VMware Storage Best Practices

***Patrick Carmichael** – Escalation Engineer, Global Support Services.*

© 2011 VMware Inc. All rights reserved

Theme

Just because you COULD, doesn’t mean you SHOULD.

Lessons learned in Storage Best Practices

Just because you Could, doesn't mean you SHOULD.

- **Storage Performance and Technology**
  - Interconnect vs IOP.
    - Disk and RAID differences.
  - SSD vs Spinning Media.
  - VAAI
    - Xcopy/write_same
    - ATS
  - VMFS5
  - Thin Provisioning
- **Architecting for Failure**
  - Planning for the failure from the initial design.
    - Individual Components
    - Complete Site Failure
      - Backup RTO
      - DR RTO

Storage Performance – Interconnect vs IOP

- **Significant advances in interconnect performance**
  - FC 2/4/8GB
  - iSCSI 1G/10G
  - NFS 1G/10G
- **Differences in performance between technologies.**
  - None – NFS, iSCSI and FC are effectively interchangeable.
- **Despite advances, performance limit is still hit at the media itself.**
  - 90% of storage performance cases seen by GSS that are not config related, are media related.
  - Payload (throughput) is fundamentally different from IOP (cmd/s).
  - IOP performance is always lower than throughput.

Factors that affect Performance

- **Hard Disks**
  - Disk subsystem bottlenecks
  - Performance versus Capacity
- **Performance versus Capacity**
  - Disk performance does not scale with drive size
  - Larger drives generally equate lower performance
- **IOPS(I/Os per second) is crucial**
  - How many IOPS does this number of disks provide?
  - How many disks are required to achieve a required number of IOPS?
- **More spindles generally equals greater performance**

Factors that affect Performance - RAID

- RAID is used to aggregate disks for performance and redundancy
- However RAID has an I/O Penalty for Writes
- Reads have an IO penalty of 1.
- Write IO penalty varies depending on RAID choice

RAID Type	IO Penalty
1	2
5	4
6	6
10	2

Factors that affect Performance - I/O Workload and RAID

- **Understanding workload is a crucial consideration when designing for optimal performance.**
- **Workload is characterized by IOPS and write % vs read %.**
- **Design choice is usually a question of:**
  - How many IOPs can I achieve with a given number of disks?
    - Total Raw IOPS = Disk IOPS * Number of disks
    - Functional IOPS = (Raw IOPS * Write%)/(Raid Penalty) + (Raw IOPS * Read %)
  - How many disks are required to achieve a required IOPS value?
    - Disks Required = ((Read IOPS) + (Write IOPS*Raid Penalty))/ Disk IOPS

IOPS Calculations – Fixed number of disks

- **Calculating IOPS for a given number of disks**
  - 8 x 146GB 15K RPM SAS drives
  - ~150 IOPS per disk
  - RAID 5
  - 150 * 8 = 1200 Raw IOPS
  - Workload is 80% Write, 20% Read
  - (1200*0.8)/4 + (1200*0.2) = 480 Functional IOPS

Raid Level	IOPS(80%Read 20%Write)	IOPS(20%Read 80%Write)
5	1020	480
6	1000	400
10	1080	720

IOPS Calculations – Minimum IOPS Requirement

- **Calculating number of disks for a required IOPS value**
  - 1200 IOPS required
  - 15K RPM SAS drives. ~150 IOPS per disk
  - Workload is 80% Write, 20% Read
  - RAID 5
  - Disks Required = (240 + (960*4))/150 IOPS
  - 27 Disks required

Raid Level	Disks(80%Read 20%Write)	Disks (20%Read 80%Write)
5	13	27
6	16	40
10	10	14

What about SSD?

- **SSD potentially eliminates the physical limitation of spinning media.**
  - Advertised speeds of 10,000 IOPS+
    - Only reached under very specific conditions.
    - Real world performance must be tested
      - Test with IO footprint as close to your intended use as possible
      - Actual values will vary, but will be significantly higher than spinning media
  - The value of testing, regardless of SSD or traditional media, cannot be understated. Every array is different.

VAAI (xcopy/write_same)

- **Advertised as a way to improve performance of certain operations**
  - Despite common belief, VAAI does not reduce load.
  - Offload to array of certain operations
    - A storage array is built to handle these operations – far more efficient, and much faster than ESX sending the commands for each individual block.
    - Still requires the disks perform the commands in question
  - In some scenarios, offloading these operations can push the array past its limits, much like doing the same sequence on the host would.
  - If your environment is at maximum performance capacity, VAAI will not allow you to do things you could not otherwise do.

VAAI (ATS)

- **A final answer to the SCSI Reservation problem.**
  - Everyone is familiar with the issues behind SCSI reservations.
    - Whole lun locking for simple metadata changes
    - Blocks IO from all other hosts
    - Lost reserves can mean downtime
    - Differing capabilities by vendor / model mean different maximums.
  - ATS instead locks (via a new SCSI spec) only the blocks in question.
  - Eliminates the design limitations of SCSI reserves
    - Capable of handling significantly larger VM/lun ratios.
    - Allows for larger luns without lost space.

VAAI (ATS) contd.

- **Remember our theme: Just because you could, doesn't mean you should.**
  - ATS will allow you to significantly increase consolidation ratios (by up to 100% in some cases) per-lun.
  - It will not, however, guarantee the underlying spindles can handle the normal IO load of said VMs.
  - Primarily a concern with linked clone environments
    - View/VCD/Lab Manager vms take up very little space (storing changes / persistent disks only)
    - Linked clones generate significant amounts of reservations
    - ATS is designed specifically to handle this, but many forget that the VMs have a normal IO load as well that can overwhelm the disks in other ways.
  - **Doubling VM count doubles IO load.**
    - Consider all the implications of what the technology will allow you to do.

VAAI (ATS) in ESX5.

- **New feature! ATS-Only volumes.**
  - Any volume created on ESX5, as VMFS5, where the array reports that it supports ATS (at the time of creation), will be created as ATS-only.
    - Flag disables SCSI-2 reservations.
    - This is good!
      - No reservation storm from ATS failures.
- **This also means that if something changes, your volume may be unreadable.**
  - SRM – does your DR site have an ATS capable array?
    - If not, volumes won't mount (different firmware revisions).
  - Some firmware upgrades on arrays disable their ATS support.
- **A global option can be set to disable this feature, or set per-volume. KB to be public shortly. Some info in KB 2006858**

VAAI (other minor considerations)

- **Performance graphs/esxtop will be skewed for VAAI commands.**
  - Unlike traditional commands that receive an acknowledgement for each command/block, the array will execute multiple commands for each VAAI command
  - This takes significantly longer, but esxtop/performance graphs expect each command to return as normal, so the values reported will be skewed.
  - Does not indicate a performance problem.

VMFS5

- **VMFS5 is the new, 3rd generation filesystem from VMware**
  - Introduced with vSphere5
  - Eliminates 2TB-512B size limit
    - Max size: 64TB
  - 1MB block size
    - File size for VMDKs still limited to 2TB currently
    - 64TB max for pRDM
  - GPT partition table (with backup copy at end of disk).
- **Allows use of truly large logical units for workloads that would previously have required extents/spanned disks.**

VMFS5 contd.

- **VMFS5, in combination with ATS, will allow consolidation of ever-larger number of VMs onto single VMFS datastores.**
  - One lun could contain the VMs previously stored on 32 (assuming max utilization).
  - While potentially easier for management, this means that 32 LUNs worth of VMs are now reliant on a single volume header.
  - When defining a problem space, you've now expanded greatly the number of items in that problem space
- **Just because you could, does that mean you should?**

VM
VM
VM
VM
VMware ESX

Thin Provisioning

- **Thin provisioning offers very unique opportunities to manage your storage "after" provisioning.**
  - Workloads that require a certain amount of space, but don't actually use it.
  - Workloads that may grow over time, and can be managed/moved if they do.
  - Providing space for disparate groups that have competing requirements.
- **The question is, where should you thin provision, and what are the ramifications?**

Thin Provisioning – VM Level

- **VM disk is created @ 0b, until used, and then grows @ VMFS Block size as needed**
- **Minor performance penalty for provisioning / zeroing.**
- **Disk cannot currently be shrunk – once grown, it stays grown.**
  - There are some workarounds for this, but they are not guaranteed

- **What happens when you finally run out of space?**
  - All VMs stun until space is created
  - Production impacting, but potentially a quick fix (shut down VMs, adjust memory reservation, etc).
  - Extremely rare to see data loss of any kind.

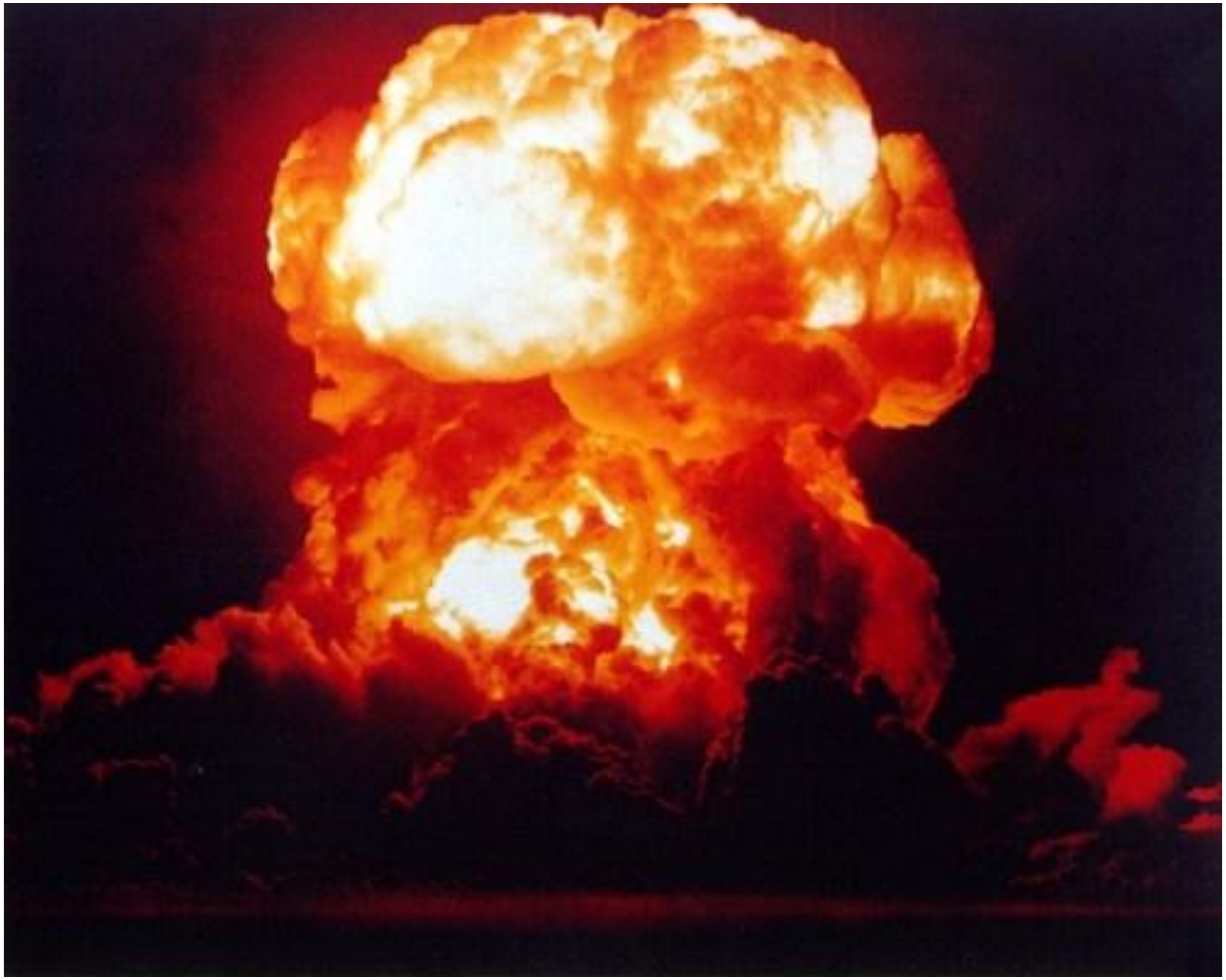
Thin Provisioning – LUN Level.

- Allows your array to seem bigger than it actually is.
- Allows you to share resources between groups (the whole goal of virtualization).
- Some groups may not use all or much of what they’re allocated, allowing you to utilize the space they’re not using.
- Standard sized or process defined luns may waste significant amounts of space, and space being wasted is $$ being wasted.
- Significant CapEX gains can be seen with thin luns.

Thin Provisioning – LUN Level - contd

- **What happens when you finally run out of space?**
  - New VAAI primitives, for compatible arrays, will let ESX know that the underlying storage has no free blocks.
    - If VAAI works, and your array is compatible, and you're on a supported version of ESX, this will result in the same as a thin VMDK running out of space – All VMs will stun (that are waiting on blocks). VMs not waiting on blocks will continue as normal.
    - Cleanup will require finding additional space on the array, as VSWP files / etc will already be allocated blocks at the lun level. Depending on your utilization, this may not be possible, unless UNMAP also works (very limited support at this time).
  - If VAAI is not available for your environment, or does not work correctly, then what?
    - On a good day, the VMs will simply crash with a write error, or the application inside will fail (depends on array and how it handles a full filesystem).
    - Databases and the like are worst affected, will most likely require rebuild/repair.
    - And on a bad day?

Thin Provisioning LUN Level – contd.



Thin Provisioning

- **There are many reasons to use Thin Provisioning, at both the VM and the LUN level.**
- **Thin provisioning INCREASES the management workload of maintaining your environment. You cannot just ignore it.**

Details for new VAAI Features

- http://blogs.vmware.com/vsphere/2011/07/new-enhanced-vsphere-50-storage-features-part-3-vaai.html
- Please note, UNMAP has been disabled by default as of P01, due to issues with some arrays. Please confirm with your vendor the support status before turning it back on.

Just because you Could, doesn’t mean you Should

- Everything we’ve covered so far is based on new technologies
- What about the existing environments?

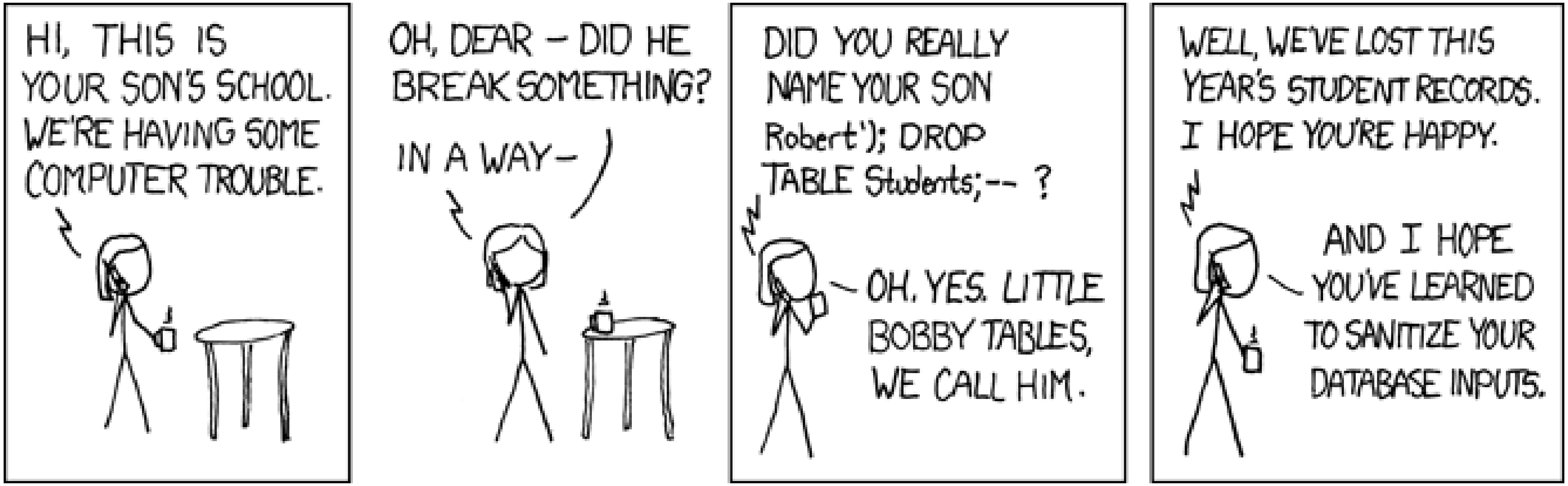
The ultimate extension of “Just because you could, doesn’t mean you should,” is what I call “Architecting for Failure”

Architecting for Failure

- **The ultimate expression of “Just because you could, doesn’t mean you should.”**
  - Many core infrastructure designs are built with tried and true hardware and software, and people assume that things will always work
  - We all know this isn’t true – Murphy’s law.
- **Architect for the failure.**
  - Consider all of your physical infrastructure.
    - If any component failed, how would you recover?
    - If everything failed, how would you recover?
  - Consider your backup/DR plan as well

Black Box Testing

- **Software engineering concept.**
  - Consider your design, all of the inputs, and all of the expected outputs.
  - Feed it good entries, bad entries, and extreme entries, find the result, and make sure it is sane.
    - If not, make it sane.
- **This can be applied before, or after, you build your environment**


HI, THIS IS YOUR SON'S SCHOOL. WE'RE HAVING SOME COMPUTER TROUBLE.
OH, DEAR – DID HE BREAK SOMETHING?
IN A WAY–
DID YOU REALLY NAME YOUR SON Robert'); DROP TABLE Students;-- ?
OH. YES. LITTLE BOBBY TABLES, WE CALL HIM.
WELL, WE'VE LOST THIS YEAR'S STUDENT RECORDS. I HOPE YOU'RE HAPPY.
AND I HOPE YOU'VE LEARNED TO SANITIZE YOUR DATABASE INPUTS.

Individual Component Failure

- **Black box: consider each step your IO takes, from VM to physical media.**
- **Physical hardware components are generally easy to compensate for.**
  - VMware HA and FT both make it possible for a complete system to fail with little/no downtime to the guests in question.
  - Multiple hardware components add redundancy and eliminate single points of failure
    - Multiple NICs
    - Multiple storage paths.
    - Traditional hardware (multiple power supplies, etc).
  - Even with all of this, many environments are not taking advantage of these features.
- **Sometimes, the path that IO takes passes through a single point of failure that you don’t realize is one**

What about a bigger problem?

- You’re considering all the different ways to make sure individual components don’t ruin your day.

- What if your problem is bigger?

Temporary Total Site Loss

- **Consider your entire infrastructure during a temporary complete failure.**
  - What would happen if you had to bootstrap it cold?
    - This actually happens more often than would be expected.
      - Hope for the best, prepare for the worst.
    - Consider what each component relies on – do you have any circular dependencies?
      - Also known as the “Chicken and the Egg” problem, these can increase your RTO significantly.
      - Example: Storage mounted via DNS, all DNS servers on the same storage devices. Restoring VC from backup when all networking is via DVS.
    - What steps are required to bring your environment back to life?
  - How long will it take? Is that acceptable?

Permanent Site Loss.

- **Permanent site loss is not always an “Act of God” type event**
  - Far more common is a complete loss of a major, critical component.
    - Site infrastructure (power, networking, etc) may be intact, but your data is not
      - Array failures (controller failure, major filesystem corruption, RAID failure)
      - Array disaster (thermal event, fire, malice)
      - Human error – yes, it happens!
  - Multiple recovery options – which do you have?
    - Backups.
      - Tested and verified?
      - What’s your RTO for a full failure?
    - Array based replication
      - Site Recovery Manager
      - Manual DR
        - How long for a full failover?
    - Host based replication.

Permanent Site Loss – contd.

- **Consider the RTO of your choice of disaster recovery technology.**
  - It equates directly to the amount of time you will be without your virtual machines.
  - How long can you, and your business, be without those services?
  - A perfectly sound backup strategy is useless, if it cannot return you to operation quickly enough.
- **Architect for the Failure – make sure every portion of your environment can withstand a total failure, and recovery is possible in a reasonable amount of time.**

The End.

- **Questions?**