

Re: How to verify/fix High Disk Read Latencies in Exch2003 ?

Source:

<http://www.tech-archive.net/Archive/Exchange/microsoft.public.exchange.admin/2006-01/msg02429.html>

- *From:* "John Fullbright" <fullbrij@xxxxxxxxxxxxx>
 - *Date:* Fri, 20 Jan 2006 10:47:24 -0800
-

All mapped to LUNs from the same disk group on a Clariion. I pity you. Who did the SAN design; Dell or EMC? You'll probably what to have a long discussion with the Storage Architect. When you finally get sick of it, give NetApp a call.

The SMTP queue directories on a bridgehead actually get quite busy. There are the actual messages flowing through, and the metadata in alternate streams that gets touched each time message state changes. In "Optimizing storage for Exchange Server 2003" MS says "You should use a RAID-1+0 array with multiple disk spindles for the SMTP queue volume. The number of disk spindles and the size of the write cache should be based on the expected SMTP message throughput of the server. " Later in the paper they give the number 500 IOPS for the SMTP directory on a bridgehead.

What's the server? A Dell 28XX with a mirror and a RAID 5? This is the worst possible configuration. For one, when you build the server using Server Assist or similar automated build program, the primary partition starts out FAT and is later converted to NTFS. Whenever you convert a partition to NTFS, the cluster or allocation unit size is 512 bytes. If you look at the physical disk counter – Split IOs per sec, you'll see that they are excessive. When an IO request is larger than the allocation unit size, the NTFS driver splits the IO into multiple smaller IOs.

Run the numbers and you'll see that a 3 or 4 disk RAID 5 set is not going to come anywhere close to 500 IOPS. I always recommend directly creating the boot partition with a 4K allocation unit size, and you can infer from all my other post that I am no fan of RAID 5. I always recommend RAID 1/10/0+1.

I recently made this same set of recommendations to one customer, a public utility company, and they took me to task on it. They did a 125 page study that compared the performance of a "stock" 28XX and one that was configured with a RAID 0+1 with all six spindles and a 4K allocation unit size. The result was a 400% performance improvement over the stock configuration.

"Marlon Brown" <MarlonBrown@xxxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message news:epFJ2leHGHA.208@xxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Re: How to verify/fix High Disk Read Latencies in Exch2003 ?

- > The pre-production Exchange servers are mapped to same LUN in which the
- > Exchange production servers are, the SAN is an EMC CX600.
- > The Exchange 2003 SMTP Connector which got same 'high latency' message has
- > local SCSI disks.

>

> "John Fullbright" <fullbrij@xxxxxxxxxxxx> wrote in message

> news:O6RUhXdHGHA.3120@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

>> Do they all map LUNs to your SAN? It could be comingling or background
>> processes on the SAN (RAID rebuild, replication, SAN to tape backup etc)
>> generating IO that the host would normally not see.

>>

>> You never did mention the brand of SAN. I'm not sure about others, but
>> Network Appliance SANs running Data ONTAP version 7.0.2 or higher allow
>> you to access the performance counters on the SAN through perfmon. Just
>> open perfmon, click the add counters button, and in the "select counters
>> from this computer" dialog box type \\filename.

>>

>>

>> "Marlon Brown" <MarlonBrown@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message

>> news:errMUPdHGHA.3408@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

>>> Thanks John ! Let me set the perfmon counters very soon.
>>> However, one more observation is that MOM is returning the same High
>>> Latency alerts against two other Exchange pre-production servers, which
>>> are virtually with 0 mailboxes on it, no usage at all. Therefore I am
>>> wondering how the disk could be highly utilized if I have no users
>>> accessing the pre-production servers. It is also returning the same high
>>> latency against the Exch2003 SMTP Connector server, which is not
>>> clustered.

>>>

>>>

>>>

>>> "John Fullbright" <fullbrij@xxxxxxxxxxxx> wrote in message

>>> news:%23klicnYHGHA.1124@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

>>>> Being on a SAN doesn't exempt you from performance problems. I guess
>>>> the first thing to do is confirm that you have a performance problem
>>>> and determine how severe it is.

>>>>

>>>> First, over a period of a few days, collect permon counters. Key
>>>> counters to collect include physical disk – Avg. Disk sec/read, Avg.
>>>> Disk sec/write, Avg. Disk sec/transfer, reads/sec, writes/sec,
>>>> transactions/sec, and split IO/sec. You also want to include
>>>> Database – log record stalls/sec and MExchangeIS – Client Latency > 5
>>>> sec RPCs.

>>>>

>>>> The Microsoft Whitepaper, "Optimizing Storage for Exchange Server 2003"

>>>>

<http://www.microsoft.com/technet/prodtechnol/exchange/guides/StoragePerformance/fa839f7d-f876-42c4-a335-338>

>>>> is an excellent place to start when interpreting the data you collect.

>>>> The paper lists two specific criteria you can use to determine if your
>>>> disk subsystem is performing poorly:

>>>>

Re: How to verify/fix High Disk Read Latencies in Exch2003 ?

>>>> 1. Average read and write latencies over 20ms
>>>>
>>>> 2. Latency spikes over 50ms that last more than a few seconds.
>>>>
>>>> The Avg. Disk sec/read, Avg. Disk sec/write, and Avg. Disk sec/transfer
>>>> counters will give you the data to compare to this standard. I believe
>>>> something close to this was the basis of the MOM alert; PhysicalDisk:
>>>> Avg. Disk sec/Read: 0 C: value = 0.059590036231884. The average over
>>>> last 10 samples is 0.05959. Once you confirm a problem exists, you
>>>> need to determine what the impact of the problem is. This is where log
>>>> record stalls/sec and client latency > 10 sec RPCs come in. Log record
>>>> stalls happen when incoming data fills the log buffer in RAM to the high
>>>> water mark. A forced commit begins writing the log buffers in RAM to the
>>>> current log file on disk and continues until data in the log buffers
>>>> falls to the low water mark. During the forced commit, all client IO
>>>> is quiesced. This is known as a log stall. KB 328466 was one of the
>>>> first to actually define when log stalls are a problem. The criteria
>>>> in 328466 are:
>>>>
>>>> 1. Average value is more than 10 per second
>>>>
>>>> 2. Spikes (maximum values) are higher than 100 per second
>>>>
>>>> MExchangeIS – Log Record stalls/sec will give you the data to evaluate
>>>> these criteria. KB 839862 says:
>>>>
>>>> "When Outlook 2002 and later versions request data from an Exchange
>>>> Server computer, Outlook calls a function that wraps the RPC to the
>>>> server. This new wrapper is the Cancelable RPC wrapper. By default, the
>>>> Cancelable RPC wrapper starts a timer and issues the RPC. When the RPC
>>>> is complete, the wrapper closes the timer, cleans up, and quits.
>>>> However, if the RPC for data takes more than 5 seconds to return the
>>>> data, the wrapper produces the Cancel Request dialog box. The Cancel
>>>> Request dialog box remains on the screen until the RPC is answered or
>>>> until the user clicks Cancel. If the action that the user performs in
>>>> Outlook causes multiple RPCs to be made, the user may receive a Cancel
>>>> Request dialog box for each RPC."
>>>>
>>>> If we see a high number of client latencies over 5 seconds, clients are
>>>> definitely seeing the dreaded "requesting data" dialog box and calling
>>>> the helpdesk. MExchangeIS – client latencies > 5 sec RPCs will give
>>>> you the data to evaluate this criteria.
>>>>
>>>> Once you determine that you have a problem and the problem is impacting
>>>> users, it's time to look for the source. KB 839869 is somewhat helpful,
>>>> and does list a plethora possible causes, in my experience disk is the
>>>> culprit 99% of the time. Slow IO times will be obvious from the
>>>> sec/transaction, sec/read, and sec/write. If spikes in log stalls
>>>> correlate with spikes in slow disk access, the log stall is occurring
>>>> because the log buffers cannot be flushed to disk fast enough. If
>>>> there is no correlation, or only a weak correlation, it is most likely

Re: How to verify/fix High Disk Read Latencies in Exch2003 ?

>>>> the impact of large messages. You can mitigate the impact of large
>>>> messages somewhat by increasing the number of log buffers per storage
>>>> group or the number of storage groups (and thereby the number of log
>>>> buffers; log buffers are set on a per storage group basis) but no
>>>> amount of buffers will solve the problem if the storage is just simply
>>>> too slow.
>>>>
>>>> A storage subsystem is either intentionally or unintentionally designed
>>>> to support a specific IO load. The closer you get to the maximum IOPS
>>>> capacity, the longer it takes for each IO to complete. When the
>>>> average read and write latencies exceed 20 ms, we say the disk
>>>> subsystem is overloaded. If we believe the disk subsystem is
>>>> overloaded, we need to know how much load we are placing on it. Avg.
>>>> Disk reads/sec, writes/sec, and Transactions/sec for each LUN that
>>>> Exchange reads or writes to will tell us that. This includes the OS
>>>> volume, the location of the page file, the temp directory location, the
>>>> location of the smtp directories, the MTA, the logs, the databases, the
>>>> SRS database if you have one, the working directory, the MSSearch
>>>> Service gatherer logs, and the Message tracking logs. After
>>>> determining what load we are placing on the various components, we need
>>>> to figure out what load the storage location of these components is
>>>> capable of supporting. You can use the following formulats for the
>>>> RAID type of each LUN:
>>>>
>>>> P = perfomance of a single spindle. For 10K RPM SCSI drives at a
>>>> target 20 ms IO use 85, for 15K use 110.
>>>> N = number of spindles in the RAID set.
>>>> N' = number of data spindles in the RAID set
>>>>
>>>> For RAID 1/10/0+1
>>>>
>>>> read performance = P*N
>>>> write performance = P*N/2
>>>>
>>>> For RAID 5
>>>>
>>>> read performance = P*N'
>>>> write performance = P*N'/4
>>>>
>>>> (depending on the controller and the caching scheme used and the amount
>>>> of write cache, you may be able to use P*N'/3 for performance
>>>> calculation)
>>>>
>>>> For RAID 4/RAID DP (as Implemented by Network Appliance)
>>>>
>>>> read performance = P*N'
>>>> write performance = P*N'
>>>>
>>>> In all cases except RAID 4/RAID DP, you will need to determine your
>>>> read/write ratio in order to apply the "write penalty. Use the
>>>> reads/sec, writes/sec and transactions/sec to determine your ratio.

Re: How to verify/fix High Disk Read Latencies in Exch2003 ?

>>>> For example:

>>>>

>>>> A RAID 5 array consisting of 3 spindles.

>>>> A workload with a 3:1 read/write ratio.

>>>>

>>>> Write performce = $85 * 2 = 170$ IOPS

>>>> Read performance = $85 * 2 / 4 = 42.5$ IOPS

>>>> Workload performance = $(.75 * 170) + (.25 * 42.5) = 127.5 + 10.625 =$

>>>> 138.125 IOPS.

>>>>

>>>> You can now determine if the load you are placing on the disk subsystem

>>>> is at or over the IOPS capacity of the disk subsystem. This works well

>>>> for direct attached sorage, but you mentioned a SAN. When you

>>>> consolidate storage on a SAN, you have to be very careful to minimize

>>>> the impact of comingling. Comingling occurs whe two or more LUNs

>>>> reside on the same set of physical spindles, and heavy IO activity

>>>> against one LUN negatively impact IO activity on another LUN that

>>>> shares the same set of spindles. Heavy database activity could impact

>>>> the logs if both LUNs share the same set of physical spindles.

>>>> Likewise, heavy IO against on server can impact another unrelated

>>>> server if both servers have disk whose LUNs reside on the same set of

>>>> physical spindles. Unless the SAN vendor has a feature to limit the IO

>>>> against comingling LUNs, you only real choice is physical isolation of

>>>> the spindles. This is the basis of the MS recommendation that log

>>>> files and databases reside on seperate spindles. The smoke and mirrors

>>>> of virtualization can result in unintended and difficult to

>>>> troubleshoot comingling situations. The only vendor that I am aware of

>>>> that has a feature to limit the impact of comingling is Network

>>>> Appliance. In Data ONTAP 7.X there is logice to distribute IO between

>>>> Flexvols that exist within a given Aggregate.

>>>>

>>>> Well, it's not quite a troubleshooting guide, but hopefully it will get

>>>> you going in the right direction.

>>>>

>>>> John Fullbright

>>>>

>>>>

>>>>

>>>>

>>>>

>>>> "Marlon Brown" <MarlonBrown@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message

>>>> news:uv4ruCVHGHA.1032@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

>>>>> The Troubleshooting Analyzer tool just showed one single user which

>>>>> presents higher RPC latency than normal, what doesn't indicate a broad

>>>>> problem in the server according to the report.

>>>>>

>>>>> Can you confirm whether high Physical Disk/%Disk Time is a good

>>>>> indication of bottlenecks ?

>>>>>

>>>>> "Marlon Brown" <MarlonBrown@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in

>>>>> message news:OTPdvwUHGHA.524@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

Re: How to verify/fix High Disk Read Latencies in Exch2003 ?

>>>>>> Darn. I think this is an actual issue.
>>>>>> I am running Perfmon Physical Disk/% DiskTime:
>>>>>> The partition in which the information store is mounted shows
>>>>>> %DiskTime utilization of 80–100 steady. I will run the
>>>>>> Troubleshooting Analyzer to see what I get.
>>>>>>
>>>>>> The servers have the databases installed on a SAN drive, Raid 1+0,
>>>>>> with 143GB free disk space (70% free disk space). Not sure what could
>>>>>> be causing such latency...
>>>>>>
>>>>>>
>>>>>> "Andy David – MVP" <adavid@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in
>>>>>> message news:5flvs1lebbmao8t6ps1eufgmt1rncukoud@xxxxxxxxxxx
>>>>>> On Thu, 19 Jan 2006 09:58:04 –0800, Marlon Brown
>>>>>> <MarlonBrown@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote:
>>>>>>
>>>>>>>MOM2005SP1 keeps warning that various Exchange 2003 Servers
>>>>>>>experience the
>>>>>>>issue below:
>>>>>>>
>>>>>>>High Disk Read Latencies for the past 10 minutes
>>>>>>>PhysicalDisk: Avg. Disk sec/Read: 0 C: value = 0.059590036231884.
>>>>>>>The
>>>>>>>average over last 10 samples is 0.05959.
>>>>>>>
>>>>>>>Can you indicate what's the best way to troubleshoot and fix this ?
>>>>>>>I think the first step would be by running Perfmon and setting
>>>>>>>manual Disk
>>>>>>>counters to see whether this is an actual problem. Can you point me
>>>>>>>a
>>>>>>>document or tell me the counter parameters I should be looking for
>>>>>>>to
>>>>>>>determine whether this an actual issue ? Also, if you know how to
>>>>>>>improve
>>>>>>>this in case it is an actual issue, I would appreciate your input.
>>>>>>>
>>>>>>>
>>>>>>> Start here:
>>>>>>>
>>>>>>><http://www.microsoft.com/Downloads/details.aspx?familyid=4BDC1D6B-DE34-4F1C-AEBA-FED1256CAF9A&d>
>>>>>>>
>>>>>>> Microsoft Exchange Server Performance Troubleshooting Analyzer Tool
>>>>>>> v1.0
>>>>>>>
>>>>>>>
>>>>>>> Make sure you have the latest firmware and updates for your Raid
>>>>>>> Controllers and disks.
>>>>>>>
>>>>>>>
>>>>>>> Have you sized your servers appropriately?
>>>>>>>

