United States).
- Root password minimum 6 characters. Best practice suggests include complexity.
- Summary screen before changes made.
- Connect to server in a browser via IP address or hostname (option to download VI client).
- VI web client can manage VMs only.
- VI client can manage ESX or VMs – also used for VirtualCenter functions by connecting to VCMS (users and groups only available when connect to ESX Server).
- Can also log on to service console (Alt-F1), Alt-F11 to return to main screen.
- Remote service console sessions are available using SSH – by default, root cannot log on via SSH, so create a non-root account and use su – (except on a server that was upgraded from ESX2).
- Prior to deployment, check compatibility guides (systems, I/O and Storage/SAN) – if not listed then no driver support (particularly I/O).
- Pre-installation checklist:
  - At least one supported NIC.
  - Network connectivity to a PC for management.
  - Network configuration details (IP address, hostname, subnet mask, gateway and DNS server).
  - VMware ESX Server installation CD and license file.
  - Guest OS media (or .iso image) and license keys.
- ESX Server sizing (VMkernel):
  - RAM – sum of maximum memory usage per VM (or minimum if over committing resources), plus overhead for each VM (79-350MB - varies according to number of vCPUs, amount of VM memory and whether the VM is 32 or 64-bit), ~50MB for VMkernel and memory for each device driver (can be read within the VI client).
  - Storage – each VM requires as much virtual disk as it would physical, plus configuration files (negligible) and a VMkernel swap file equivalent to the VM's maximum RAM.
  - CPU – Total number of CPU cycles for all VMs – more if Gigabit Ethernet in use.
  - NIC – sum of bandwidth used by all VMs.
- ESX Server sizing (service console):
  - RAM – up to 800MB, 272MB default (recommended).
  - Storage – file systems as above.
  - CPU – one, shared with VMs.
  - NIC – one, shared with VMs or dedicated.
- Boot from SAN can be useful – DR, cost reduction, hardware restrictions, cold boot any server to VI3; but only 1 HBA can be the primary and it is a primitive connection until the HBA driver is loaded (i.e. no multipathing).
- Both Emulex and QLogic have supported HBAs – QLogic uses the BIOS for configuration so a single driver instance is used for multiple HBAs and HBAs can be transferred between servers – Emulex holds configuration in the driver, so one driver instance is required per HBA.
- FC HBA BIOS must be used to initiate a connection to the target boot LUN and the server BIOS must identify the FC HBA as first boot device.

Networking

- ESX Server networking is based around virtual switches (vswitches). pNICs are associated with vswitches and VMs gain access to the network via vswitch ports. A vswitch must exist between a VM and a pNIC.
- vswitches are also used by VMkernel for VMotion and to access iSCSI/NAS-based storage.
- vswitches may be internal-only (i.e. not connected to any pNICs), or may connect to multiple pNICs for NIC teaming with automatic distribution of packets and failover.
- vswitches also used to allow access from the service console to a management LAN.
- New vswitches default to 56 ports (although vswitches created at installation time only have 24 ports). The maximum is 1016 ports per vswitch. Increasing the number of ports on a vswitch will necessitate restarting the server.
- Internal-only vswitches offer cost savings, are efficient, and allow isolation – networking limited to single ESX server but no collisions and communication occurs at bus speed (traffic shaping is not supported).
- The simplest way to connect a VM to a network is to create a vswitch and bind one pNIC to it – each vNIC will have a MAC address and outbound bandwidth can be controlled with traffic shaping.
- A VM can be multi-homed on an internal switch and a switch with an outbound adapter to act as a router (e.g. one-box firewall environment).
- Create a NIC team with a vswitch connected to multiple pNICs – the act of adding the NIC will automatically set up the team to balance the network load – each vNIC will have a MAC address and outbound bandwidth can be controlled with traffic shaping. Connection to an 802.3ad NIC team is possible (but the ESX Server

implementation is not compliant with 802.3ad).
- There are three type of network connection – defined when create switch using the VI client – can add more connections later:
  - Service console port – access management network.
  - VMkernel port – VMotion, iSCSI, NFS/NAS.
  - VM port group – VM networks.
- Multiple connection types can exist on a single vswitch but will share same pNICs as this is defined at vswitch level (which NICs are active can be configured at port group level).
- When creating a service console port, define network label, VLAN tag (optional) and IP settings.
- When creating a VMkernel port, define network label, VLAN tag (optional), whether or not to enable VMotion and IP settings.
- Only one VMkernel port can be created per server.
- Separate IP stacks are configured for service console and VMkernel (i.e. each has their own IP settings).
- DHCP possible for service console but not recommended.
- When creating a VM port group, define network label and VLAN tag (optional) – IP settings are controlled within the VM.
- VMotion traffic is not encrypted, so it is recommended to use own pNIC (as well as for performance reasons).
- pNICs always run in promiscuous mode (configured by ESX Server) – hence limited number of supported NICs.
- NIC location identifier is made up of the PCI bus, slot, and function (e.g. port number for a multi-port NICs) – e.g. PCI 02:02:0.
- NIC driver e1000 indicates an Intel NIC and tg3 is Broadcom.
- All vswitches are known as vswitch# where # is a number starting at 0 and working up sequentially.
- Every port or port group has a network label (user defined).
- Service console ports are named vswif#.
- Every vswitch has properties for general (number of ports) and network policies.
- The associated network adapters can be configured within the virtual switch properties (link speed and duplex – default is autonegotiate).
- Network policies are defined at the vswitch (default policies), port or port group level (effective policies overriding defaults) – there are four policies:
  - VLAN – creation of multiple logical LANS within or across physical segments (operate at layer 2).
  - Security – configuration of layer 2 Ethernet security policies.
  - Traffic shaping.
  - NIC teaming.
- VLANs improve security (the switch only presents frames to nodes in the correct VLANs), improve performance (each VLAN is its own broadcast domain) and reduce cost (less hardware required).
- With Cisco equipment, VLAN 1 has access to all areas – because vswitch VLAN numbers correspond to pswitch VLAN numbers, vswitch VLAN 1 will also have access to all areas.
- ESX Server supports 802.1q VLAN tagging (industry standard for implementation).
- To extend VLANs across switches at trunk line is required. Trunk frames include an additional 4 bytes after the source and destination MAC address – the first 2 bytes indicate that this is an 802.1q frame and the next 2 contain the 12-bit VLAN ID. Sniffing is possible by plugging into the trunk port.
- ESX Server provides VLAN support through vswitch tagging (giving a port group a VLAN ID) – VMkernel then tags and untags packets as they pass through the vswitch with little impact on performance. The connection to the physical NIC operates as a trunk port.
- Three security policy exceptions are available:
  - Promiscuous mode (i.e. see everything on the wire) – when set to reject (the default setting), the vNIC will appear to go into promiscuous mode but won't receive additional frames.
  - MAC address changes – when set to reject, if the guest OS attempts to change the MAC address it stops receiving frames (default is accept).
  - Forged transmits – when set to reject, drop any frames that the guest sends where the source MAC address differs from the virtual hardware (default is accept).
- Outbound traffic can be controlled using traffic shaping, based on average and peak bandwidth, specified in Kbps as well as burst (bandwidth x time), specified in KB – inbound must be controlled by the router or by load-balancing.
- Traffic shaping is disabled by default and although defined at switch or port group level is applied per-VM and per-vNIC.
- NIC teaming can be used to distribute traffic across pNICs and reroute in the event of failure. Options are:
  - Load balancing (outbound only) – route based on:
    - Originating port ID (default) – simple and fast – no need for VMkernel to view frames – once assigned,

remains in place until hardware changes (e.g. failure).
    - IP hash (hash of source and destination IP address) – slight CPU overhead (hash calculation) and requires the switch to support 802.3ad link aggregation (because vNIC MAC can appear via multiple routes and hence on multiple switch ports) but has better distribution of traffic across pNICs.
    - Source MAC address – low overhead and compatible with all switches but may not spread traffic evenly (i.e. the same source address will follow the same route and if more pNICs than vNICs, some pNICs will be unused) – can create problems with unicast traffic (as every node uses an identical, spoofed, MAC address).
  - Network failure detection – detected by VMkernel – link state (not very sophisticated and will not detect blocked switch ports of remote cable pulls) or link state and beaconing (listening for probe packet).
  - Notify switches – whenever there is a failover event of a new vNIC is connected to the vswitch – do not use with unicast most Microsoft NLB as vMAC address is used rather than NLB MAC address, causing dropouts when VMotion or otherwise move NIC (multicast NLB is not affected) – workaround is to create own port group and disable switch notification.
  - Rolling failover – preferred uplink list sorted by uptime (recently failed adapter remains inactive until another fails).
  - Failover order – explicit list of links (uses highest priority link which is available).
- Multiple policies can be applied to the same pNICs, so different port groups can apply their own set of active and standby adapters, traffic shaping, security and VLAN IDs.

Storage

Fibre Channel

- A fibre channel (FC) SAN consists of the following components:
  - Storage system, with arrays for physical hard disks exposing themselves as logical units (LUNs), each with a LUN identifier.
  - Storage processors (SPs).
  - Fibre channel switches and cables (the interconnection of which forms a fabric) – multiple switches allow for redundancy.
  - Servers with FC host bus adapters (HBAs).
- Technically, connected devices (servers, storage systems or tape drives) are known as nodes.
- Unique, 64-bit world wide names (WWNs) are assigned by SAN equipment manufacturers to HBAs and SPs – they are used by SAN administrators to identify equipment for zoning purposes.
- Host access to LUNs is controlled by zoning (SP-based soft zoning of LUN visibility on a per-WWN basis or switch-based hard zoning of SP visibility on a per-port basis) or LUN masking (usually at SP-level but possible at server-level) to make a LUN invisible during scanning – masking is important on a shared SAN (e.g. to prevent Windows administrators from initialising VMFS LUNs) – zoning effectively splits the fabric and prevents incompatible OSs from interrogating one another's devices – soft zoning allows cable replacement/movement but requires reconfiguration if an HBA is replaced – hard zoning can use any HBA but breaks if cables/ports are changed – a node may exist within multiple zones but cannot communicate outside its zone – zones can cross switches.
- World wide node name (WWNN) is the node, world wide port name (WWPN) is the port on a node (e.g. in multi-port HBAs).
- VMkernel addresses disk partitions as *vmhbaadapter*:*target*:*LUN*[:*partition*] – the target is the SP or disk array on the SAN (there may be multiple paths to the same LUN).
- In common with all supported PCI devices, the FC storage adapter is recognised by VMkernel at boot time. ESX Server supports 256 LUNs in the range 0-255; however the ESX installer will only see the first 128 (the rest will be detected by a rescan).

iSCSI

- Internet SCSI (iSCSI) provides transport for SCSI block storage over standard TCP/IP networks (cf. SCSI over FC)
- Initiators (e.g. an iSCSI HBA) send SCSI commands to targets, located in iSCSI storage systems. Arrays of physical hard disks are represented as LUNs via SPs (as for FC storage).
- Key features:
  - Lower cost than FC and uses existing NICs.
  - Use existing infrastructure (i.e. CAT5 wiring and switches) and knowledge (iSCSI routing is identical to regular Ethernet).
  - Authentication and encryption – provided by IP or VPNs.

  - Internet-ready – can be used for long haul data transfers, e.g. between two geographically-separated data centres (FC requires a gateway to tunnel through or convert to IP over extended distances).
- ESX Server supports iSCSI for VMFS, raw disk access, VMotion and boot-from-SAN.
- Although ESX Server can be booted from iSCSI, it is only possible with a particular iSCSI hardware initiator (QLogic QLA4010) – the boot order must also be set in the BIOS and hardware initiator support is experimental, using a TCP offload engine to manage its own IP stack.
- There is also a software initiator, implemented inside the VMkernel which uses standard NICs and a Cisco iSCSI initiator command reference implementation (working with a daemon inside the service console); however this cannot be mixed with the hardware initiator inside the same server.
- Both the service console (initiator daemon) and VMkernel (I/O) need access to the iSCSI storage – either by sharing a switch and subnet, or by using routing to allow both access to the storage – outgoing port 3260 also needs to be enabled in the service console firewall (using the VI client).
- iSCSI naming is specified in RFC3270 and all nodes (target and initiator) require names for the purpose of identification – the iSCSI qualified name (IQN) is made up of iqn.*year-month.topleveldomain.subdomain*[:*uniqueid*] – the year and month represent the registration date and the uniqueid is assigned by the organisation – iSCSI aliases and IP addresses will also be defined.
- iSCSI LUNs are discovered by static or dynamic (send targets) configuration – static configuration includes IP address, TCP port number and iSCSI target name (suitable for small networks only) whereas with dynamic configuration the initiator issues a SendTargets command to a known IP address and TCP port and the iSCSI device responds with the available targets – the software initiator only supports SendTargets.
- It is good practice to create a separate IP network or VLAN for iSCSI traffic as data is unencrypted. Challenge handshake authentication protocol (CHAP) is used to authenticate initiators trying to access a device (and can be enabled on the initiator too), sending a random hash challenge, which is encrypted using the password and sent as a response – if this response matches the device's own calculation of encrypted password then the password must be correct and has been validated without sending the actual password in clear text across the network.
- iSCSI software initiator must be enabled, after which a default iSCSI name and alias are chosen –dynamic discovery requires the name and port for a server to send targets – static discovery allows the manual addition of targets that are accessible to the ESX Server but is not supported for the software initiator – CHAP authentication may also be enabled using a name (e.g. the initiator name) and secret (password) – once configured, a rescan should return a list of iSCSI LUNs.
- Although iSCSI can be routed across subnets, doing so will impact performance.
- An iSCSI appliance is available on the VMTN website.

VMFS

- Virtual machine file system (VMFS) is a file system optimised for ESX Server VMs that can be deployed on a variety of SCSI-based devices (including FC and iSCSI SANs) – accessible in the service console under /vmfs/volumes or addressed by the volume label, data store name and physical address (vmhba*A*:*T*:*L*:*P*).
- Unlike NTFS, VMFS is designed as a clustered file system and so multiple hosts can address the same LUN. VMFS is used by VMkernel and is accessible in the service console; however VMs cannot see the VMFS.
- The maximum file size is linked to the block size; however, sub allocation allows multiple small files to be placed in once block.
- It is not possible to tell whether the VMFS is located on iSCSI or FC without looking at the storage adapter configuration – may be advisable to include an identifier within the data store name.
- VMFS can be extended in order to span multiple LUNs (an extent), or to exceed the 2TB limit during creation – there are some limitations in that it writes to disk 1, then disk 2 (so there is no performance increase) and a disk fails then the entire volume is lost.
- Cannot create an extent on a volume with an existing VMFS – if the chosen LUN already has data (e.g. NTFS) then a warning is displayed that data will be permanently lost.
- Removing an extent requires removal of the entire volume.
- VMFS can be mounted read-only using VCB tools.

Multipathing

- MPIO is required for multiple HBAs otherwise data is presented multiple times – traditionally provided via expensive software – ESX Server and Windows Server 2003 can manage out of the box.
- FC multipathing ensures access to a LUN in the event of hardware failure – automatic failover after a configurable delay (configured at HBA) – only one path is active at a time (and individual paths can be enabled or disabled) – 2 failover policies exist – the default is most recently used (MRU) and works well for active/passive devices, alternatively fixed (preferred) path may be used (e.g. for active/active) whereby a preferred path is set

and will be reverted to when becomes available – preferred and active paths may be set for each LUN.
- iSCSI multipathing is facilitated using dynamic IP routing and multiple paths are recognised from a send targets discovery – MRU and fixed paths are both available; however only active-passive configurations are supported – software initiator looks like a single HBA; however it can present multiple paths at the networking layer as it can use multiple NICs.
- FC and iSCSI HBAs can exist in the same server but are not supported for sharing a LUN.
- Paths may be managed within the VI client (within the properties for the VMFS).
- If multiple paths are available to a LUN ESX server will recognise this by the UUID – LUN known by first found path.

NAS

- Network attached storage (NAS) is accessed across the network at file system level (cf. iSCSI at block level), providing a low cost solution with moderate performance and a lower infrastructure investment than FC.
- NAS either uses the network file system (NFS) or server message block (SMB/CIFS) – many appliances support both but VMkernel only supports NFS v3 over TCP (service console can use Samba for SMB/CIFs access.
- NFS volumes are treated just as VMFS volumes in FC or iSCSI storage – can hold VMs, .iso images and templates and also support VMotion – performance is not great for large files.
- ESX Server uses a VMkernel port on a vswitch (new or existing) to access the NFS server which contains a directory to share with ESX – /etc/exports defines the systems that are allowed to access the shared directory including name or directory, subnet(s) allowed access, access (e.g. rw), no_root_squash (i.e. don't limit access from the root user – required for ESX Server) and sync (commit all file writes to the disk before regarding a request as complete).
- NFS volumes are addressed by their IP address and folder name – can be mounted read-only (e.g. if all that they contain is .iso images for software installation).
- By default, ESX Server supports 8 NFS mounts – this can be increased.
- For performance reasons, VMs should not be configured to swap to NFS volumes (use VMFS instead) – this is controlled by editing the sched.swap.dir= entry within the VM's .vmx file.

Further information

- Comparison – FC provides high performance due to dedicated network, whereas iSCSI and NAS performance is dependant upon the condition of the IP network – FC and iSCSI operate at block level, NAS is at file system level (i.e. no direct LUN access is available) – iSCSI does not support VM clustering or VCB – NAS does not support any features which require block-level access (boot-from-SAN, VMFS or RDM), VM clustering or VCB.
- In general, one LUN should be used for one purpose – where used for shared storage (e.g. VMotion), multiple servers will require visibility.
- Only one VMFS is supported per-LUN but separate VMFS should be used (on separate LUNs) to support separate environments (e.g. test and production).
- RDMs should be used in physical-to-virtual (standby host) clusters or cross-host clustering as well as to support hardware snapshotting within a SAN.
- Snapshots of raw disks, RDM physical mode disks or independent disks are not supported – revert to snapshot goes to parent snapshot – go to snapshot reverts to another snapshot.
- For improved iSCSI security and performance it should run on a separate and isolated IP network.
- For improved NAS security and performance it should run on a separate and isolated IP network.
- Each LUN should have the appropriate RAID level and storage characteristics for the applications running within the associated VMs – preferred paths should be used to spread the I/O and RAID volumes with more than 7 disks may suffer reduced performance due to parity calculations.
- Either use separate volumes for each RAID level and assign system/data vdisks accordingly or use large LUNs and monitor performance to identify VMs with performance issues.
- vdisks can only be read by mounting them within a VM – once mounted they are locked – multiple VMs on a single LUN is not a security issue – recommended to host multiple VMs on a single VMFS in order to identify VMs that compete for disk access.
- Up to 32 low I/O VMs per LUN, 8-16 medium I/O VMs per LUN or provided dedicated LUN or RDM for VMs with high I/O requirements.

VirtualCenter

- VC allows central management of multiple ESX hosts and VMs – also allows use of VMotion, HA, DRS, etc. – maximum 10 tasks concurrently (after which it queues tasks) – alternative is to schedule tasks.
- Architecture:

  - Core services – resource and VM inventory management; alarms and events management; statistics logging; VM provisioning; task scheduler; host and VM configuration.
  - Distributed services – DRS; HA and VMotion.
  - Database interface.
  - ESX Server management – accesses ESX Server via VC Agent (communicates via VI API to host agent).
  - User access control and Active Directory interface for access to domain user accounts.
  - VI API for third party application support (using SOAP).
- Software components:
  - VC server – service to direct actions to be taken within VMs on ESX Servers – recommended that hosted on a physical server.
  - VMware License Server – server based licensing for VC and ESX Server functionality – recommended that co-hosted with VC server.
  - VI client – GUI interface to VC/ESX Server.
  - Web client – web interface for managing VMs.
  - VC database – repository for VC information including performance and configuration data.
  - VC agents (vpxa) – processes on ESX Servers used to receive tasks initiated by the VC server
- VI client or web client can communicate with VC or with the hostd process on ESX Server. VC communicates directly with the VC agent.
- Installation order is database (Microsoft SQL Server 2000 SP4 or Oracle 9iR2/10gR1 – 10.1.0.3 or later – MSDE 2000A for evaluation/demonstration only and not supported in production), License Server, VC Server, VI client.
- VC server has 3 services:
  - VMware Virtual Infrastructure Web Access.
  - VMware Virtual Mount Manager Extended – used during guest OS customisation (cloning or deploying from template).
  - VMware VirtualCenter Server.
- Security uses Windows accounts – if server is a member of a domain then access is automatically available from accounts in that domain and all trusted domains.
- If firewall between VC and ESX Server, ensure that port 902 is open.
  - Can change port for managing VC but not ESX (always 902).
- VC inventory is a hierarchy of objects (containers or objects to manage – such as hosts and VMs) – topmost entry is a folder (root) under which a datacenter is created containing all the different types of object needed to work in a VI (hosts, VMs, networks and data stores).
- Datacenters are the primary organisational structure – managed objects belong to a single datacenter – group hosts in a datacenter under single administrative control and to meet VMotion requirements – folders can be used to group datacenters (e.g. Americas, EMEA and Asia-Pacific) – multiple levels of folders can be used.
- VMs, templates and hosts can also be organised into folders – organise VMs based on business unit/function and hosts based on CPU family
- Hosts can be standalone or clustered in a group to form a single pool of resources – clusters used for HA, DRS or both.
- There are 4 inventory hierarchies available: hosts and clusters; virtual machines and templates; networks; data stores.
- VC Server with minimum hardware support 20 concurrent client connections, 50 managed hosts and 1000 VMs – using dual CPUs and 3GB RAM can scale to 50 concurrent clients, 100 hosts and 2000 VMs
- If VC server fails, ESX Servers will continue to run but no VMotion/DRS and HA may fail, no scheduled tasks and no monitoring – when VC is restarted it can reconnect to running hosts and synchronise the state of hosts and VMs.
- ESX Server can only be managed by 1 VC server at a time and if build a new VC server will lose statistics
- Recommended to create a standby server (powered off) until it needs to take the place of the primary server and to use the clustering capabilities of the database to provide resilience – standby server does need to be licensed though.
- Database maintenance activities include: monitoring the growth of the log file and compacting as needed; scheduling regular backups of the database; and backing up the database prior to any VC upgrade.
- Custom attributes can be used for VC objects – always a string – stored in the VC database (i.e. not used by ESX).
- Scheduled tasks are timed according to the VC server (not the VI client) Only one timing frequency can be set for a task – if tasks need to occur at other intervals, set up addition tasks – removal of a task stops future occurrences (whereas cancelling applies to a running task).

Create and manage VMs

- VM consists of set of discrete files:
  - *vmdisplayname*.vmx – configuration file – describe virtual hardware (CPU, memory, HDD, NIC, CD-ROM drive, FDD, etc.).
  - *vmdisplayname*.vmdk – virtual disk description (~350bytes).
  - *vmdisplayname*-flat.vmdk – virtual disk.
  - *vmdisplayname*.nvram – Phoenix BIOS 4.0 release 6.
- Avoid special characters in display name as will affect the file names (and hence access to the files).
- Hardware is uniform (old but solid – provides legacy support):
  - Mouse.
  - Keyboard.
  - Up to 16GB RAM.
  - VM chipset (440BX and NS338 SIO) and 1 CPU (2 or 4 CPUs with Virtual SMP).
  - 6 PCI slots:
    - 1 Video adapter.
    - 1-4 SCSI adapters with 1-15 devices each.
    - 1-4 Ethernet NICs.
  - Up to 4 IDE CD-ROMs.
  - Up to 3 parallel ports.
  - Up to 4 serial ports.
  - 1-2 FDDs.
- MAC address calculated using service console IP address.
- Windows NT guests only support 3.4GB RAM.
- Multiple vCPUs requires Virtual SMP; however a VM with 2 or 4 vCPUs will have to wait for 2 or 4 cores to be available simultaneously – resulting in scheduling inefficiencies – many guest OS/application combinations are not enhanced with additional CPU – in general, P2V dual-CPU servers to single vCPU.
- vdisks are monolithic and pre-extended – i.e. a 6GB disk will be a 6GB file – maximum size is 9TB.
- Adding the first vdisk automatically adds a vSCSI HBA – can chose whether LSILogic or BusLogic adapter is used.
- Disk modes include:
  - Snapshotting is supported by default – allow return to known state – e.g. training/test and development – can also be used for backups.
  - Independent modes – persistent (changes written to disk immediately) and non-persistent (changes discarded when VM is powered off).
- NIC is vlance – connect to virtual switch.
- CD-ROM drive – connect to physical device or .iso image.
- FDD drive – connect to physical device or .flp image.
- SCSI HBAs may be used for generic SCSI devices such as tape libraries.
- Access VM via console (in VI client) – access BIOS (F2) or boot menu (ESC), power on/off and reset VM (Ctrl-Alt-Insert), access guest OS.
- Install OS from virtual CD-ROM – .iso file stored on VMFS, NFS or in /vmimages within service console – VMFS or NFS recommended as accessible to all hosts – /vmimages is part of SC root file system (may be better to create on separate partition).
- VMware Tools (VMtools) should be installed into guest OS post-installation – device drivers for virtual video card (VMware SVGA II), mouse (VMware Pointing Device) with hardware acceleration (need to set to full under Windows Server 2003) and ability to move mouse outside of console, optimised SCSI driver (VMware SCSI Driver) and memory management (vmmemctl) – support for quiescing a file system (used by VCB) – optionally enable time synchronisation between guest and host (do not use if AD or NTP services are also doing this) – can also define scripts for event handling (e.g. suspend guest OS).
- It is best practice to install VMtools on every VM.
- Templates used for commonly-deployed VMs – VM that is marked as never to be powered on – disk can be monolithic or sparse (no pre-allocated space, cut into 2GB chunks) – can be stored on VMFS, NFS or within the service console file system – as for .iso (and .flp) files, VMFS and NFS are recommended. Sparse files can be used for templates but not for VMs.
- Creation of templates in the GUI is a VC operation – for ESX-only environments then needs to be performed at the command line.
- Create a template (.vmx -> .vmtx) by cloning or converting (right-click VM in VI client) – use clone to retain original VM.
- Update a template by placing it on an isolated network (prevent user access), convert to VM, make changes and

convert back to a template.
- Deploy VM from template by right-clicking on template in VI client.
- Guest OS customisation allows customisation of a clone to prevent software and network conflicts – enable for Windows VMs (2000, XP and 2003) by downloading the Microsoft System Preparation Tool (sysprep) and extracting to %allusersprofile%\Application Data\VMware\VMware VirtualCenter\sysprep\1.1 (Application Data is usually a hidden folder) – for Linux VMs, open source components are installed with VirtualCenter Server – more convenient than sysprepping the OS in the template and having to supply information for each deployment.
- Moving a VM to another server via a cold migration (powered off) may or may not require movement of disks, depending on where the files are stored – use cold migration when different processor family or moving to a non-shared data store (may still fail with certain Linux distributions that are compiled for a particular CPU).
- Most devices must be added to VMs whilst they are powered off – Floppy Drives and CD-ROM Drives can be mounted whilst the VM is online but only hard disks are hot pluggable – there are some exceptions as this requires a VM3 format VM (i.e. not an ESX Server v2.5 VM) and hard disks cannot be hot-removed.
- VMs can access raw SAN LUNs using a raw device mapping (RDM) – this is a special file on a VMFS or NFS data store that points to the actual SAN LUN – takes no space on the disk but a directory listing will report it as taking the same size as the physical target.
- Tape devices and some storage processors have LUN IDs (so can be accessed via RDM); however this is not supported.
- Can directly connect VMs using a serial port as a named pipe (need to identify client and server).
- Wake On LAN can be configured in power management properties for a VM.
- VM creation and deployment best practice:
  - Plan load mix, understand goals and expectations, understand requirements (and how success is defined), avoid mixing VMs with competing resource requirements, test before deploying – virtualisation allows sharing of resources but has some overheads – does not create new resources.
  - Size VMs according to needs – over-configured VMs waste shareable resources – disable unused devices (e.g. COM ports) and install VMtools (to optimise performance).
  - Tune the guest OS as it would be for a physical computer registry, swap space, etc.) – disable unnecessary programs such as screen savers – keep the guest OS up-to-date with the latest patches.
  - Tune and size applications on VMs as they would be for physical computers – don't run single-threaded applications in an SMP VM.
  - Avoid high memory reclamation (ballooning and swapping) by correctly-sizing VMs and avoiding memory over-commitment.
  - For NUMA systems, memory should be evenly balanced across NUMA nodes (memory and pCPU) to allow the NUMA scheduler to co-locate VM memory and vCPUs, avoiding remote memory access.

VM access control

- The VI security model is based around users (login accounts), roles (groups of privileges), privileges (a task that a user may perform, grouped by categories and subcategories) and permissions (the pairing of a user and a role).
- VC users and groups are inherited from the VC server's domain – ESX Server users and groups are defined in the service console – no attempt is made to reconcile (synchronise) VC users with ESX Server users.
- Privileges apply to the VI client and SC (ESX commands only – not standard Linux commands). Within a VM, access is controlled by the guest OS.
- ESX Server default roles are: no access; read-only; and administrator – VC includes these three, plus sample roles (i.e. can be modified) for: virtual machine administrator; datacenter administrator; virtual machine power user; virtual machine user; and resource pool administrator – custom roles may also be defined for VC or ESX Server:
  - No access (system role) – cannot view or change the object (default for all non-Administrators).
  - Read-only (system role) – view state and details – view all except console – cannot perform any actions.
  - Administrator (system role) – all privileges for all objects – can add, remove and set access rights/privileges for all users and objects (default for members of the Administrators group).
  - Virtual Machine User – perform actions on VMs only – interact with VMs but cannot change configuration (including scheduled tasks and some global items but no privileges for folders, datacenters, datastores, networks, hosts, resources, alarms, sessions, performance and permissions).
  - Virtual Machine Power User – perform actions on VMs and resources only – interact with VMs and change most configuration items, take snapshots and schedule tasks (including scheduled tasks and some global items but no privileges for folders, datacenters, datastores, networks, hosts, alarms, sessions, performance and permissions).
  - Resource Pool Administrator – perform actions on datastores, hosts, VMs, resources and alarms – resource

delegation for resource pool objects (including folders, VMs, alarms and scheduled tasks, some global items, datastore, resources and permissions but no proviledges for datacenter, network, host, sessions or performance).
  - Datacenter Administrator – perform actions on global items, folders, datacenters, datastores, hosts, VMs, resources and alarms – set up datacenters but limited ability to interact with VMs (including folders, datacenters, datstores, networks, resources, alarms and scheduled tasks, some global items, hosts and VMs but no session, performance and permissions).
  - Virtual Machine Administrator – perform actions on global items, folders, datacenters, datastores, hosts, VMs, resources, alarms, and sessions but cannot change permissions.
- Roles may be cloned for editing.
- A role may be propagated down to child objects – permissions may be overridden at a lower level by adding a new permission for the same user:

```
Container 1
   -> Item 1          Inherit from container 1
   -> Item 2          Inherit from container 1
   -> Container 2     Not inherited from container 1 unless propagation is enabled
      -> Item 3       Inherit from container 2
```

- Datacenters, folders, clusters and resource pools are all containers – VM is an item (items have no children and therefore do not propagate).
- For VC, the local Administrators group is assigned the Administrator role at the topmost level in the inventory – if the VC server is a domain computer then this means that Domain Admins all have full administrator rights over VC.
- For ESX Server, the service console users vpxuser and root are assigned the administrator role at the host level in the inventory – vpxuser is an account used by VC to identify itself when sending tasks to the ESX Server – root performs the requested tasks.
- VC users only see UI elements that they have rights to.
- Web Access is a browser interface for managing VMs (not ESX or VC) – no need to install the VI client and can use local floppy/CD-ROM – runs under Apache Tomcat on either VC or ESX and requires a plug-in within the browser – supported with IE6 and Firefox 1.0.8 for Linux or Windows – if logging on to ESX, use ESX credentials, for VC, use Windows credentials.
- Can disable web access post installation for ESX or during installation for VC.
- Cannot create new VMs with web access but can display list of VMs, access console, view status, power on/off and edit the VM configuration – list of VMs presented depends on whether the connection is to VC (all VMs) or ESX server (that host's VMs).
- It is possible to generate a remote console URL (and to limit the view) – not a replacement for access control as user can edit the URL and bypass the customisation.

VM resource management

- CPU resource:
  - Limit – cap on CPU time in MHz (default=unlimited).
  - Reservation – number of CPU cycles reserved for VM in MHz (default=0) – VMkernel chooses CPU and may migrate – if not used, then available to others – VM will only start if reservation can be guaranteed.
  - Shares used to compete for CPU time between reservation and limit (default=1000 per vCPU).
- All vCPUs must be scheduled simultaneously (so 1000MHz reservation may be generous for a single vCPU, but not for 4).]
- Memory resource:
  - Limit – cap on memory in MB (default=unlimited).
  - Reservation – memory reserved for VM in MB (default=0) – never donated to another VM.
  - Shares used to compete for memory between reservation and limit (default=10 per MB).
- VMkernel allocates a per-VM swap file to cover range between limit and reservation – used as a last resort if RAM is unavailable.
- Shares only apply when competing for CPU or memory resource – guarantee a certain fraction of the available resource (as a minimum) – shares can be added to a VM whilst running, allowing increased resource access but also affected by VMs being powered on/off.
- CPU affinity may cause issues by avoiding the VMkernel scheduler's ability to automatically balance load or ESX host's ability to meet resource requirements – in addition, CPU admission control does not consider affinity – VM with manual affinity set may not get all of its reserved resources – affinity may no longer apply after move VM between hosts with a different number of CPUs – NUMA scheduler may not be able to manage a VM with manual

affinity.
- HT cannot be enabled on a system with >16 pCPUs because there is a limit of 32 lCPUs – otherwise, HT enabled by default (if visible in BIOS).
- lCPUs on the same core have adjacent numbers.
- Recommended to set a limit when starting out with a small number of VMs in order to maintain user expectations; similarly it may be advisable to set a reservation to ensure that a VM receives sufficient resources.
- Resource Pools allow resources to be allocated to VMs, other resource pools or users – can have associated access control and permissions – control aggregate CPU and memory of the compute resource (either host or DRS cluster) – reservation, limit and shares just as for individual VMs – expandable reservations may be set for resource pools (not VMs) allowing VMs and sub-pools to draw resources from the parent pool if available – creating a child pool reserves resources from the parent, irrespective of whether any VMs are powered on.
- Admission control is used for any action that changes a VM or resource pool's reservation (i.e. power on a VM, create a new sub-pool with its own reservation or change a pool's reservation) – check that sufficient capacity available to satisfy the reservation before carrying out the action.
- VMotion allows live migration of VMs between hosts – entire VM state (memory content – including transaction data, operating system and applications – and configuration information to identify the VM – such as data to map the virtual hardware) moves whilst data remains in the same location on the SAN:
  1. Virtual disk and configuration held on SAN – dedicated VMotion network used to pre-copy memory from host to host.
  2. User access continues as normal logging ongoing changes into a memory bitmap on the first host (not the contents, just a list of pages).
  3. Once copied, VM is quiesced (taken to a state where no additional activity will occur – i.e. frozen/paused with file lock released) on the first host and the memory bitmap copied to the second host (very short break in service).
  4. VM is available on second host and ARP request notifies the network that the MAC address is now found at a different switch port.
  5. Memory pages from the bitmap are transferred, with priority given to any that are accessed on the second host during the transfer.
  6. To guard against failure, the source VM remains in place until the target is successfully running – only at the end of a successful migration is the VM deregistered from the first host.
  - Error if:
    - VM has an active connection with an internal switch.
    - VM has an active connection to a local CD-ROM or floppy drive or image.
    - VM has CPU affinity set to run on one or more specific pCPUs.
    - VM is in a cluster relationship (using MSCS) with another VM.
  - Warning if:
    - VM configured to use an internal switch but not active (connected) on it.
    - VM configured to access a local CD-ROM or floppy drive or image but is not connected to it.
  - Source and destination hosts must have:
    - Visibility of all SAN LUNs (FC or iSCSI) and NAS devices used by the VM.
    - Gigabit Ethernet backplane.
    - Access to the same physical networks (i.e. have networks in common – not necessarily identical – i.e. one of both could have access to additional networks not used by the VM).
    - Consistently labelled vswitch port groups.
    - Compatible CPUs (i.e. same manufacturer and instruction set) – some instructions can be hidden from VM to aid VMotion compatibility (e.g. NX/XD) and CPU mask can be edited if required – VMware provides a CPU compatibility tool.
- Topology maps are useful to check if all hosts have access to the required resources for VMotion and other services such as HA and DRS – if VM is shown on a green background then can be migrated using VMotion – the map relationship window can be used to customise the map.
- DRS allows several hosts to be placed into one pool – a DRS cluster is implicitly a resource pool which may be divided into sub-pools – all the resources for a VM are only owned by a single host at any one time – load is balanced across all hosts in the cluster, considering resource policies, affinity and anti-affinity rules.
- Turning on DRS disables CPU affinity.
- DRS is CPU/memory based and ignores disk/network.
- Initial placement provides a recommendation for the host on which a VM should be placed – for dynamic balancing, DRS monitors key metrics and resource policies to determine the appropriate resource allocations.
- After creating a DRS cluster, automation level needs to be defined – manual (DRS displays recommendations),

partially automated (DRS automatically places the VM at power-on, but displays recommendations for virtual machine migration), or fully automated (DRS migrates VMs to ensure a balanced use of cluster resources) – migration thresholds can also be set, with 5 levels from conservative through to aggressive – automation level can also be overridden at a per-VM level.
- In practice it is the lightly-loaded VMs that move, not the ones causing the problems.
- Affinity rules are used to ensure that certain VMs are placed together (e.g. for performance) and anti-affinity rules will separate VMs (e.g. two servers that share the load on a particular service.
- It is recommended to follow all DRS recommendations with 4 or 5 stars to avoid deterioration in service – it may be most effective to use some automation with manual or partially-automated migration for key VMs (overriding the cluster defaults).
- Tiered resource pools can be used to delegate administrative responsibility – e.g. a datacenter administrator for the cluster (root resource pool), resource administrators for the next level of resources pools and a virtual machine power user for the sub-pool.
- Use expandable reservations for administrators within an IT group (i.e. for top-level resource pools to use resources from elsewhere within the DRS cluster) – do not use expandable reservations for customers who pay for a specific amount of compute resource (else set expectations to high on a lightly-loaded server).
- VC reports the state of a cluster within the VI client – valid (all resource constraints satisfied), no colour; yellow (some resource constraints not satisfied), e.g. a host is down; red (cluster is internally inconsistent), e.g. a DRS or HA violation (DRS overcommitted due to host failure; DRS invalid if VC unavailable and VMs powered on by directly accessing the ESX host; HA invalid if failover capacity is lower than configured failover capacity or if hosts are not responding; DRS or HA invalid if user reduces reservation in parent resource pool whilst VM is failing over).
- When add a new host or moving an existing host into the cluster, resource pool placement can be decided (by default, existing resource pools on the host will be discarded but they can be grafted onto the cluster's hierarchy.
- Before removing a host from a DRS cluster it must be placed into maintenance mode – VMs will continue to run but no VMs will be migrated to the host (and no VMs can be powered on) – once all VMs have been shut down (manually) or migrated away (VMotion), the machine will enter maintenance mode – note that a fully-automated DRS cluster migrates VMs to different hosts as soon as maintenance mode is enabled (a partially-automated DRS cluster will display recommendations).
- Resource management best practice:
  - If frequent changes to available resource are expected, use shares, not reservation.
  - Use reservation for minimum acceptable memory (not desirable amount of memory) – setting reservation to high will limit the number of running VMs.
  - Always leave some room in the system – do not reserve all resources (otherwise difficult to make changes without violating admission control).
  - Use resource pools for delegated resource management – make the pool fixed and apply reservation and limit to fully isolate the pool.
  - Group VMs for a multi-tier service in a resource pool – allows ESX to assign resources for service as a whole.

Monitoring

- Optimising resource use:
  - CPU load balancing:
    - VM can have up to 4 vCPUs (with a Virtual SMP license) – guest must support SMP too – 2 or 4 vCPU VMs use 2 or 4 pCPU cores at a time (or none).
    - CPU failure will cause a system crash (VMware assumes correctly functioning hardware).
    - VMkernel dynamically schedules VMs and SC for CPU time – looks for VMs to migrate every 20ms but SC always runs on CPU core 0.
    - HyperThreading (HT) is an Intel technology to provide 2 logical CPUs (lCPUs); however it is not dual-core and simply allows idle cycles to be used by another thread – if VMkernel detects 2 CPU-intensive VMs using lCPUs from the same pCPU it will migrate one off.
    - NUMA architecture (e.g. AMD Opteron) has memory controllers within each pCPU – ESX server optimises itself for processor and memory affinity on NUMA hosts (whilst each CPU can access another's RAM, there is a performance hit in doing so – should be avoided).
  - Transparent memory page sharing:
    - VMkernel detects identical pages in VM memory an maps them to the same page in physical memory – no change to guest OS required (up to 30% memory shared with an idle Windows 2000 VM) – shared pages are copied on write so read-only whilst shared and private after write.

    - Note that when Windows VMs are started, the OS writes zeros across memory so memory utiisation will rise and status will show as red/yellow before dropping off to green later when transparent memory page sharing takes effect.
    - Page sharing is always active unless disabled by the administrator.
    - Performance implications of transparent page mapping:
        - None for regular guest memory access.
        - Additional time to map memory (e.g. when setting up mappings or context switching between address spaces).
        - Memory virtualisation overhead depends on workload.
  - vmmemctl (balloon-driver mechanism):
    - Used to deallocate memory from selected VMs when RAM is scarce.
    - When ample memory, balloon remains deflated, then inflate balloon as driver demands memory from guest OS – guest is forced to page out to it's own paging file, VMkernel reclaims memory – deflate balloon as driver relinquishes memory and guest may page in.
    - Advantage is that guest OS chooses the pages to swap out, not VMkernel.
  - VMkernel swap file:
    - A last resort (e.g. if the balloon driver cannot free enough memory), VMkernel will copy pages to the VMkernel swap file.
    - Size determined at power-on as difference between memory reservation and limit – stored in same location as the VM's boot disk.
    - Noticeable performance degradation – configure servers so that normal memory requirements can be accommodated using physical memory.
- VM owner can affect performance with Virtual SMP, maximum RAM size, LUN placement, vswitch and NICs (e.g. teamed).
- Administrator can affect performance with limits, reservations, share allocation, affinity and traffic shaping.
- When tuning, take a logical, step-by-step approach as one small change affects many VMs – a well tuned server will provide maximum performance to high-priority VMs (possibly at the expense of low-priority VMs) – benchmark before and after change.
- Performance chart can be torn off (into a separate window) or exported to Excel.
- Improved VM's CPU and memory performance – modify VM's limits and reservations – modify resource pool's limits and reservations – add capacity to DRS cluster.
- CPU-constrained VMs have high CPU ready time (i.e. ready to use CPU but waiting to be scheduled) and high CPU usage within Task Manager in the VM – CPU ready time is calculated per vCPU and only visible in real-time graphs.
- Memory-constrained VMs have high levels of ballooning activity (as shown in the real-time graphs with the memory balloon counters) and will show a high number of page faults within the VM's Task Manager – if CPU Ready is 5% greater than CPU used then there is an issue (10% represents a big issue) – if evidence of ballooning but VM memory usage is not high then this is by design (idle taxing); however ballooning plus high memory usage and page faults within the VM indicate an issue (use of VMkernel swap file is a big issue).
- Disk-constrained VMs will have low disk read/write rates – get an idea of what is possible by running a tool such as IO Meter out of hours and compare actual disk bandwidth with performance graphs – improve performance by moving VM to another LUN, changing path to the LUN or changing the RAID level of the LUN.
- Network-constrained VMs may be affected by issues outside ESX Server (e.g. WAN) – rule out WAN by measuring effective bandwidth between VM and peer on an internal vswitch (i.e. at bus speed) – Network Usage in performance chart and IO Meter – problems here indicate a VMkernel issue – VMkernel cannot offload network processing to NIC for multiple VMs so if CPU constrained, will also affect network.
- Alarms can be used to indicate status-based notifications – within VI client and onwards via SNMP or SMTP – status indicated in inventory and also in list of running VMs within the VI client – create alarm (VM or host-based) using on triggers – use reporting to limit multiple alarms (tolerance or frequency) – configure actions (e.g. send a notification e-mail, run a script, or power off a VM – for VM alarms only).
- Default alarms at Hosts and Clusters level in the inventory – can define new alarms at any level (e.g. add a folder and define alarms on that folder) – SNMP and SMTP settings must be defined for VC under Administration, Server Settings in the VI client – MIBs for ESX and VC available for many third party management tools (e.g. VMM for HP SIM).
- Generate quick trap to test by setting alarm for VM state changing to powered on.

Data availability and protection

- Backup strategies for VMs:

  - File level backup using backup agent within VM.
  - Windows VM file-level backup using VCB.
  - Any OS full VM backup using VCB.
- VMware recommends storing application data on separate disks to system images – use file-level backups for application data and full VM backups for system images (or redeploy from template).

VCB

- LAN-free, online backup – offloaded to Windows Server 2003 SP1 backup proxy over FC SAN network – file system consistent – works with major 3rd party backup software.
- Requires VCB Framework (vcbMounter and vLUN driver) + supported backup software + integration module.
- Turn off automatic mounting of disks for Windows servers (Windows Server 2003 Enterprise does not automount disks).
- File level backup:
  1. Backup software runs the pre-backup script from the VCB integration module.
  2. Pre-backup script asks hostd to quiesce VM's file system.
  3. hostd asks VMware Tools to run the pre-freeze script and quiesce the Windows file system on the vdisk.
  4. VMware Tools tells Windows to hold all writes to the file system.
  5. hostd takes a snapshot of the VM and informs VMware Tools when complete.
  6. VMware Tools tells Windows to allow queued writes through (diverted to a disk-write buffer).
  7. hostd tells the pre-backup script that quiescing is done.
  8. Pre-backup script calls vcbMounter (part of VCB framework) to mount the vdisk.
  9. vcbMounter mounts the .vmdk at the defined junction point.
  10. Backup software backs up the files via the junction point.
  11. Backup software runs the post-backup script from the VCB integration module.
  12. Post-backup script tells vcbMounter to unmount the vdisk and tells hostd to flush the disk write buffer.
  13. Once hostd has flushed the disk write buffer it removes the snapshot and tells the post-backup script that it is done.
  - Pros:
    - LAN-free online backups of vdisks.
    - No backup client required per-VM.
    - Possible to backup to FC-attached tape drives.
    - Integrated with backup solutions.
    - No resource contention on ESX Server host (as backup is proxied).
    - Better performance – shorter backup window.
  - Cons:
    - Requires the use of separate physical server running Windows (the backup proxy).
    - Cannot restore directly into the vdisk (would result in corruption).
    - Differential/incremental backups not available as setting archive bit would involve writing to the vdisk.
    - FC-only (no iSCSI/NAS).
  - Restoration – trade-off backup agent requirements vs. ease of restoration:
    - Centralised – restore to separate disk area on proxy and copy files across the network (high degree of administrative intervention) – proxy server used for backups and restores.
    - Per-group – 1 VM with a backup agent in each logical group of VMs – restore there and then move files within VI – backup agent only used for restoration (backup on proxy).
    - Self-service – each VM has own backup agent (expensive) – only used for restore (backup on proxy).
- Full VM backup:
  - vcbMounter and vcbRestore (independent of backup software).
  - vcbMounter needs to know host (or VC) IP address, username, password, address or name of VM, type of backup (full VM or file), path to backup location (cannot already exist) and media type (e.g. SAN).
  - vcbRestore removed from final release but on ESX server – use Samba client to access data on backup proxy (open firewall) – just needs to know where the media is mounted and will restore to the original ESX Server.
- One approach may be to VCB to staging area and stream to tape at leisure.
- Can also use vcbMounter to mount a full VM backup and extract a file (Windows or Linux versions).
- If VM is powered off then VCB can perform full VM backup of any OS.

High availability

- High availability can be achieved using clustering VMs (MSCS or similar) or VMware HA (optional VC feature).
- VMware HA – provide HA to VMs – failover within cluster of ESX Servers – if host fails, another host detects

this and will bring VM up from shared storage – some downtime, but minimal (assuming VM recovers from crash).
- HA does not manage individual VM failure – that can be addressed with an alarm.
- HA prerequisites include – access to common resources (VM can be powered on from any host in the cluster) and host should be configured for DNS (HA is very reliant on DNS – or hosts files).
- HA advantages include minimal setup, reduce hardware cost and increased application availability; however limitations include loss or run-time state and longer application downtime than a traditional cluster – it may be appropriate to use a combination of HA and traditional clustering.
- Set up networking to avoid single points of failure – two network paths for heartbeats: 2 SC ports on different vswitches; or single SC port with NIC teaming on the vswitch.
- HA is enabled within the properties for the cluster (as for DRS) – use both as HA is reactive and DRS is proactive – unlike DRS, HA does not use VMotion – it is based on Legato AAM.
- Configuration settings: number of host failures allowed (1-4); admission control (i.e. will a VM be powered on if it will affect resource reservations for other VMs – which is more important: uptime or resource fairness?); restart priority for each VM (set to disabled if HA should not be used); isolation response (what to do if isolated – power off VM will release lock on disks).
- Add host to cluster via drag and drop within VI client or right-click cluster and select add host.
- HA architecture – VC agent (vpxa) communicates with VMs and VMAP (list of which servers are running which VMs) – AAM (HA agent) monitors AAM on other hosts (heartbeat via SC network) – to avoid split brain when nodes become isolated (e.g. network failure), each node has an isolation address that it pings to see if it is isolated or the others (can be set as the advanced option das.isolationaddress – cluster-wide setting) – waits 15 seconds before deciding that host is isolated (non-configurable) but will start to power off VMs after 12 seconds (if network connectivity restored before 15 seconds, then VM powered off but not failed over) – will only ping if it can't see another node.
- Troubleshooting HA – look at IP connectivity, DNS (host files can help as a reserve) – ensure that storage and networks are visible throughout the cluster – ensure that host has not been managed directly, bypassing VC (out of date VMAPs), which will also affect DRS clusters – check logs at /opt/LGTOaam512/log/* and /opt/LGTOaam512/vmsupport/*.
- Turn off HA temporarily by putting ESX into maintenance mode.
- Turn off DRS temporarily by making it partially automated.

Clustering

- Computers may be clustered to improved reliability, scalability, or both.
- Applications need to be cluster-aware – examples include stateless web applications.
- Applications with built-in recovery features, such as database, e-mail and file servers.
- Although other clustering arrangements are available, VMware only supports the Microsoft Cluster Service (MSCS), which provides failover support for database, e-mail and file servers – limited to 2-node active/passive (even though MSCS has greater capabilities).
- VMware does not test Microsoft Network Load Balancing (NLB).
- Typical setup – shared disk (i.e. FC-connected SAN) and extra network connectivity between nodes for monitoring heartbeat status.
- Hot clustering possible using MSCS – cold standby with VMware HA and VC.
- Can cluster 2 VMs on the same ESX Server (cluster in a box) using local or remote storage; however no protection against hardware failure – useful for testing cross-host clustering before distributing the VMs across hosts.
- Cross-host clustering provides protection against hardware failure (except within the shared storage). By using ESX with MSCS it is possible to consolidate many nodes onto a few servers.
- ESX can also operate as a standby host – with one VM representing each physical server – if the physical server fails then the VM on the standby host can take over.
- Hardware requirements:
  - Cluster in a box – ESX host with 2 pNICs (one for SC and one for client communications), SAN storage (recommended) or local disks and a local SCSI controller – virtual disks or remote LUNs can be used (using RDM in virtual-compatibility mode – non-passthrough RDM) – VMs require 2 vNICs and at least 2 disks (quorum and data).
  - Cross-host clustering as for cluster in a box but 3 pNICs required (1 SC and 2 for cluster), shared storage must be on a SAN and virtual disks cannot be used – RDM can be used in physical or virtual compatibility mode (pass-through or non-pass-through RDM) but physical (pass-through) is recommended.
  - Standby host (N+1) clustering as for cross-host but RDM must be in physical compatibility mode (non-pass-through RDM), multipathing software cannot be used on the physical machine and the FC HBA on the physical server must use the SCSIport driver (not STORport).

- Restrictions:
  - Each VM has 5 available PCI slots by default – with 2 NICs and 2 SCSI controllers that leaves 1 spare.
  - VMs can only use SCSI-2 (not SCSI-3) for disk reservation.
  - 2GB HBAs must be used with a supported driver.
  - The LSILogic virtual SCSI adapter must be used.
  - VMXNet must be used (not vlance).
  - VMs can only be 32-bit.
  - Only 2 node clusters are supported.
  - iSCSI clustering is not supported on iSCSI or NFS disks.
  - STORport driver is not supported (only SCSI miniport driver).
  - NIC teaming is not supported.
  - VM boot disk must be on local storage.
  - Boot from SAN for the ESX host is not supported.
  - Mixed HBAs are not supported on the same host.
  - Mixed ESX Server environments (2.5 and 3) are not supported.
  - Clustered VMs cannot use VMotion, DRS, or HA.
- Potential issues after setup include:
  - Disk I/O timeout (should be set to 60 seconds).
  - Cloning of RDM disks will result in virtual disks – remove RDMs before clone and remap after.
- Other advice:
  - Create all disks before set up networking – if add a disk later, MSCS will consolidate IP settings and reconfiguration will be required.
  - QLogic HBAs must be set to enable target reset and full LIP login but not full LIP reset.
- Setup cluster in a box:
  - Create node 1 with 2 vNICs and boot disk on local storage – install guest OS and power off the VM.
  - Clone node 1 to create node 2.
  - Add 2 remote disks on a separate SCSI controller (not SCSI 0) and configure guest's public and private IP addresses – disks must be zeroed using vmkfstools in preparation for use (unless using RDM in virtual compatibility mode) – bus sharing must be virtual and the LsiLogic controller must be used.
  - Install (Windows 2000) and configure MSCS (Windows Server 2003 includes MSCS)
- Setup cross-host cluster:
  - Create node 1 with 2 vNICs and boot disk on local storage – install guest OS and power off the VM.
  - Clone node 1 to the second host server to create node 2.
  - Add 2 remote disks on node 1 for quorum and shared storage using a separate SCSI controller (not SCSI 0) – bus sharing must be physical and the LsiLogic controller must be used.
  - Configure guest OS public and private IP addresses on node 1.
  - Configure disks and guest OS public and private IP addresses on node 2.
  - Install (Windows 2000) and configure MSCS (Windows Server 2003 includes MSCS)
- Setup standby host clustering:
  - Install operating system on node 1 (physical), ensuring that it has at least 2 NICs, access to the same SAN as the ESX Server and does not have multipathing software installed.
  - Create node 2 with 2 vNICs – install guest OS.
  - Install (Windows 2000) and configure MSCS (Windows Server 2003 includes MSCS) – disable storage validation heuristics to avoid issues with the same LUN presenting different IDs on different computers.
  - Create additional physical/virtual pairs as required.
- Upgrading clusters is only supported from v2.5.2 to v3.0.
- Upgrade cluster in a box:
  - Power off VMs.
  - Upgrade ESX hosts.
  - Upgrade VMFS2 storage holding .vmdk files to VMFS3.
  - If necessary, repeat for volume with RDM files.
  - Upgrade virtual hardware for each VM – if the RDMs and boot disks are on the same volume then this will produce an error for the second VM which can safely be ignored; however it will be necessary to edit the .vmx file to point to the relocated RDM files that are located with the first node.
  - Power on VMs and verify cluster setup – if there are problems, import the virtual disk using vmkfstools and edit the .vmx file to point to the new location.
- Upgrade cross-host cluster:
  - Use vmkfstools to change shared VMFS2 volume from shared to public.
  - Upgrade the ESX hosts.

  - Upgrade the VMFS2 volume with the cluster .vmdk files to VMFS3.
  - Create LUNs for each shared disk, then create RDMs and import the virtual disks using vmkfstools.
  - Edit the .vmx file to point to the RDM.
  - Power on the nodes and verify that the cluster service starts.
- Upgrade standby host clustering, the public disk is mapped using RDM and the upgrade process converts VMFS2 disks to VMFS3 (or they can be explicitly upgraded later).

Troubleshooting

- VM issues: not enough resources; guest OS or application failures; misconfigurations (under sizing virtual resources, VMotion requirements not satisfied, insufficient resources in resource pool or HA cluster, not using FQDNs for ESX Servers) – VI client error messages in tasks and events (can be a bit vague); virtual machine monitor (VMM) may panic and create core dump on same volume as .vmx file.
- Methodology: identify specific failure (capture error messages); know what failure depends on; avoid changing the system before identify underlying fault (try tests that will rule out the most common or largest number of underlying faults); record changes before making them (make changes one at a time); confirm that the problem is fixed.
- Cannot start a VM:
  - Insufficient resources (memory or CPU)? Check resource allocation for cluster, resource pool and VM.
  - Insufficient privileges? Check user permissions on the VM.
  - Insufficient resources to satisfy HA? Check resource allocation of the cluster and each host within the cluster.
- VM blue screens or hangs:
  - Problems in guest OS or application? Reproduce on physical hardware (V2P software may help).
  - Physical failures (e.g. bad RAM)? Run hardware diagnostics.
  - Bugs in VMware software? Open a support case.
- Poor application performance:
  - User confusion between console performance and application performance? Benchmark the application – try RDP connection to VM (more direct than workstation -> VC -> SC -> VMkernel -> VM).
  - Environmental issues? Rule out software misconfigurations in the guest OS (e.g. wrong HAL, wrong DNS servers); also check speed and duplex of vswitch and pswitch.
  - Key application is limited by a resource? Check for saturated CPU, disk, memory and network bandwidth.
- Cannot log in using VI client:
  - "No connection could be made because the target machine actively refused it" – VC Server service (vpxd) not running (or hostd failure on ESX Server – check by running ps -ef | grep hostd) – possibly caused by issues accessing database (authentication?) – restart service.
  - "Logon failed due to a bad username or password" – incorrect username or password (Windows credentials for VC and ESX for the host).
  - "Connection Failed" – ESX server management agent (hostd) not running.
  - "A connection attempt failed because the connected party did not properly respond after a period of time, or established connection failed because connected host has failed to respond" – IP connectivity issues – check address/host name – try pinging the VC or ESX Server.
- Cannot add host to inventory:
  - "Unable to access the specified host. It either does not exist, the server software is not responding, or there is a network problem" – Incorrect host name or IP address? Try pinging it.
  - "Login failed due to a bad username or password" – Incorrect user name or password? Try logging on to ESX Server from the command line.
  - "Unable to access the specified host. It either does not exist, the server software is not responding, or there is a network problem" – ESX Server management agent (hostd) is not running? Run ps -ef | grep hostd at the command line. – if hostd PID keeps changing then restarting in a loop – try service mgmt-vmware restart (service mgmt-vpxa would be the agent).
- VI client shows server as not responding – check service console network connectivity.
- VI clients reports "connection to the server has been lost. The application will not exit" – VC Server service may have stopped.
- ESX Server problems generally caused by: hardware problems (test for 72 hours before deployment, install a dummy OS to check out hardware, ensure shown on hardware compatibility guide); misconfigurations or inadequate planning.
- ESX Server will display PSOD if it cannot continue without data loss (VMkernel panic) – most commonly: general hardware problem; NMI ECC or parity error – VMware support can help to pinpoint the failing subsystem or

memory bank based on contents of core dump – copy screen display (grab via iLO or KVM, take a photo, copy down contents manually) in case core dump not generated – check: environmental factors (e.g. humidity); detached devices; recent hardware changes – use Vi client to export diagnostics data (option to include VC information too) or run vm-support from SC command line.

- If VC Server service fails to start, check VC Server service (vpxd) logs for clues (after get it going or in %temp%\vpx), use Windows Event Viewer to identify problems – possibly caused by database problems (check authentication, or for full transaction logs).
- Enable host services for remote access – inbound and outbound connections blocked by default – connections required for VI client are opened by the ESX installer during installation – can also open services in ESX Server security profile within VI client (open individual ports from the SC CLI using esxcfg-firewall).
- SSH access to SC as root is not allowed by default – either login as another user and su – or edit /etc/ssh/sshd_config and change PermitRootLogin entry from no to yes, save the file and restart the service (service sshd restart) – alternatively access the console via iLO or similar.
- SC CLI allows access to esxcfg-* commands:
  - Networking: esxcfg-firewall; esxcfg-route; esxcfg-vmknic; esxcfg-vswif; esxcfg-vswitch; esxcfg-nics.
  - Storage: esxcfg-dumppart; esxcfg-mpath; esxcfg-nas; esxcfg-swiscsi; esxcfg-vmhbadevs.
  - Misc: esxcfg-info; esxcfg-advcfg; esxcfg-resgrp.
- Swap files may remain after an ESX Server failure – to remove them start the VM and stop it explicitly.
- esxtop command can be used to examine how resources are used on real time.

ESX and VC upgrades

- Why upgrade?
  - Larger VMs and better performance (need ESX3, VMtools3 and new VM hardware).
  - Resource pools and DRS (VC2 and ESX3 required).
  - HA (ESX3, VC2 and VMFS3 required).
  - Improved manageability (with VC2).
- Mixed environment:
  - Manage ESX2 with VC2? Yes, but no access to new features.
  - Manage ESX3 with VC1? No.
  - VMotion from ESX2 to ESX3? No.
  - Upgrade VM on ESX3 and boot on ESX2? No.
  - Store ESX2 and ESX3 VMs on the same VMFS? No.
- Each step of upgrade is irreversible – order of steps is significant – downtime is required whilst VMFS is upgraded – make backups first and treat process as a project (plan):
  1. Make backups:
     - VC database – using native techniques – if Access, export for later import into MSDE/SQL Server/Oracle.
     - VM configurations – offline copy of .vmx files.
     - vdisks (.vmdk or .dsk) – using archive software or SAN snapshots – don't allow ESX2 to see VMFS volume and its clone.
     - ESX system configuration – SC file-level backups or vm-support.
     - Use clean backups – not crash-consistent backups created with vmsnap.pl.
     - Local images, templates, exported VMs and .iso files.
     - (an alternative to backing up VMs and their vdisks is to clone them; however this creates a new UUID and so is not an exact copy of the VM.
  2. Confirm ESX is upgradeable – run the pre-upgrade script (mount /mnt/cdrom perl /mnt/cdrom/scripts/preupgrade.pl) – alerts of any problems (e.g. 1000MB available disk space on SC, 1200MB available disk space on VMFS, VMFS2).
  3. Choose a strategy – in-place or migration – look for common downtime windows – may be less disruptive to migrate; however in-place is simpler if the downtime can be tolerated.
  4. Prepare for changes – migrate clustered VMs using vdisks to raw LUNs (using Symantec Ghost or similar) and connect VMs to raw LUNs using RDMs (ESX3 does not support cluster across boxes with vdisks) – commit or discard all redo logs (ESX 2 redo logs are unusable with ESX3) – check whether hardware supports NX/XD (ESX2 hid this, ESX3 does not by default – will affect VMotion).
  5. Upgrade VC components:
     - Install license server and add ESX3/VC2 licenses (unless not using VC – host based licensing is available for ESX).
     - VC2 detects previous releases and uninstalls them – no reboot required (note that VC1 used TCP port 905 – VC2 uses 902) – upgrade requires VC to be unavailable for approximately 10 minutes.

    - VC1.0 and 1.1 must be upgraded to 1.2 if the database is to be preserved.
    - MDAC2.6 is required on Windows 2000 SP4 servers.
    - VC2 cannot co-exist with a web server, GSX or VMware Server due to conflicts on ports 80/443/902.
  - Upgrade VC database – in place as part of upgrade except Access (not supported for VC2 – import old data into new database, or repopulate as only suitable for test/demonstration).
  - Install VI client – use to access VC2 – can co-exist with VC client v1.x – VI client can access VC or ESX (ESX management user interface, known as the MUI, no longer exists in ESX3) – VC2 datacenters replace VC1 farms as organisational and VMotion boundaries.
6. Upgrade ESX hosts and data stores:
  - ESX upgrade requires a reboot – upgrade all ESX servers before upgrade VMFS – ESX2 does not support VMFS3 – some ESX releases are not supported for upgrade (supported releases are 2.1.1-2.1.3/2.2/2.5.1-2.5.3).
    - NFS mounts are lost during the upgrade – they can be restored from /etc/fstab.save.
    - VMs with SCSI passthrough will need to be re-attached to their disks (renumbered during upgrade).
  - VMFS upgrade requires VM downtime (all vdisks powered off) and 1200MB free space – VMFS3 includes support for subdirectories – necessary for shared storage (e.g. for HA) – VMFS2 is read only with ESX3 and VMs cannot be powered on – may be less disruptive to migrate (build a new VMFS3, copy ESX2 VMs to VMFS3 when suitable to shut down, then power on using ESX3, when all complete, remove/replace empty VMFS2) – VMFS conversion takes approximately 15 minutes – performed within VI client.
    - ESX3/VC2 templates are VMs marked not to be powered on – ESX2/VC1 templates were exported vdisks stored on VMFS2 or NTFS volume on the VC server – upgrade in VI client.
    - Also need to consider .iso images that may be on the VMFS.
    - VM disks can be relocated in the VI client.
    - Workstation/GSX disks need to be imported using vmkfstools as they are not supported with ESX3.
7. Upgrade VM hardware:
  - In-place upgrades of VMFS will upgrade VMs; however manual VMFS upgrades necessitate manual VM upgrades (can be performed in VI client).
  - Upgrade to VM3 takes approximately 30 minutes and supports: snapshot of VM state (disk and memory, as for VMware Workstation 5); 16GB RAM and 4 vCPUs per VM; improved network performance (especially for Windows Server 2003); hot-plug SCSI (where supported by OS).
  - Can upgrade multiple VMs and VMtools simultaneously – VMs must be on VMFS3 and must be powered off (upgrade will power them on and off briefly) – use vmware-vmupgrade.exe from Windows with username and options for password, vmname (or all VMs on host), pot number (if not 902), maximum number of simultaneous conversions, skip tools and quiet mode – only suitable for Windows 2000/XP/2003 and Linux VMs.
  - VM2s can run on ESX3 as long as moved to VMFS3 – no snapshotting, no hot-plug SCSI, limited to 3.6GB RAM and 2 vCPUs.
8. Upgrade VMtools inside VMs – at same time as upgrade VM (using vmware-vmupgrade.exe) or from within guest OS – upgrade supports quiescing file system and enhanced network performance (vlance can switch to vmxnet, driver improvements for Windows 2000/2003).