

Analyzing Storage Performance



clinth@microsoft.com 18 Mar 2009 7:54 PM

0

Introduction

The purpose of this article is to provide prescriptive guidance on how to troubleshoot logical and physical disk response times in regards to Windows performance analysis.

Start with the following performance counters to analyze disk response times:

- \LogicalDisk\Avg. Disk Sec/Read
- \LogicalDisk\Avg. Disk Sec/Write
- \LogicalDisk\Disk Bytes/Sec
- \LogicalDisk\Disk Reads/Sec
- \LogicalDisk\Disk Writes/Sec
- \LogicalDisk\Split IO/sec
- \LogicalDisk\Disk Transfers/sec

These counters are generally the first ones to look at because we are looking for the following attributes of the Input/Output profile:

- **Average disk seconds/read and average disk seconds/write (response times):** Are the users having to wait for application responses? Are we exceeding established thresholds for disk drive performance degradation (generally > 15 ms)?
- **Throughput:** Are we saturating any of the pipes, such as mainboard bus, SCSI connection, SAN connection, or other link between servers and storage. Are we reaching the throughput limit of the disk subsystem?
- **Transfers/second:** Is the server and its applications generating more I/O than the disk subsystem can keep up with? For example, suppose you have allocated 4 disk drives to a single logical disk group that is configured as RAID 1+0. Assuming 200 Input/Output Operations per second (IOPs) capability of a given disk drive, that RAID group will be capable of around 400 IOPs (cache and read-ahead may increase that number somewhat). Even being very generous and saying that with cache and optimizations a disk drive can perform 400 IOPs, the most that could be hoped for in write operating on a 4 disk RAID 1+0 is 800 IOPs. If transfers/second exceeds that number at the same time as response times are deteriorating, chances are there just are not enough disk drives to back that logical disk and the assigned workload.
- **Reads/second and writes/second:** Gives you an indication as to the mix of workload that you are dealing with. Certain disk subsystem types handle certain workloads better than others. For example, some RAID-5 controllers can handle large I/O writes and sequential reads relatively well.
- **Split I/O:** Does the operating system have to perform more than one command for each I/O? Split I/O is a good indicator of fragmentation, which can reduce performance by causing excessive seek time.

This article is grouped by symptoms, then by possible causes.

Symptoms: Long disk response times and High I/O

Applies to:

- Windows Server 2003 (all editions) unless otherwise specified
- Windows XP (all editions) unless otherwise specified
- Windows Server 2000 (all editions) unless otherwise specified

Symptom Details:

- **Long disk response times:** A "\LogicalDisk\Avg. Disk Sec/Read" or "\LogicalDisk\Avg. Disk Sec/Write" value greater than 15ms though occasional spikes are not necessarily cause for immediate concern.
- **High I/O:** "\LogicalDisk\Disk Transfers/sec" is at or near the number of I/O operations per second that each physical spindle is designed to handle which is typically between 80 to 180 per disk.

Possible Causes	How to Diagnose	Possible Solutions and/or Recommendations
Storage response time reduced because of misaligned partitions.	<ul style="list-style-type: none">• A misaligned partition is the result of creating a partition that is cylinder aligned, versus sector aligned. Windows, up until Vista, aligned partitions using Cylinder, Head, and Sector addressing. Most storage controllers will still report some value for C/H/S, even though that scheme is likely not used for disk	<ul style="list-style-type: none">• DiskPart: To resolve this issue, use the Diskpart.exe tool to create the disk partition and to specify a starting offset of 2,048 sectors (1 megabyte). A starting offset of 2,048 sectors covers most stripe unit size scenarios. For more information, go to "Disk performance may be slower than expected when you use multiple disks in Windows Server 2003, in Windows XP, and in Windows 2000" http://support.microsoft.com/kb/929491

	<p>addressing. Windows will therefore create a partition on sector 63. You can see this in MSINFO32, or various other methods (look for hidden sectors, partition offset, etc).</p> <ul style="list-style-type: none">• If the starting offset for a partition is 63 sectors (32,256 bytes), you can guarantee misalignment.• Though the alignment issue predominately affects RAID disks, where a volume cluster might cross a RAID chunk boundary, there are other boundaries that could be misaligned such as cache lines.	
--	--	--

Symptoms: General poor response from storage subsystem

Applies to:

- Windows Server 2003 (all editions) unless otherwise specified
- Windows XP (all editions) unless otherwise specified
- Windows Server 2000 (all editions) unless otherwise specified

Symptom Details:

- **Long disk response times:** A "LogicalDisk\Avg. Disk Sec/Read" or "LogicalDisk\Avg. Disk Sec/Write" value greater than 15ms though occasional spikes are not necessarily cause for immediate concern.
- **Low I/O:** "\LogicalDisk\Disk Transfers/sec" is well below the number of I/O operations per second that each physical spindle is designed to handle which is typically between 80 to 180 per disk.
- **Split I/O:** "Physical Disk\Split IO/sec" can be an indicator of volume fragmentation.
- **High Queue Lengths, poor response times.** LogicalDisk\Average Disk Queue Length is averaging higher than 2-3 plus the number of spindles. At the same time, "LogicalDisk\Avg. Disk Sec/Read" or "LogicalDisk\Avg. Disk Sec/Write" value greater than 15ms are observed.
- **Low throughput, high number of transfers.** "Transfers/second" counter is relatively high, but the overall "Disk Bytes/sec" is low.

Possible Causes	How to Diagnose	Possible Solutions and/or Recommendations
High disk fragmentation	<ul style="list-style-type: none">• Use the Windows Disk Defragmentation tool to analyze the disk fragmentation. If the disk is more than 20% fragmented, then consider running the defragmentation tool during your next maintenance period.• In Performance Monitor, look at the "Split I/O" counter. This indicates that a single request was split into multiple requests, likely as the result of fragmentation.	<ul style="list-style-type: none">• Defragment the Disk: Use the Windows Disk Defragmentation tool to defragment the disk during your next maintenance period. Note: Some products like Microsoft SQL Server have clustered indexes that are mapped to hard disk clusters, therefore consult your database administrator before defragmenting.
Lack of free space	<ul style="list-style-type: none">• Use the "\LogicalDisk\% Free Space" counter to determine if the disk has less than 30% free space. % Free Space is the percentage of total usable space on the selected logical disk drive that was free. Performance is not really affected until the available disk drive space is less than 30 percent. When 70 percent of the disk drive is used, the remaining free space is located closer to the disk's spindle at the center of the disk drive, which operates at a lower performance level. Lack of disk free space can cause severe disk performance. Note: Low disk free space disk performance can vary on hardware RAID solutions depending on how the hardware spreads the data on the	<ul style="list-style-type: none">• Remove Files: Move or delete unnecessary files from the disk drive.• Add Physical Spindles: Add additional disks to the LUN or disk volume.• Increase the Partition Size: If contiguous disk space is available the partition can be expanded using Diskpart. This is often an option when disks are added to disk groups in the underlying storage device.

	spindles.	
Insufficient number of disks	<ul style="list-style-type: none"> In many cases the workload presented to a storage device is greater than was originally designed for. Look at the disk queue length counters in Performance Monitor. <p>The rule is that you want:</p> <ul style="list-style-type: none"> 2 outstanding requests for low-performance disks 3 outstanding requests for high-performance disks (such as Fibre Channel storage devices) 	<ul style="list-style-type: none"> Add Physical Spindles: Most of the time it is a matter of adding physical disks to the disk group that is suffering. The Windows side of the story will vary depending on maintenance windows and so forth. The volume can be extended in Windows using Diskpart, because the new disk capacity will show up as free disk space at the end of the current disk device. The only other option is to backup or move the data, delete and then recreate the partition, making sure to align the partition using Diskpart unless using Windows Vista or later.
Flooding the I/O channel and causing retries or "busy" from the storage device	<ul style="list-style-type: none"> Windows may be sending too many I/Os at one time to the storage device, resulting in "BUSY" being returned by the storage device until buffers are freed. There is no easy way to determine if this is the case, with the possible exception of iSCSI and using Network Monitor. With direct attached storage you would need a bus analyzer or advanced tracing in order to find out if busy is being returned to the host. With Fibre Channel the only way to determine if a port is being overrun is with a Fibre Channel trace. 	<ul style="list-style-type: none"> Adjust the HBA Queue Length: Fibre Channel: Fibre Channel Host Bus Adapters (HBAs) have settings that control the number of outstanding I/O sent to the storage device. A Storport miniport HBA driver architecture provides a "per LUN" queue of 255 outstanding requests. Most HBAs have a default setting of 32, though OEMs can change this and often do, usually to 16. What we do here is check the current setting, find out how many disk (LUNs) we are sending I/O to, and calculate the overall I/O load. We can try reducing the queue length setting at the HBA and see if this helps improve response time. If lowering the queue depth helps, we know that we were sending too many I/O through the HBA. If this setting makes no change, then there is likely some other issue we need to look at.
Low throughput, high number of transfers	<ul style="list-style-type: none"> Examine the "Disk Bytes/Transfer", "Disk Bytes/Read", and "Disk Bytes/Write" counters. You may find that the overall request size is small, say 9 KB, or that writes are small (8 KB) and reads are as expected. The one thing lacking in Performance Monitor is to tell us where on the disk the bytes are being written. If the disks are constantly performing random reads or write to wildly different locations on the disk, the benefits of cache are lessened. 	<ul style="list-style-type: none"> Adjust Read/Write Ratio: It may be possible to adjust the read/write ratio settings on the storage subsystem cache. This is something the storage vendor would handle, not Microsoft. Adjust Short-Stroking: It may be possible to lessen the spatial effects of random data by a technique known as "short-stroking". What this means is that in Windows you create a partition that only fills up ½ of the available disk space. The physicals of the disks mean that average seek time will be lessened when there is less distance for the armatures & disk heads to cover. This does mean that some disk space will be sacrificed in order to gain performance. Adjust Element Size: If using RAID, the stripe-unit size, or element size, may not be sized for a smaller workload. The result is that not enough physical disks are involved in the workload. The storage vendors will normally have a handle on how to size the stripe units in the storage groups. If a storage group needs to have its stripe unit size resized, the group usually will need to be recreated. This means that the data will be lost, so be sure it is backed up first. Also, and this point cannot be stressed enough, be sure that the partition is created/recreated to a stripe unit boundary using Diskpart.

More Information

Perfmon Log Capture Interval: Generally speaking, if capturing performance data on a live system using the Windows Performance Monitor, the sampling interval should be kept fairly non-intrusive, such as every 10 seconds. The problem with sampling at 10 seconds or longer is that we tend to miss a lot of data. If in a testing environment we should set the capture interval to 1 second and capture both Physical Disk and Logical Disk counters. If we are capturing at short intervals like 10 seconds or less, we may not want to capture other counters at the same time so as to not impose too much overhead on the system for performance monitoring.

When capturing performance data, there is sometimes a concern about the size of the capture file. If capturing only Physical Disk and Logical Disk counters, even at a 1 second interval, the resulting file will not get to be excessively large. For a cost of 100 MB or so, and depending on the number of disk devices. If capturing only Physical Disk and Logical Disk counters, even at a 1 second interval, the resulting the counter log file will typically not grow, excessively large, perhaps 100 MB or so, depending on the number of disk devices.

References

- Ruling Out Disk-Bound Problems
<http://technet.microsoft.com/en-us/library/5bcdd349-dcc6-43eb-9dc3-54175f7061ad.aspx>
- How to Identify a Disk Performance Bottleneck Using the Microsoft Server Performance Advisor (SPA) Tool
[http://www.codeplex.com/PerfTesting/Wiki/View.aspx?title=How%20To%20Identify%20a%20Disk%20Performance%20Bottleneck%20Using%20SPA1&referringTitle=How%20To%20Download%20Details%20for%20Microsoft%20Service%20Performance%20Advisor%20\(SPA\)](http://www.codeplex.com/PerfTesting/Wiki/View.aspx?title=How%20To%20Identify%20a%20Disk%20Performance%20Bottleneck%20Using%20SPA1&referringTitle=How%20To%20Download%20Details%20for%20Microsoft%20Service%20Performance%20Advisor%20(SPA))
- Download Details for Microsoft Service Performance Advisor (SPA)
<http://www.microsoft.com/downloads/details.aspx?FamilyID=09115420-8c9d-46b9-a9a5-9bfcd237da2&DisplayLang=en>

Written By: **Robert Smith**

Contributors: **Clint Huffman, Jimmy May, and Ken Brumfield**