vmware®

VMware Infrastructure 3

Recommendations for Aligning VMFS Partitions

Partition alignment is a known issue in physical file systems, and its remedy is well documented. The goal of the testing reported in this paper was to validate the assumption that unaligned partitions also impose a penalty when the partition is a VMware Virtual Machine File System (VMware VMFS) partition.

This paper lists a summary of the results of our testing, recommendations for VMware VMFS partition alignment, and the steps needed to create aligned VMware VMFS partitions. It covers the following topics:

Executive Summary

Our testing validates the assumption that VMware VMFS partitions that align to 64KB track boundaries result in reduced latency and increased throughput. Partition alignment on both physical machines and VMware VMFS partitions prevents performance I/O degradation due to unaligned tracks. Creating VMware VMFS partitions using the Virtual Infrastructure Client (VI Client) that is part of VMware Infrastructure 3 results in a partition table aligned on the 64KB boundary as storage and operating system vendors recommend.

Throughput Increase		**Latency Decrease**	
Min = 2%	Average = 12%	Min = 7%	Average = 10%
Max = 62%		Max = 33%	

Note: These recommendations are for block-based storage solutions, not those that are IP-based. I/O characteristics on NFS are different from those of Fibre Channel and iSCSI storage systems. Though partition alignment eliminates track crossings and benefits performance on all

storage platforms, the throughput improvements in specific types of I/O between SAN, NFS, and iSCSI are different.

See the section titled Performance Results on page 8 for more details.

Recommendations

We recommend against wholesale migration of all unaligned VMware VMFS partitions for the following key reasons:

- Partition alignment requires cold migration, which is disruptive. To align a partition, you use `fdisk` on the ESX Server host or use the VI Client to delete then recreate the VMware VMFS partition. This requires shutting down, backing up, restoring, then restarting the virtual machines that reside on the target VMware VMFS partition.
- The performance degradation of unaligned partition occurs during intensive I/O workloads rather than on those with low to moderate I/O activity.

Based on our studies we make the following recommendations for VMware VMFS partitions:

- Carefully evaluate the I/O workload against your unaligned VMware VMFS partitions. For example, workloads consisting of mostly of sequential reads gain the most from partition alignment. If the workload is light, the system already meets service level agreements, or the workload contains many random writes, consider delaying the partition alignment procedure until a scheduled outage or not migrating at all.
- The best practice for adding VMware VMFS storage to ESX Server is to use the VI Client, because it automatically aligns VMware VMFS partitions when it creates them.
- To manually align your VMware VMFS partitions, first check your storage vendor's recommendations for the partition starting block. For example, in the *EMC CLARiiON Best Practices for Fibre Channel Storage* guide available at EMC Powerlink, EMC recommends a starting block of 128 to align the partition to the 64KB boundary. If your storage vendor makes no specific recommendation, use a starting block that is a multiple of 8KB.

Also note that an in-place migration for unaligned VMFS partitions (such as an upgrade from ESX Server 2.x to ESX Server 3.x) does not align partitions automatically.

Instructions for VMware VMFS Partition Alignment Using fdisk

To check that your existing partitions are aligned, issue the command:

```
fdisk -lu
```

The output is similar to:

```
Device boot   Start     End         Blocks      Id   System
/dev/sdj1     128       167766794   83883333+   fb   Unknown
```

Aligned partitions start at 128. If the Start value is 63 (the default), the partition is not aligned.

If you choose not to use the VI Client and create partitions with `vmkfstools`, or if you want to align the default installation partition before use, take the following steps to use `fdisk` to align a partition manually from the ESX Server service console:

1. Enter `fdisk /dev/sd<x>` where <x> is the device suffix.

2. Determine if any VMware VMFS partitions already exist. VMware VMFS partitions are identified by a partition system ID of `fb`. Type `d` to delete to delete these partitions.

   Note: This destroys all data currently residing on the VMware VMFS partitions you delete. Ensure you back up this data first if you need it.
3. Type `n` to create a new partition.
4. Type `p` to create a primary partition.
5. Type `1` to create partition No. 1.
6. Select the defaults to use the complete disk.
7. Type `t` to set the partition's system ID.
8. Type `fb` to set the partition system ID to fb (VMware VMFS volume).
9. Type `x` to go into expert mode.
10. Type `b` to adjust the starting block number.
11. Type `1` to choose partition 1.
12. Type `128` to set it to 128 (the array's stripe element size).
13. Type `w` to write label and partition information to disk.

Instructions for Guest File System Alignment

Once you have aligned your VMware VMFS partitions, you also need to align the data file system partitions within your virtual machines.

Note: Aligning the boot disk in the virtual machine is neither recommended nor required. Align only the data disks in the virtual machine.

The following sections discuss how to align guest operating system partitions in Linux and Windows environments.

Linux

A best practice for Linux physical as well as virtual machines is to align file system partitions using `fdisk`. Use the `fdisk` procedure in the previous section of this paper, and instead of setting the partition system id to `fb`, set it to `83` (Linux) or other appropriate partition system ID.

Windows

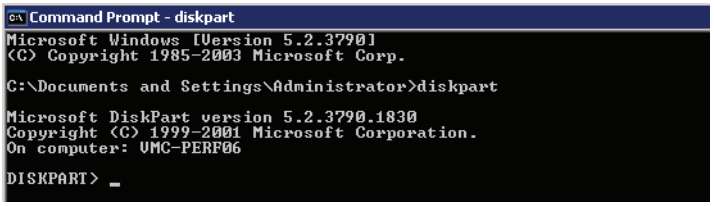
For Windows virtual machines there is an additional layer of NTFS partitioning that requires alignment using the `diskpart.exe` tool from the Microsoft Download Center or MSDN.

Also, using a larger allocation unit (also called cluster size) improves performance of NTFS volumes.

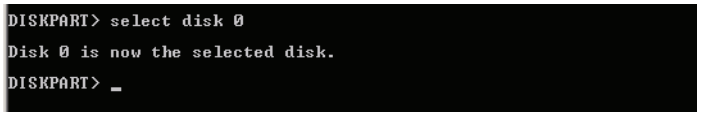
Note: The previous version of `diskpart.exe` was `diskpar.exe`. The main difference is that `diskpart.exe` creates partitions in sectors (512 bytes) while diskpart uses KB.

This example assumes the new disk is disk 0. To create an aligned partition, take the following steps:

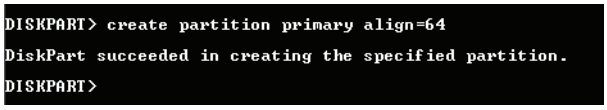
1. 1.Ensure that no data exists on the disk. Then open a command prompt and start the disk partitioning utility.
   `diskpart`



2. Enter the command to select disk 0.
   `select disk 0`

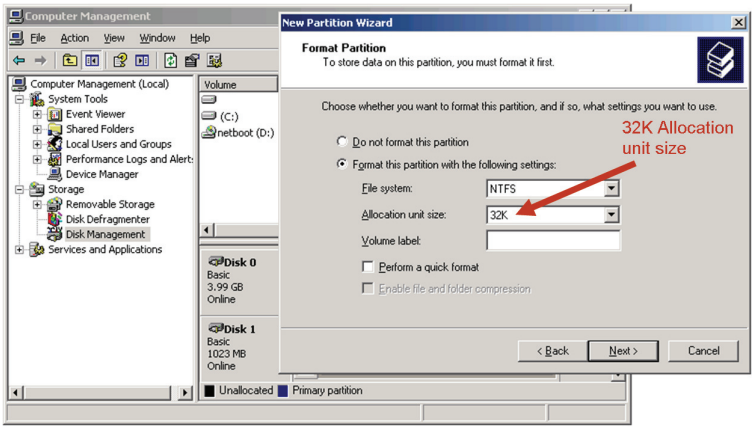


3. Create the aligned primary partition.
   `create partition primary align=64`



4. Exit the diskpart.exe utility.
   `exit`
5. Close the Windows command prompt window.
6. Format the drive with a 32K allocation size.

   Start Windows Disk Manager by right-clicking My Computer on the desktop, then choosing **Manage**. From **Computer Management**, choose **Disk Management**.
7. Select the new unformatted disk, then right-click and choose **Format**.
8. When asked for the allocation unit size, choose 32K.



Caveats

Different user scenarios present different challenges, therefore we present some caveats associated with these recommendations. Some of the following conditions could change the performance of VMware VMFS partitions.

- Though these tests took place on ESX Server 3.0, these recommendations also apply to ESX Server 2.0 and later versions.
- We obtained the results on an isolated configuration. The VMware VMFS volume was not shared, and the array was zoned such that no other hosts could access the LUNs in the SAN.
- The results assume that all disks are in persistent mode.
- These recommendations assume there are no VirtualCenter operations or user-initiated copy operations (for example, from the service console) in the same VMware VMFS volume running in the background. Such operations share bandwidth with the virtual machines and may affect performance.
- ESX Server swap files reside on the VMware VMFS volumes by default. For the purposes of these experiments, we did not host the swap files on the VMware VMFS volume so that we could isolate the performance impact of alignment.

Background

In a SAN environment, the smallest hardware unit used by a SAN storage array to build a LUN out of multiple physical disks is a called a chunk or a stripe. To optimize I/O, chunks are usually much larger than sectors. Thus a SCSI I/O request that intends to read a sector in reality reads one chunk.

On top of this, in a Windows environment NTFS is formatted in blocks ranging from 1MB to 8MB. The file system used by the guest operating system optimizes I/O by grouping sectors into so-called clusters (allocation units).

Figure 1 shows that an unaligned structure may cause many additional I/O operations when only one cluster is ready by the guest operating system.

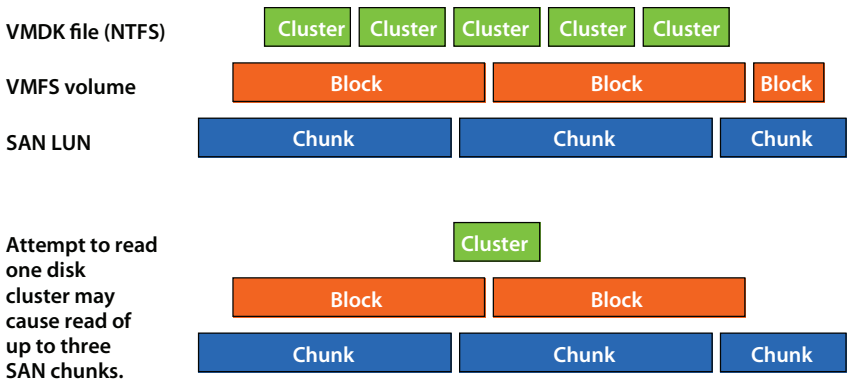


Figure 1: Unaligned partitions result in additional I/O

Figure 2 shows I/O improvements on a properly aligned Windows NTFS volume in a VMDK on a SAN LUN.

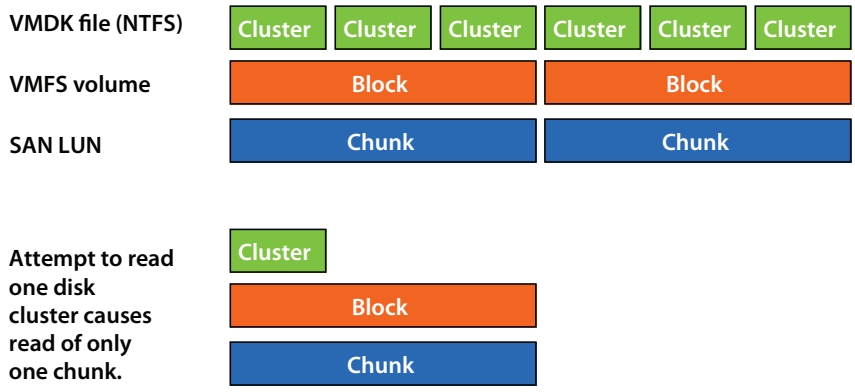


Figure 2 Aligned partitions reduce I/O

Also, operating systems on x86 architectures create partitions with a master boot record (MBR) that consumes 63 sectors. This is due to legacy BIOS code from the PC that used cylinder, head, and sector addressing instead of logical block addressing (LBA). Without LBA, the first track is reserved for the boot code, and the first partition starts at cylinder 0, head 1, and sector 1. This is LBA 63 and is therefore unaligned.

An unaligned partition results in a track crossing and an additional I/O, incurring a penalty on latency and throughput. The additional I/O (especially if small) can impact system resources significantly on some host types. An aligned partitions ensures that the single I/O is serviced by a single device, eliminating the additional I/O and resulting in overall performance improvement.

During ESX Server 3.0 installation, however, the installation procedure creates a default VMware VMFS partition that is unaligned. Administrators should consider manually aligning the default VMware VMFS partition with `fdisk` before use.

Test Environment

Storage Layout

In the context of this benchmark testing, storage layout refers to the location and type of the disk used in the tests. Tests were conducted against a 20GB virtual disk located on a five-disk RAID 5 LUN in an EMC CLARiiON SAN array. Virtual disks are implemented as files on the underlying storage. From the perspective of the virtual machine, the disk appeared to be a physical drive.

Software Configuration

Unless stated otherwise, all ESX Server and guest operating system parameters were left at their default settings.

I/O Workload Characteristics

Servers typically run a mix of workloads consisting of different access patterns and I/O data sizes. Within a workload there may be several data transfer sizes and more than one access pattern.

There are a few applications in which access is either purely sequential or purely random. For example, database logs are written sequentially. Reading this data back during database recovery is done by means of a sequential read operation. Typically, online transaction processing (OLTP) database access is predominantly random in nature.

The size of the data transfer depends on the application and is often a range rather than a single value. For Microsoft Exchange, the I/O size is generally small (from 4KB to 16KB), Microsoft SQL Server database random read and write accesses are 8KB, Oracle accesses are typically 8KB, and Lotus Domino uses 4KB. On the Windows platform, the I/O transfer size of an application can be determined using Perfmon.

In summary, I/O characteristics of a workload are defined in terms of the ratio of read operations to write operations, the ratio of sequential accesses to random accesses, and the data transfer size. Often, a range of data transfer sizes may be specified instead of a single value.

Test Cases

The primary objective was to characterize the performance of virtual machines on both unaligned and aligned partitions for a range of data sizes across a variety of access patterns. The data sizes selected were 1KB, 4KB, 8KB, 16KB, 32KB, 64KB, 72KB, and 128KB. The access patterns were restricted to a combination of 100 percent read or write and 100 percent random or sequential. Each of these four workloads was tested for eight data sizes, for a total of 32 data points per workload.

Load Generation

The Iometer benchmarking tool, originally developed at Intel and widely used in I/O subsystem performance testing, was used to generate I/O load for these experiments. A well-designed set of configuration options allows a wide variety of workloads to be emulated and executed. Since this investigation was intended to characterize the relative performance of the two partition configurations, only the basic load emulation features were used in these tests.

Iometer configuration options used as variables in these experiments:

- Transfer request sizes: 1KB, 4KB, 8KB, 16KB, 32KB, 64KB, 72KB, and 128KB
- Percent random or sequential distribution: for each transfer request size, 0 percent and 100 percent random accesses were selected
- Percent read or write distribution: for each transfer request size, 0 percent and 100 percent read accesses were selected

Iometer parameters that were held constant for all tests:

- Size of virtual disk: 20GB
- Number of outstanding I/O operations: 16
- Runtime: 4 minutes
- Ramp-up time: 60 seconds
- Number of workers to spawn automatically: 1

With a 512MB read cache in the array, the virtual disk size of 20GB minimizes the effect of array caching.

vmware®

Performance Results

In general, all workloads show an increase in throughput when the partition is aligned.

Figure 3 shows improvement for read I/O (sequential and random) on RAID 5 aligned and unaligned partitions. Figure 4 shows improvement for write I/O (sequential and random) on RAID 5 aligned and unaligned partitions.

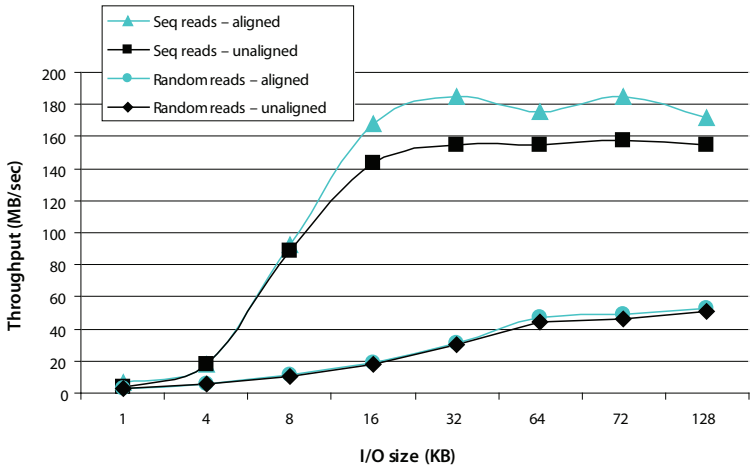


Figure 3: Reads from RAID 5

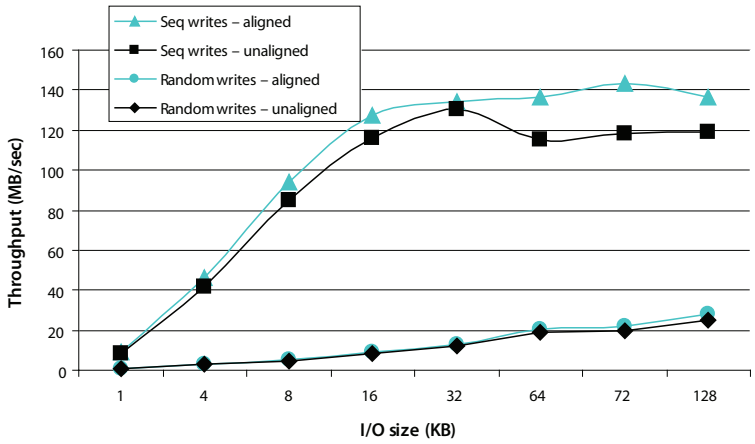


Figure 4: Writes to RAID 5

Figure 5 shows the average throughput (MB per second) at various I/O block sizes for RAID 5 unaligned partitions.

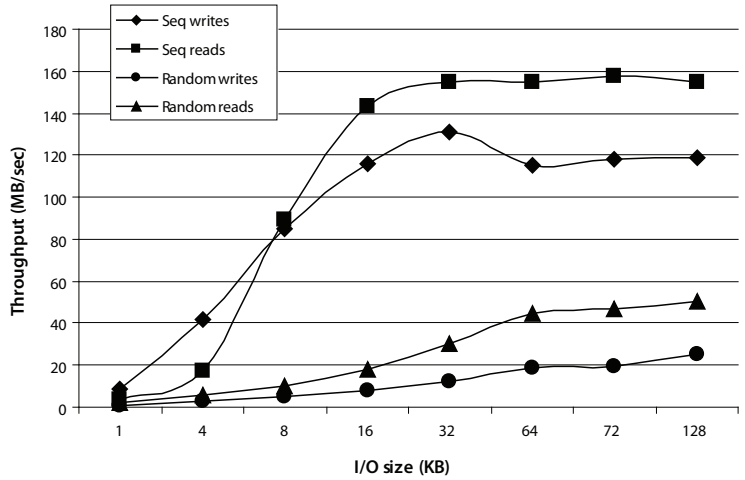


Figure 5: Average throughput for RAID 5 unaligned partitions

Figure 6 shows the average throughput (MB per second) at various I/O block sizes for RAID 5 aligned partitions.

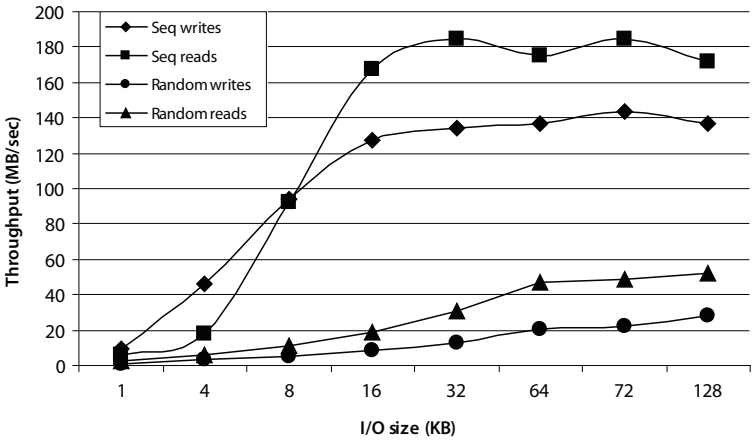


Figure 6: Average throughput for RAID 5 aligned partitions

Conclusion

Track alignment for both physical machines and VMware VMFS partitions yields I/O performance improvements such as reduced latency and increased throughput. Creating VMware VMFS partitions using the VI Client results in a 64KB-aligned partition table and provides the foundation for a best practices storage layout.

References

EMC PowerLink
powerlink.emc.com/

Microsoft TechNet: How to Align Exchange I/O with Storage Track Boundaries
www.microsoft.com/technet/prodtechnol/exchange/guides/StoragePerformance/0e24eb22-fbd5-4536-9cb4-2bd8e98806e7.mspx

Microsoft Windows 2000 Resource Kit: Examining and Tuning Disk Performance
www.microsoft.com/resources/documentation/Windows/2000/server/reskit/en-us/prork/pree_exa_xiep.asp

A Description of the Diskpart Command-Line Utility
support.microsoft.com/kb/300415

If you have comments about this documentation, submit your feedback to: docfeedback@vmware.com
VMware, Inc. 3401 Hillview Ave. Palo Alto, CA 94304 www.vmware.com