VMware vSphere™ 4

The Best Platform for Building Cloud Infrastructures

vmware

VMware Virtual Disk Development Kit

Sudarsan Piduri, VMware

Outline of the talk

- Introduction
- A bit about physical disks
  - Disk terminology (Geometry, MBR, Partition Table, Boot Sector)
  - Dynamic disks
- Virtual disk basics
  - Parent, Child disks, Disk geometry, VMDK format
- Why VDDK
- VixDiskLib – API
  - Local and Remote disks
  - Transport methods and how they work
  - Snapshots
- VixMntapi – API
  - Mounting on Windows and Linux
  - Pre-requisites on Linux
- References and Links
- Q&A

Hard Disk terminology – A brief recap

- Disk is made of many platters and heads
- Sector – Smallest addressable unit on the disk (512 bytes)
- Geometry – CHS - Cylinders/Heads/Sectors (per track)
  - Cylinder/Head/Sector used to be enough to address any data item on the disk – not any more
- Master Boot Record – first sector of the disk
  - Contains partition table and boot code
  - Holds disk signature on Windows
- MBR based partitioning
  - Legacy, GPT based partitioning is modern
  - Boot sector - first sector of the partition

Partitions and Volumes

- Volume
  - One or more partitions that a file system can own
    - A single file system instance can’t span volumes
  - Simplest – One to one relation between volume and partition
  - Volumes can span partitions
    - Logical Disk Manager (LDM) - Windows
    - Logical Volume Manager (LVM) – Linux
  - Volume may or may not be initialized / formatted
    - A volume needs to be mounted to be accessible
    - Volumes (typically) have unique identifiers
      - Disk Signature + Partition begin offset
      - UUID
      - …

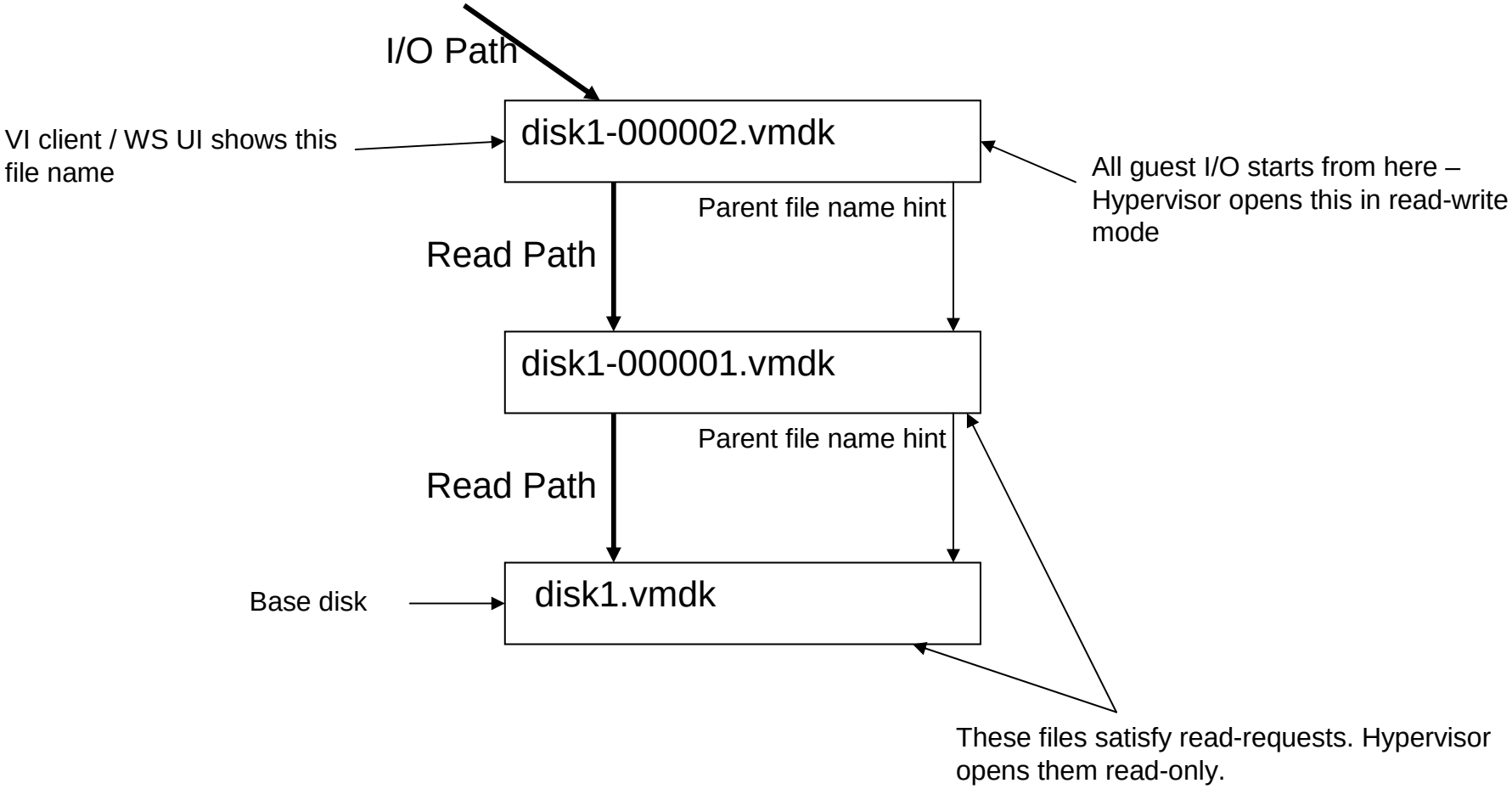
Virtual Disk

- Abstraction that represents a physical disk
  - Stored as a set of files on the host
  - Together with disk emulation code in Hypervisor, implement the disk for the guest
- Virtual Disk
  - Metadata (Descriptor file)
    - Contains Geometry, UUID, Hardware Version, Data filename
    - e.g. WindowsXp.vmdk
  - Disk Data
    - Data corresponding to the disk by sector
    - e.g. WindowsXp-flat.vmdk
  - The above two may be in different files or in the same file.

Virtual Disk Types

- Virtual Disk Types
  - Flat
    - Simplest
    - Fastest access
    - Some metadata (Geometry, Id …)
    - Raw disk sector data
    - File size on host slightly more than the size of disk that it represents
    - Can either be one file (monolithic) or multiple (2GB size each) files to accommodate older file systems (split)
  - Sparse
    - Contain all the sectors that were ever written to this disk
    - Size on host is slightly more than the initialized data
    - Slower access
    - Can be split
  - Raw Disk Mapping (RDM)
    - Use a physical disk as backing
    - Can be in either Physical or Virtual compatibility mode
  - Both Workstation and ESX server support these types
    - Formats slightly different

Parent and Child Disks



Parent and Child Disks

- Child Disk
  - Contains data of the Virtual Disk that has changed since creation of this child disk
  - Also called Copy On Write (COW), Delta, Snapshot, Differencing disk (MS)
  - Contains the filename (+ path) of the parent disk
  - May be split into 2GB segments
- Parent and Child disks form a tree
- Base Disk
  - The root of the tree
- Each disk contains a content ID
  - Used to verify if the parent has changed – to avoid disk corruption

Why VDDK

- VMDK format public but we need an API level access
  - Avoid everyone implementing VMDK writer/reader
  - Avoid updating the writer/reader for ISV's for format updates
  - Optimized code from VMware which has already done all this work
  - Low threshold to let ISV's use VMDK as their native format (e.g. backup)
- Users need a set of utilities to manipulate virtual disks like
  - Creating new virtual disks
  - Mounting a volume / disk

VDDK – Components

- VixDiskLIb
  - C (dynamic) library implementing virtual disk manipulation functions
    - Create / Unlink
    - Read / Write
    - CreateChild
- VixMntapi
  - C (dynamic) library implementing virtual volume mounting functions
    - Parse for volumes
    - Mount / Unmount volume
- VMware-mount
  - Command line tool to mount a volume
- VMware-vdiskmanager
  - Command line tool to create / manage virtual disks

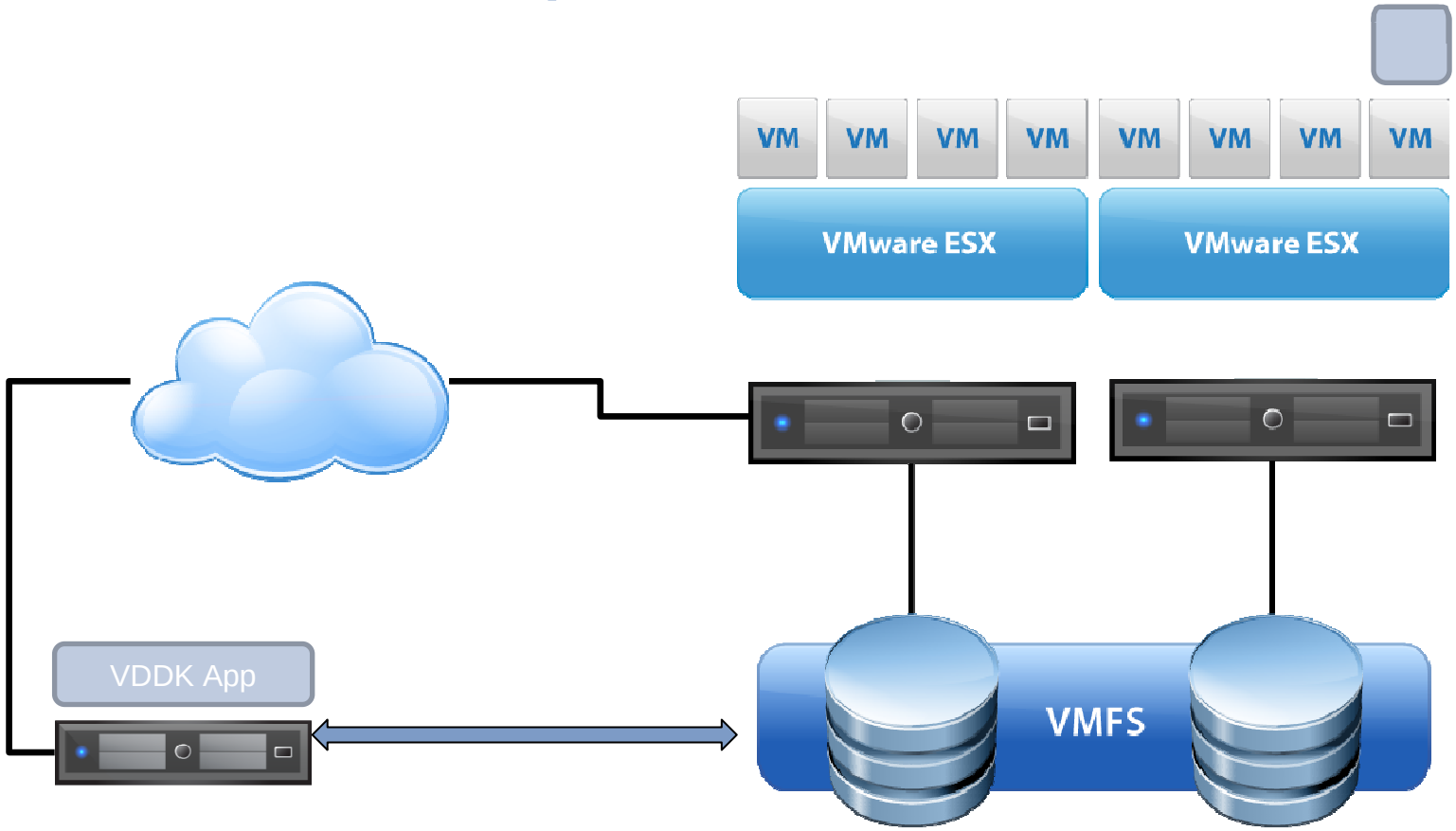
VixDiskLib API descriptions

- Program flow
  - VixDiskLib_InitEx(…)
  - … use VI SDK to get hold of the vm / disk details …
  - VixDiskLib_ConnectEx -> returns a handle to a connection
  - VixDiskLib_Open -> returns a handle to the disk
  - VixDiskLib_Read / Write (handle) -> read / write data of the disk
  - VixDiskLib_ReadMetaData/Write (handle) -> read / write meta data of the disk
  - Close (handle)
  - Disconnect (connection)
  - Exit

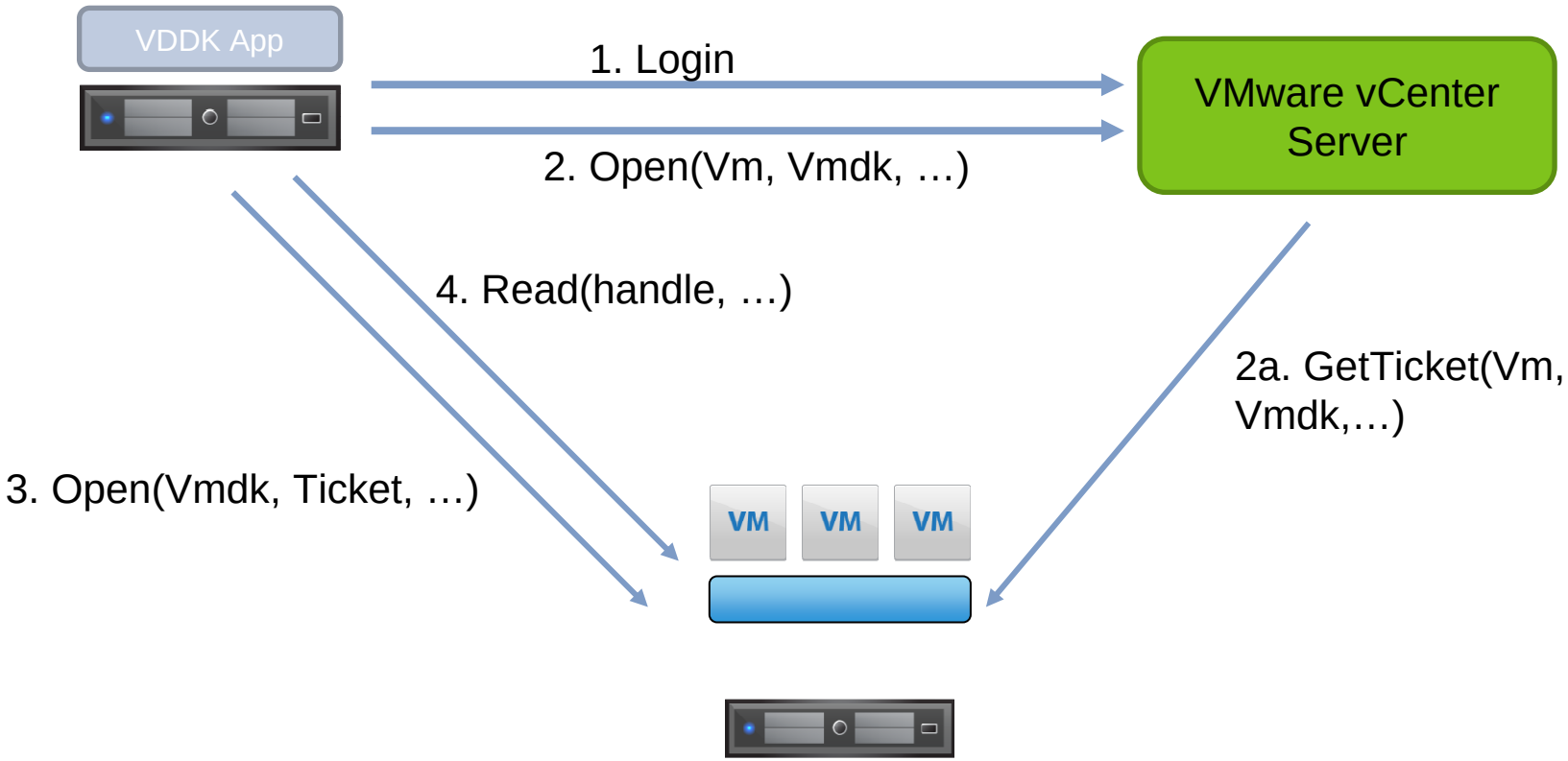
VixDiskLib API – some details

- VixDiskLib_ConnectEx needs
  - Host (ESX or VC) details – name, credentials
  - VM to which the disk belongs
  - Snapshot corresponding to the disk(s) that you want to manipulate
  - Allowed transport modes
    - If none of the specified is possible, always defaults to NBD
- VixDiskLib_Open
  - Needs the full path name of the disk
    - Of the form “[<DatastoreName>] path/to/vmdk.vmdk
    - Disk must belong to the snapshot referenced in the ConnectEx call

VDDK - Transports



VDDK – NBD transport



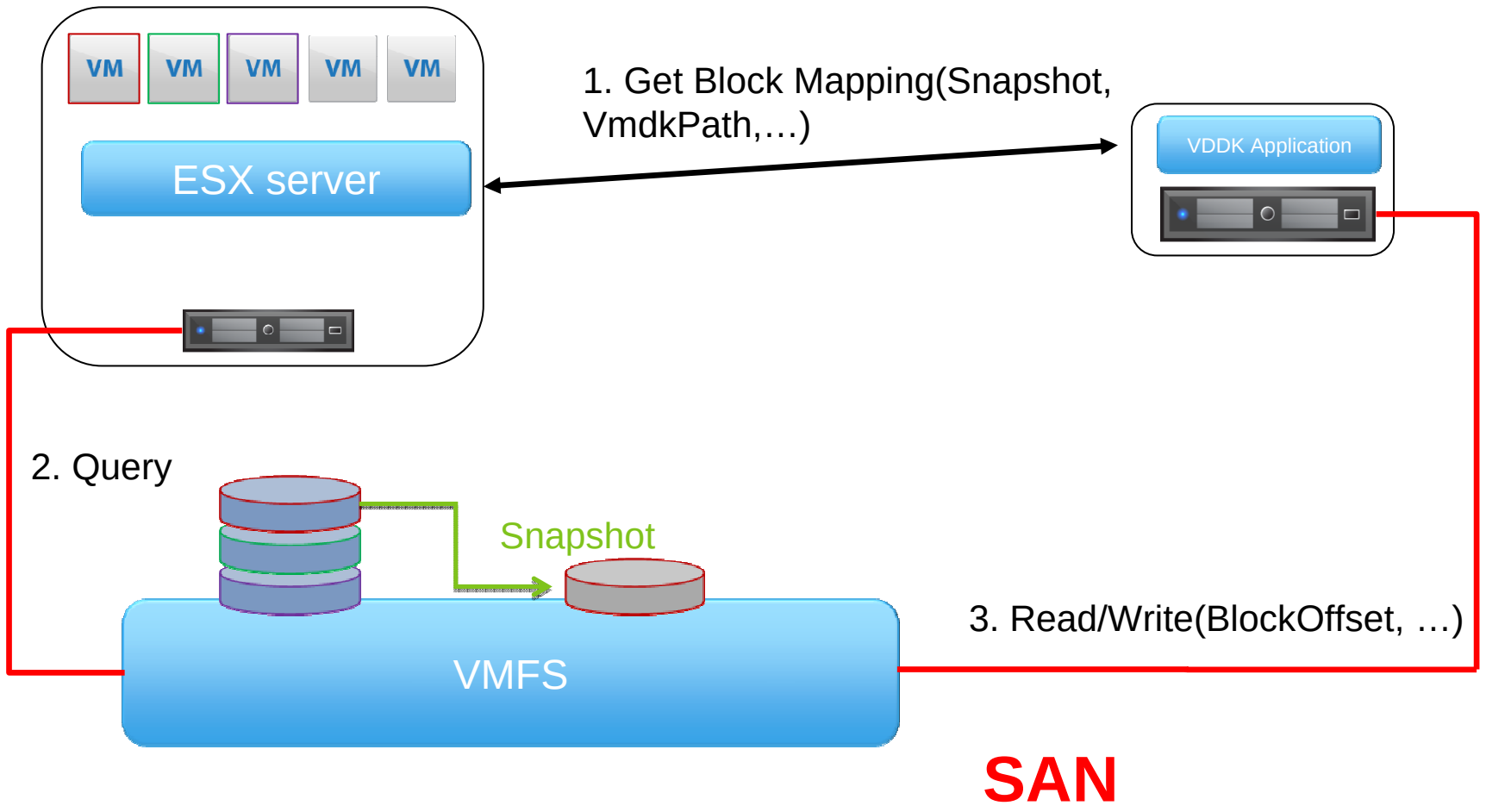
VDDK – HotAdd transport (new in 1.1)

- Uses SCSI hot-add capability of ESX to attach disk to be accessed to the VM running the VDDK application (Virtual Appliance – VA)
- Disk has to be on storage accessible by ESX host running VA
- For read-only access an additional redo log is created (disk config looks similar to linked clones)
- Descriptor file in VA holds virtual disk metadata, points to in-VA device node of hot-added disk (looks like “raw disk” in Workstation to software inside VA)
- Internally, VixDiskLib_Open on this descriptor file is used to access disk
- Needs a registered, helper VM with “VCB-HELPER(<VA name>)” for anything other than vSphere 4.0 Virtual Center.

VDDK – SAN transport (new in 1.1)

- Efficient access to virtual disk when VDDK application is running on physical hardware (“proxy”)
- Proxy needs direct access to VMFS LUNs storing virtual disks
- ESX/VC acts as metadata server, providing information about layout of virtual disk on SAN LUNs
- SAN transport uses this information to directly read data off SAN LUN
- No load on ESX host, “LAN free” transfer possible

VDDK – SAN transport



VixDiskLib – Some tips

- VDDK Logs
  - Provide callbacks to VixDiskLib_InitEx
    - No log level, but different callbacks for warning and informational logs
- Third party dependencies
  - OpenSSL, Glib, …
  - Use the VDDK versions
- Try the command line utilities
- Verify parameters
  - Use the managed object viewer (https://<host>/mob)
- Multi-threading is supported with limitations

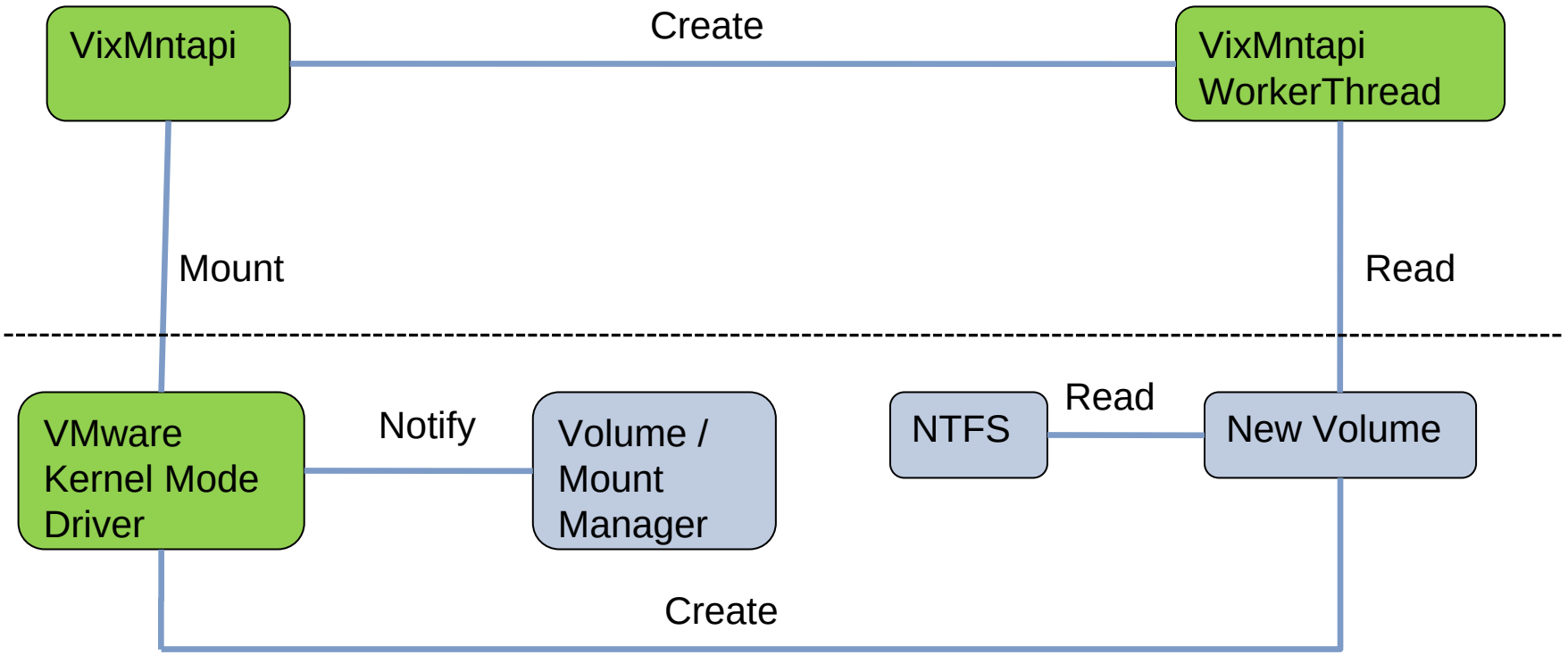
VixMntapi (new in 1.1)

- VixMntapi allows the developer to
  - treat a number of disks as a DiskSet
  - look for volumes in the DiskSet
  - support (Windows only)
    - Simple volumes – limited to a single partition
    - Spanned volumes – spans multiple partitions, possibly on different disks
    - Mirrored volumes
    - Striped volumes – data alternating among many partitions
  - mount or unmount a Volume
- VixMntapi does not understand file systems
  - File system support is assumed to be present in the OS

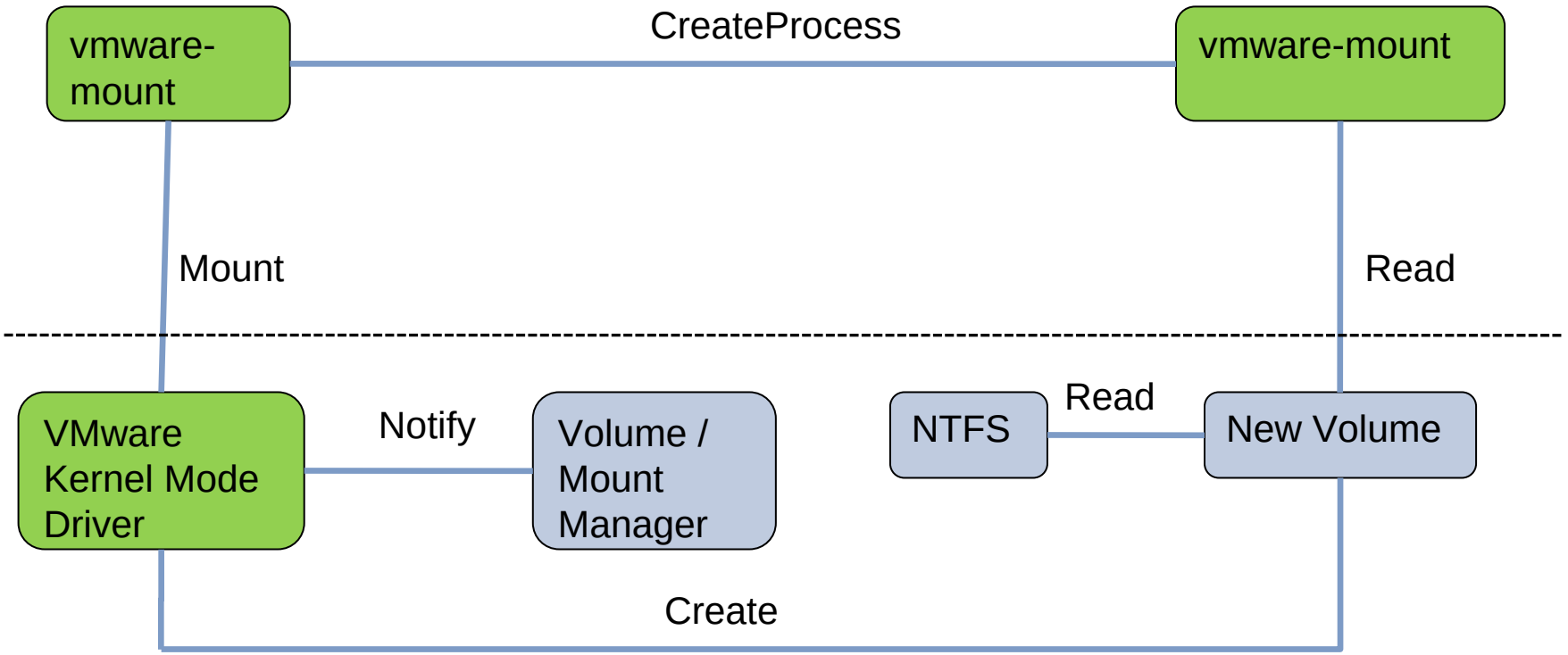
VixMntapi - program flow

- VixMntapi_Init(…)
- VixDiskLib_Connect(…) – returns a connection
- VixMntapi_OpenDisks(connection, disks,…) – returns a DiskSet handle
- VixMntapi_GetVolumeHandles
- VixMntapi_MountVolume(volumeHandle)
- VixMntapi_GetVolumeInfo(volumeHandle)
- …. OS functions to manipulate the mounted file-system …
  - CreateFile etc. on Windows
  - fopen etc. on Linux
- … Cleanup …

VixMntapi Internals – Windows



vmware-mount Internals – Windows



VixMntapi internals - Windows

- After mounting the volume
  - Use the device's symbolic link
    - CreateFile, ReadFile. Create reparse point, Map to drive letter
    - Get more details from Windows (total space, space used …)
- Make sure you unmount the volume cleanly
  - Windows may show device not available errors
  - Don't stop the thread that services the kernel mode request
- Volume objects are session specific
- vmware-mount always maps to a drive letter
  - You can remove the mapping using a Win32 program
  - Directly use the device name

VixMntapi internals - Linux

- Use FUSE – File system in user space
- FUSE
  - Kernel module + User land component
  - Kernel module preinstalled on many Linux distro's – may not be enabled
    - modprobe fuse
  - User land component (libfuse.so)
  - You can get it from an rpm like http://www.atrpms.net/dist/el5/fuse/
- FUSE is used to present a flat file representation of the Disk
  - You can do this trivially if the disk is flat
- Volumes are mounted using loop back device and the mount command
- After mounting
  - Use the mount point like any other volume mount point

References and Links

- VDDK main landing page
  - http://communities.vmware.com/community/developer/vddk
  - http://www.vmware.com/support/developer/vddk/
- VMware SDK main landing page
  - http://www.vmware.com/support/pubs/sdk_pubs.html
- VI SDK
  - http://communities.vmware.com/community/developer/managementapi
- VDDK blog
  - http://blogs.vmware.com/vddk/
- Useful Windows resource
  - Windows Internals series
    http://www.microsoft.com/learning/en/us/books/12069.aspx

Q & A

Thank you

VMware vSphere™ 4

The Best Platform for Building Cloud Infrastructures

vmware

VMware Virtual Disk Development Kit

Sudarsan Piduri, VMware