

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

Source:

<http://www.tech-archive.net/Archive/Exchange/microsoft.public.exchange.design/2007-08/msg00005.html>

- *From:* "John Fullbright" <fjohn@donotspamenetappdotcom>
 - *Date:* Fri, 3 Aug 2007 12:33:55 -0700
-

If I get this right,

1. You are assuming .85 IOPS per user
2. You will not be using caching on outlook clients
3. You're assuming a read/write ratio of 3:1
5. In an A/A/A/P cluster, each active node will host 8000 or so mailboxes
6. You're assuming 50% concurrency
7. This is an Exchange 2003 Design
8. You're active user count is above 100%

Take a look at Optimizing Storage for Exchange Server 2003. Start about page 19. If you're going to assume load instead of measuring (I wouldn't recommend that, but just for the sake of argument), then you need to understand the impact that scalability will have on that assumed IOPS/user number. You'll notice that as the number of users increases you multiply the base IOPS/user number by some offset. This is because the more users you have on a server the less database cache there is per user. In exchange 2003 with 4GB of RAM and the /3gb switch in the boot.ini, you have something like 898MB of database cache. If I pile 8000 users on a single EVS, even at that mythical 50% concurrency(which I have to assume to meet the recommended max concurrent users for a single node). That gives me about 229K of cache per active user. That amount of cache isn't even close enough to cache the 11 views for the default folders. If you're going to scale to that level, client side caching is a must. In your case, you're measuring upwards of 10000 active users. There are many potential reasons for this. To name a few:

1. Your concurrency assumption is wrong.
2. Users may be connecting to Exchange from more than one device
3. Blackberry, goodlink, etc are essentially an extra mapi session (for BB that sees is at 3.64 times the load of a normal session by the way)
4. Users are disconnecting and reconnecting multiple times within the session timeout (20 minutes or so)

Any way you look at it, cache is still used per session and you're looking at 10,000 active sessions. From your measurement (not your consultant's assumption; there's a big difference) you'll get more like 91K of cache per

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

user. You're off the end of the tables in "Optimizing Storage for Exchange Server 2003" To get to 8000 users, following the MS guidance on storage group layout, you have 4 storage groups and I'd assume 4 stores in each storage group. At the very minimum, using the table on page 20 you should add 38% to that .85. The size of the mailboxes is also a factor that could increase that IOPS/user number further. Using outlook in cached mode shifts many of the read operations to the client and will help. Typically the read/write ratio drops from about 3:1 to about 2:1 when you go to cached mode on all the clients. That's a 25% reduction in IO, although all of it is read reduction. On the extra client connection front, if you're clients are using desktop search engines then this will shift the IO from the Exchange server to the desktop where it belongs. Blackberry polls, so it won't help you there.

When calculating the write penalty for RAID 10, each read requires one operation and can occur from either side of the mirror, and each write requires 2 operations, one to each side of the mirror. That's where $P*N$ and $P*N/2$ come from. You then apply the read write ratio to determine the number of composite IOs (mixture of reads and writes at the specified read/write ratio) an array will support. The difference between this route and $\text{Penalty} = (R + W)/(R + 2W)$ is well... If I have 100 and subtract 20 percent, then I have 80. I can't simply add 20 percent of back to get 100; I end up with 96 if I try that. The correct way is to add 25% of 80 back to 80 to reach 100 where I started. If I assume 2 spindles at 130 IOPS/spindle and a 2:1 read write ratio, then with $\text{Penalty} = (R + W)/(R + 2W)$ I get 195. If I figure out writes supported and reads supported then apply the read write ratio, then I get 216.67. In addition, the math for RAID 5 in the paper you cite neglects to subtract out the parity spindles, skewing the results. For example, in a RAID5 vraid volume on an EVA, 1 out 6 stripes is parity. Read/write ratios are another potential pitfall; a small change can make a big difference in the IO that hits the storage. When dealing with Netapp storage, it's a whole different set of math. There is no write penalty which makes things much simpler.

"TonyP" <TonyP@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message
news:D7370B76-8263-4A3E-825B-D7FA76C5584C@xxxxxxxxxxxxxxxxxxxx

Hi John

I added some comments etc below on the areas you have mentioned I am unsure about:

"John Fullbright" wrote:

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

For shared disks:

1. Put the DTC and the quorum on the same drive.
2. Put the databases for each storage group on a drive (4 storage groups 4 drives). This will give you a granularity of restore at the SG level.
3. Put the transaction logs for each storage group on their own drive (4 storage groups 4 drives).
4. Put the SMTP directories on a drive. It's important for the SMTP queues to be on shared disk in 2003. If you leave it local, every time you fail over you will strand messages. When you failback they will mysteriously reappear.. maybe days or months later.
5. The MTA run and database directories are on the first database drive path by default. Unless you have some abnormally high MTA activity (mixed mode with 5.5 and this is a bridgehead for example) leave it there.

Ok, you have a theoretical .85 IOPS/user. Measure a test group or a subset of your user base. 50% concurrency? The concurrency ratio is a very common pitfall. Unless your users are on shifts, in the air and can't access a computer, etc., you'll get burned. Along will come the end of the quarter, everyone will be burning the midnight oil, you'll have 90% of your user on at once, performance will dive, and key staff will miss end of quarter reporting deadlines. Result: you'll be looking for a new job. Assume 100% unless it's physically impossible. Measure (over a long period – several quarters – and use the peak value+ 20%) if you'll be using anything less. Read/write ratio is important also.

Concurrency is an Issue, they are assuming 50 percent based on a present 4–node AAAP cluster they have which hosts roughly 25,000 users , 8000 users per node. They are informing me that only 50 percent are concurrent at any one time.

I have monitored a series of Perfmon counters on a montly basis and on a daily basis also when I look at the MExchangeIS counter "Active User Count" on the busiest node it approaches 10,000 users even though the node holds only 8000 users.

I am assuming 100 percent concurrency for the Storage Design.

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

Internal Technical lead in the team is providing input saying it is MExchangeIS counter "User Account" which is important showing 50 percent concurrency?

They are saying some users are showing more than one session to the information store?

Hence they are saying User Account is a more accurate figure to use then Active User Account which shows some users have more than one connection?

RAID 10 has a write penalty of 2, so the impact of increases in the percentage of writes is amplified. Assume 2:1 in Exchange 2003 with Outlook cached mode clients. Assume 3:1 if the outlook clients are not cached. Make sure you add in IOPS for online maintenace and backups.

Whe you build your arrays, make sure the IOPS/spindle number is @ 20ms response time or less. Sure, a 15K spindle can reach a maximum IOPS of 300 or so, but the response time at that level can me measured in seconds. You want an average response time less than 20 ms with no spikes greater than 50ms lasting more than a few seconds. For 4K random IOs, use the following:

10K RPM SCSI – 90 IOPS/spindle @ 20ms response time
15K RPM SCSI – 130 IOPS/Spindle @ 20ms response time
7200 RPM SATA – 40 IOPS/Spindle @ 20 ms response time

Where P is the performance of a single spindle, and N is the number of spindles in the RAID set, for Raid 1/10,

Read performance = $P*N$
Write performance = $P*N/2$

So if I have 4 10K SCSI drives in a RAID 10 array,

Read performance = 360 IOPS
Write performance = 180 IOPS

Applying a 2:1 read write ratio, the composite performance is $(360+360+180)/3 = 300$ IOPS.

NOTE: Just as a comparative reference point, a RAID 5 array with the same 4

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

disks would have a composite performance of 201 IOPS; that's why you don't use RAID 5. At a 1:1 read/write ratio, RAID 5 has less than half the performance of RAID 10, so don't consider it in an Exchange 2007 solution either.

I am VERY confused on this area about COMPOSITE performance I don't know the number of spindles I require YET?

I am trying to work out the number of drives (spindles) required to meet my performance needs?

John I used this article before your post and followed out the below:

http://www.petri.co.il/sizing_exchange_part_2.htm

(number of disks) = (IOPS/mailbox × number of mailboxes) ÷ (IOPS/disk × RAID penalty factor)

Raid 10 Penalty = (R + W)/(R + 2W)

Again since I have no sound statistical data due to latency on the present 4-node cluster I will assume 3 Reads for every 1 Write since Outlook clients are not cached

Raid 10 Penalty = (3+1) / (3+2(1))

Raid 10 Penalty = 0.8

Hence

So each Storage group which host 1875 users will need

= IOPS/mailbox * number of mailboxes
= 0.84 * 1875
= 1575 IOPS

Recommended to handle spikes we add a 20 percent buffer to the storage design to handle these peaks:

Peak Storage Group DB IOPS

= 1575 * 20%
= 1890

Now standard calculation:

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

(Number of disks) = (IOPS/mailbox * number of mailboxes) / (IOPS/disk at 20ms * RAID penalty factor)

Number of disks = $1890 / 130 * 0.8$

Number of disks = $1890 / 104$

Number of disks = 18.18

Since we are using Raid 10 we must round up to the nearest even number.

Number of disks = 20

Thus

Number of disks required per Storage Group to host 1875 users is 20 15K RPM
SCSI Drive in Raid 10

Database Storage Group size = number of users * mailbox size

Client has defined the mailbox size to 180MB.

Hence

Database Size = $1875 * 180 \text{ MB}$
= 329.59 GB

So each Database Storage Group is required to be no bigger than 330 GB

Disks are 146GB in size recommended by HP

Total Storage generated to accommodate our Performance for the Storage Group comes to

Total Storage for Performance = $10 \text{ disk} * 146 \text{ GB}$
= 1460 GB

Note: 20 Disk in Raid 10 to meet IOPS requirement, hence 10 discs available for storage

Our previous Capacity figure above suggests we only need 330 GB per Storage Group.

But assigning 1460 GB for each LUN using traditional Storage methods leads to a huge waste in disk storage space.

Virtualized Storage techniques we can create 4 LUNS from our 10 Disks:

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

= 1460 GB / 4
= 365 GB – size of each LUN

Hence we meet our Capacity requirement since each LUN created is greater than 330 GB and also we effectively use our 10 disks more efficiently.

Previous figures we calculated for performance related IOPS was based on all physical spindles within each LUN created dedicated to the Storage Group.

Since the physical spindles are NOT now dedicated to a single LUN but are shared amongst 4 LUNs is a loss in Performance IOPS?

So is each LUN carved from the array set now is not giving me the performance I defined for 1875 users – 1890 IOPS??

Is this a case of Comingling where IO against one LUN negatively impacts the performance of other LUNs that share the same physical spindles?

I am now seriously concerned about my reasoning since you have talked about "P is the performance of a single spindle, and N is the number of spindles in the RAID set" and working out composite performance etc???

To determine the IOPS for the transaction logs, you divide the database IOPS by anywhere from 8 to 12, with 10 being common. I tend to use the 8 figure to stay on the conservative side. The size of the log lun is another story. What is your average 24 hour change delta. What is the peak? If you collect change delta information over a period of time, what is the mean and the standard deviation of the dataset?

How are you collecting this data? Perfmon counters?

How often do you do a full backup and truncate the logs? What is your backup failure tolerance in days (how long should the system stay operational if backups starts failing? Generally 4–7 days to cover a long weekend and troubleshooting)? What level of reliability is required? The answers to these questions will tell you

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

how

large to make the log LUNs. For example, let's assume:

1. Mean change delta 9GB
2. SD of sample set 1GB
3. Backup failure tolerance 7 days.
4. 99.9% reliability

We start with the mean change delta, then add enough standard deviations to

reach or exceed the required level of reliability (3 in this case), so

our

Change delta size is $9+(1*3) = 12\text{GB}$. Now, we take this figure and multiply

by our backup failure tolerance and our LUN is 84GB. I can say with 99.97559% accuracy that an 84GB LUN will withstand 7 consecutive days of backup failure before the drive fills and the stores dismount.

You can take a similar statistical approach to sizing the SMTP LUN; take

a

sample set of max size of a long collection period.

How are you collecting this data? Perfmon counters?

Figure out the mean and SD, then add enough standard deviations to the mean to reach the desired

level of reliability. A lot of folks don't bother, and just allocate an overly large disk (50 – 100GB) to cover normal traffic and any potential loops/chain mails/store outages/etc.. without going offline. I believe

Optimizing Storage for Exchange Server 2003 says 500 IOPS, however, I would

measure. The number of IOs depends on the number of messages, message sizes, destination, retries, etc. On average, the categorizer touches

an

eml file in the queue directory 7 times.

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

"TonyP" <TonyP@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message
news:3A3EFF63-5E9C-4E58-996F-CF2137F7D4FB@xxxxxxxxxxxxxxxxxxxx

Hi

Currently trying to design a 3 node cluster comprising of 2
Active
nodes
and
1 Passive node. Exchange 2003 environment

Will have 4 Storage Groups per Node which will have there
only
dedicated
drives.

Transaction logs for each transaction drive will also have
there own
drive
letters

As will SMTP , MTA and Quorum drives all on separate
drive letters will
LUNs
carved out of the SAN

Used a theoretical value for IOPS per user as 0.85 and user
mailbox
limits
where decided as 180 MB.

Each Storage Group will hold 1875 users, did the standard
calculation
to
to
work out size for each Database drives. Hence each node
holds 7500
users
but
there is only 50 percent concurrency.

Database Drives are in RAID 10

Transaction drives where taken as 1/10 of IOPS requirement
of Database
Drives , will also be in RAID 10

Re: Sizing Exchange Transaction, SMTP, MTA and Quorum Drives???

How do you determine a safe size for the Transaction Log drives?

Also what is the standard calculation to work out

SMTP drive size?

MTA drive size?

Quorum size ?

Would you use Raid 1 for the SMTP, MTA and Quorum?

what about IOPS for

these

also?

help greatly appreciated as always

Tony