

Re: Huge Av Disk Queue Length

Source:

<http://www.tech-archive.net/Archive/Exchange/microsoft.public.exchange.admin/2007-01/msg00728.html>

- *From:* "John Fullbright [MVP]" <fjohn@donotspamnetappdotcom>
 - *Date:* Sun, 7 Jan 2007 15:47:41 -0800
-

If we assume 15k SCSI drives, 130 IOPS @ 20ms response time

Read Performance = 260 per mirror

Write performance = 130 per mirror

Mixed performance at a 1:1 R:W ratio = 195 IOPS per mirror.

With 150 users total, that works out to about 10 users/store or mirror.

There are a lot of indicators of heavy use, 500mb + mailboxes, and Cisco unity.

It seems that most stores are ok, except the one that hosts Unity.

One problem with the storage design is that it has gone too granular, and a lot of IOPS capacity is wasted. In a Netapp design, this would be avoided by creating two aggregates (pools of disk) and using one for logs and one for everything else. If, for example, I assume 5 IOPS/user in lieu of measurement, then with 150 users my database IOPS requirement is $5 * 150 * 1.2$ or 900 IOPS. Logs would need an additional 900/8 or 120ish IOPS. Then there's SMTP LUNs etc...

If I use the current 32 spindles dedicated to databases, and a RAID DP raid group size of 16, then My database aggregate has 28 data spindles and can support 3640 IOPS (there is no write penalty on netapp thanks to the virtualization layer WAFL). From this aggregate, you carve volumes (a slice across all the spindles in the aggregate) and in each volume you can have one or more LUNs. I prefer a 1:1 mapping of volumes to LUNs in that it preserves the greatest flexibility in backup/restore (you can do a clone split, a single file snap restore or a volume snap restore this way.) There should be enough excess IOPS capacity in the aggregate (3640 - 900 = 2740 IOPS) to handle any spikes from unity. If you wish to add reliability, you can use the Data Ontap 7.2 feature "Flexshare" to set priorities on volumes (sort of like weighted fair queuing for disk performance).

As for what FAS system to use - it depends on the space requirement. From a performance point of view, a single shelf in a FAS270 would meet the requirement. If you can go larger disks and meet the space requirement, this would be sufficient. Certainly, at the most you'd use a 3020. The

Re: Huge Av Disk Queue Length

3000 series frames can support SATA or FCP shelves. If you have more than one location, and are mirroring snapshots between the locations, you would put a 3020 in each location with FCP shelves for the primary data and SATA shelves for the replication targets.

"Kirill Palagin" <kpalagin@xxxxxxxxxxxxxxxxxxxx> wrote in message
news:%23G05%23jcMHHA.4384@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

Mark Arnold [MVP] wrote:

Yes indeed.

Looking at the OPs disk configuration I'm stunned that he gets any kind of performance whatsoever out of it.

You (JustCol) need to get away from this hideous waste of disk.

Mirroring is a nice efficient method up to about three or four pairs of disks with a couple of SGs. Once you get up to 16 stores you really have to get involved in, at the very least, an iSCSI SAN. You're suffering from something I have recently helped a customer resolve; i.e. they were badly disk bound. They had the same configuration as you, hundreds of BlackBerry users, and awful disk performance. They now have very underutilised servers and are actually going to consolidate because they have reduced their previously busy servers to a virtually idle state.

Mr Fulbright won't tell you because modesty forbids but what you need is a FAS3020 from his employer (NetApp) and a couple of shelves of disks. At the very least you want a FAS270 which is a entry SAN with integrated disk shelf. Obviously you can choose from any vendor but you can take those models and translate them into something else.

If you simply cannot go with a separate storage solution then a complete redesign of your environment is necessary. Give the logs the usual set of RAID1 but look at a couple of RAID5 arrays with hot spares. Your writing is a major sequential load but your reading is going to be very much more random, favouring the NetApp RAID-DP (an implementation of RAID6) that will blow you away. Get someone who actually knows what they're doing rather than a conventional Exchange architect (who will actually cause more problems than solve)

Speaking personally I'd hold off on actually performing the tests that Dave from Microsoft suggests because your storage is just so plainly wrong for you. He, again, won't say as much because he has to maintain a certain detachment but believe me, your storage is very very wrong at the moment. Get the software and check out what it does, but don't act on anything yet until you've spoken to someone. (If you're in the UK you can email me and my company might offer me up at reduced rates)

Re: Huge Av Disk Queue Length

LOL!!

If you were a politician, I would vote for you instantly after such speech. :-)

But how come 32 spindles can't support 150 users?

(At the very least each disk can do 100 IO op/s, combined (with 1:1 read:write ratio) – 2400 IO op/s.

Heavy users generate 2 IO op/s, 150 users will generate 300.)

Does voicemail generate such a big load?

—

If my message is helpful, please help me by registering at <http://www.openoffice.org/servlets/Join> and voting for the following issues:

http://www.openoffice.org/issues/show_bug.cgi?id=24969

http://www.openoffice.org/issues/show_bug.cgi?id=29807

http://www.openoffice.org/issues/show_bug.cgi?id=51564

http://www.openoffice.org/issues/show_bug.cgi?id=70753

http://www.openoffice.org/issues/show_bug.cgi?id=15220

http://www.openoffice.org/issues/show_bug.cgi?id=10931

http://www.openoffice.org/issues/show_bug.cgi?id=35579

http://www.openoffice.org/issues/show_bug.cgi?id=32785

http://www.openoffice.org/issues/show_bug.cgi?id=1035

http://www.openoffice.org/issues/show_bug.cgi?id=67838

http://www.openoffice.org/issues/show_bug.cgi?id=39527

http://www.openoffice.org/issues/show_bug.cgi?id=64785

Thank you very much!