

Storage - the final frontier

(c) Colin Leversuch-Roberts April 2009



The slide includes a small logo of a knight on a horse in the top left corner. The text identifies the speaker as "Colin Leversuch-Roberts" with the nickname "aka Grumpy Old DBA". It lists the company "Kelem Consulting Limited" and provides three website URLs: "www.kelemconsulting.co.uk", "http://sqlblogcasts.com/blogs/grumpyolddbba/", and "www.grumpyolddbba.co.uk". A copyright notice is at the bottom.

The slide lists ten topics for the presentation, each preceded by a square bullet point. The topics cover various storage concepts and operational concerns. A copyright notice is at the bottom.

Storage is a very "dry" subject so I will do my very best to avoid having the audience nod off.

Storage is key for a well performing SQL Server Database, the impact of storage performance or lack of it, will increase as your database/server becomes larger and/or performance increases are required.

There's not a scale of size to which any of this applies, in my experience applications face many challenges and I will consider any attempts at generalisation with a fair level of contempt, as readers of my blog may have noted in my post concerning parallelism.

What I'm hoping to achieve is to relate some of the aspects of storage to SQL Server, I'm not interested in any other areas of storage here.

It is not my intention to delve into detail of vendor specifics with storage, most vendors offer the same features for their storage, the price of these features may vary widely dependent upon the actual storage platform/product.

Although I've arranged topics into an order in actual fact there is a fair degree of overlap in subject matter so there will be references which will be clarified later.

In as much as I can I have scripted the presentation into a document you can download - there's a lot of stuff I hope you'll find useful and I wanted to keep the slides simple.



ATWAA - All The World's An Acronym

It is absolutely critical to know your acronyms to be taken seriously in storage and you should be able to add the latest hot buzzwords into the mix to impress listeners.

In the best traditions of the TV reality show, in no particular order :-

ATA	AT Attachment	
IDE	Integrated Drive Electronics	
SAS	Serial Attached Scsi	Actually the interface or bus for moving data. Also used to describe disks.
SATA	Serial ATA	Another bus. Replacement for Parallel ATA (PATA) mainly for PC's.
SCSI	Small Computer System Interface	Parallel bus. Server technology. Being replaced by SAS
FC	Fibre Channel	Networking technology for storage
ISCSI	Internet SCSI	The ability to network SCSI mainly for storage.
SAN	Storage Area Network	Note the term Network.
NAS	Network attached Storage	Shared storage placed upon a network (essentially iscsi)
HBA	Host Bus Adapter	Usually refers to a Fibre Channel Network Card.
RAID	Redundant Array of Inexpensive Disks	
SNIA	The Storage Networking Industry Association (2)	
JBOD	Just a bunch of disks	(a storage unit with no assigned arrays is just a bunch of disks)
DAS	Directly Attached Storage	Generally refers to storage dedicated to one host
FCoE	Fibre Channel over Ethernet	Piggy back FC onto Ethernet for wide area storage networking.
SSD	Solid State Disk	No moving parts and lower power.
HPC	High Performance Computing	Big Fast Servers.
PATA	Parallel ATA	aka IDE on home PCs





What's a BUS ?

Serial Attached SCSI (SAS)

Uses the standard SCSI command set .

Is point to point.

Full Duplex

3 Gbit/s, 6 Gbit/s now available, 12 Gbit/s coming.

16k devices possible

Supports 3 Gbit/s SATA disks on same backplane.

8m cable length.

Multipath support.

(1)

Serial ATA (SATA)

Replacement for PATA which only supported 2 devices.

Almost unlimited number of devices.

SATA 1.5 Gbit actually no better than PATA

SATA 3.0 GBit also wrongly called SATA 2

Lower cabling cost. point to point.

Infinband

High speed point to point switched interconnect.

Full Duplex

Low latency

Up to 120 Gbit/s

Designed (?) for HCP

Multipath

SCSI

Original parallel interface for storage.

Fractionally faster than 3Gbit SAS

16 device limit

Replaced with SAS

Legacy storage choice

Fibre Channel

Storage Network

Used as arbitrated loop in FC storage arrays

FC is the solution of choice for most companies.

4 Gbit FC disks available

switched network fabric

Full duplex

Multipath

Bandwidth, how does it affect me?

Device	Bits	Bytes
Ultra DMA ATA	528 Mbit/s	66 MB/s
Ultra-2 wide SCSI (16 bits/40 MHz)	640 Mbit/s	80 MB/s
Ultra-3 SCSI (Ultra 160 SCSI; Fast-80 Wide SCSI) (16 bits/40 MHz DDR)	1,280 Mbit/s	160 MB/s
Ultra-320 SCSI (Ultra4 SCSI) (16 bits/80 MHz DDR)	2,560 Mbit/s	320 MB/s
Serial ATA 2 (SATA-300)[36]	2,400 Mbit/s	300 MB/s
Serial Attached SCSI (SAS)[36]	2,400 Mbit/s	300 MB/s
Serial Attached SCSI 2[36]	4,800 Mbit/s	600 MB/s
Fibre Channel 4GFC (4.25 GHz)[35]	3,400 Mbit/s	425 MB/s
Fibre Channel 8GFC (8.50 GHz)[35]	6,800 Mbit/s	850 MB/s
iSCSI over Gigabit Ethernet	1,000 Mbit/s	125 MB/s
iSCSI over 10G Ethernet	10,000 Mbit/s	1,250 MB/s
100G Ethernet		12,500 MB/s
12 x Infiniband		

© Cole Lennsch-Roberts Kalam Consulting Limited March 2008

Bandwidth - How does it affect me then ? (3)

Device	Bits	Bytes
Ultra DMA ATA	528 Mbit/s	66 MB/s
Ultra-2 wide SCSI (16 bits/40 MHz)	640 Mbit/s	80 MB/s
Ultra-3 SCSI (Ultra 160 SCSI; Fast-80 Wide SCSI) (16 bits/40 MHz DDR)	1,280 Mbit/s	160 MB/s
Ultra-320 SCSI (Ultra4 SCSI) (16 bits/80 MHz DDR)	2,560 Mbit/s	320 MB/s
Serial ATA 2 (SATA-300)[36]	2,400 Mbit/s	300 MB/s
Serial Attached SCSI (SAS)[36]	2,400 Mbit/s	300 MB/s
Serial Attached SCSI 2[36]	4,800 Mbit/s	600 MB/s
Fibre Channel 4GFC (4.25 GHz)[35]	3,400 Mbit/s	425 MB/s
Fibre Channel 8GFC (8.50 GHz)[35]	6,800 Mbit/s	850 MB/s
iSCSI over Gigabit Ethernet	1,000 Mbit/s	125 MB/s
iSCSI over 10G Ethernet	10,000 Mbit/s	1,250 MB/s
100G Ethernet		12,500 MB/s
12 x Infiniband		

- For ballpark figures on usable bandwidth, e.g. how long to copy a backup;
- divide the Gbits / 10 and then take 80% of that figure for usable bandwidth.

e.g.

10 Gbit Ethernet = 800 Mb/sec , time to copy 1TB backup file assuming disks can support read/write rate > 800 Mb/sec, approx 22 minutes. (nearly 4 hours on 1Gb Ethernet)

Bandwidth is all about performance, you need sufficient bandwidth to support the storage subsystem and you need sufficient storage to support the bandwidth.

Implementations of iscsi or NAS storage over 1GB Ethernet were obviously slow, Ethernet being three times slower than standard SCSI and here the bandwidth quickly became saturated.

I'll return to the subject of bandwidth in later discussions.

Disks – big isn't always beautiful



Rotational latency	How long it takes for a revolution of the disk e.g. 15k disk = 15,000 rotations/min $15,000/60 = 250$ rotations/sec $1,000/250 = 4.0$ ms The average is 50% of the full rotation = 2.0 ms
Track to Track Seek	How long it takes the head to move from one track to another Sequential disk access Again this is an average
Average Seek	How long it takes to find data on the disk Random access Again this is an average

How to calculate io for a disk.

IO calc is 1 second divided by the rotational latency plus the seek time

(You should then apply the 85% factor for a reasonable working figure)

Disk Type	Latency	Average Seek Read/Write	Track to Track Read/Write	Random Read/Write	Sequential Read/Write
15k 2.5" SAS	2.0	2.9 / 3.3	0.2 / 0.4	204 / 188	454 / 416
72k 3.5"	4.16	8.5 / 9.5	0.8 / 1.0	79 / 73	201 / 194
10k 3.5"	3.0	3.8 / 4.4	0.2 / 0.2	147 / 135	312 / 312
15k 3.5"	2.0	3.5 / 4.0	0.2 / 0.4	182 / 167	454 / 416
10k 2.5" SAS	3.0	3.6 / 4.2	0.2 / 0.3	151 / 138	312 / 303

Disks - Big isn't always better.

I don't want to dwell too much on actual hard disks because there is a danger of being seriously sidetracked.

Essentially all disks are the same, to clarify, a 300GB disk is a 300GB disk, it's only the interface which changes, so the same disk will come with SAS, SCSI, SATA, FC interface; the basic characteristics of the disk will remain the same.

The actual sustained throughput for a disk is usually unaffected by the actual interface, thus a 4GB FC disk cannot write to the disk quicker than the same disk with 3GB SCSI interface

What concerns the DBA - spindle speed, there's no rocket science to this, the faster the spindle speed the lower the latency (response) and the higher the supported io.

There are some myths, sata disks do fail more often than sas/fc/scsi disks, that's why they are cheaper. LSI have some research into disk failure rate, available on their web site, what it showed was that when sata disks were used for infrequent data access the failure rates were similar to fc disks, however, when sata disks were used for online data access (they said "incorrectly") then the failure rate was over double that of a fc disk.

So why isn't big better? Well performance is achieved with multiple spindles not by capacity, so to get 2 TB of storage which is best, ignoring raid ?

- 1 x 2TB disk
- 2 x 1TB disks
- 7 x 300GB disks
- 28 x 73GB disks

The 28 disks will be the best solution by far although it will be the most expensive

Now the other point about disks is that all the heads move together, hold onto that thought for later.

How to calculate io for a disk.

Rotational latency	How long it takes for a revolution of the disk e.g. 15k disk = 15,000 rotations/min $15,000/60 = 250$ rotations/sec $1,000/250 = 4.0$ ms The average is 50% of the full rotation = 2.0 ms
Track to Track Seek	How long it takes the head to move from one track to another Sequential disk access Again this is an average
Average Seek	How long it takes to find data on the disk Random access Again this is an average

IO calc is 1 second divided by the rotational latency plus the seek time

(You should then apply the 85% factor for a reasonable working figure)

(Times in ms)

Disk Type	Latency	Average Seek Read/Write	Track to Track Read/Write	Random Read/Write	Sequential Read/Write
15k 2.5" SAS	2.0	2.9 / 3.3	0.2 / 0.4	204 / 188	454 / 416
7.2k 3.5"	4.16	8.5 / 9.5	0.8 / 1.0	79 / 73	201 / 194
10k 3.5"	3.0	3.8 / 4.4	0.2 / 0.2	147 / 135	312 / 312
15k 3.5"	2.0	3.5 / 4.0	0.2 / 0.4	182 / 167	454 / 416
10k 2.5" SAS	3.0	3.6 / 4.2	0.2 / 0.3	151 / 138	312 / 303

- If you're trying to convince yourself that SATA disks are viable for your server then you might want to think again.
- Note that the 10k raptor SATA disks have typically the same performance as the 3.5" 10k disk
- You get what you pay for

What's IOPS then guv?

Be very wary, it's a measure but largely meaningless

Response Time - Throughput Data

By Request Percentage	10% Load	50% Load	80% Load	90% Load	95% Load	100% Load
Average Response Time (ms):	2.27	3.81	5.35	6.89	8.43	9.97
All ASUs	1.99	3.72	5.47	6.08	6.49	7.22
ASU-1	2.73	4.78	6.69	7.31	7.72	8.40
ASU-2	2.04	5.03	7.70	8.61	9.13	9.94
ASU-3	0.42	0.88	1.90	2.36	2.74	3.54
Reads	4.51	8.26	11.38	12.34	12.89	13.72
Writes	0.36	0.76	1.62	2.00	2.32	3.00

See demo of random vs sequential io !!

© Colin Leesruch-Roberts, Kitem Consulting Limited March 2008

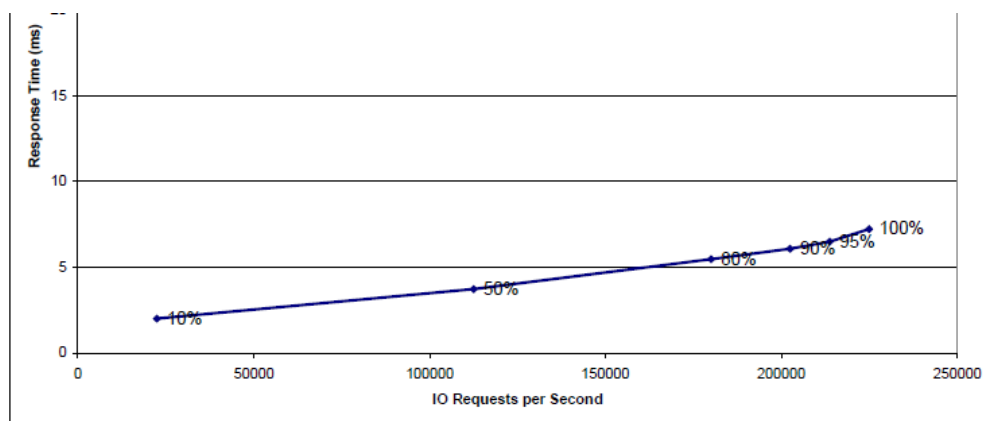
Live demo of random vs sequential io !!

© Colin Leesruch-Roberts, Kitem Consulting Limited March 2008

What is IOPS and does it have relevance?

To be Frank, instead of Grumpy, this is a measure that is, in my opinion, largely pointless. I explained how to calculate theoretical IO figures for a disk earlier, so when an array is put together a figure for IOPS is often produced. For example a vendor may say that an array can support 120,000 iops - but what does this figure actually mean?

Well actually very little; we already know that a disk can support sequential and random io, the actual figures being dependent upon the rotational speed of the disk. What we don't know is what type of mixture of io SQL Server will perform, there is no perfmon counter to tell us if it is a sequential or random io. We can look at average io size, but this is always an average and an average can blur the reality.



Response Time - Throughput Data

I/O Request Throughput	10% Load	50% Load	80% Load	90% Load	95% Load	100% Load
Average Response Time (ms):	2.27	3.81	5.35	6.89	8.43	9.97
All ASUs	1.99	3.72	5.47	6.08	6.49	7.22
ASU-1	2.73	4.78	6.69	7.31	7.72	8.40
ASU-2	2.04	5.03	7.70	8.61	9.13	9.94
ASU-3	0.42	0.88	1.90	2.36	2.74	3.54
Reads	4.51	8.26	11.38	12.34	12.89	13.72
Writes	0.36	0.76	1.62	2.00	2.32	3.00

- (disclaimer) I use this extract from a storage performance council full disclosure spec - 1 test result as an example - I am in no way seeking to express any views concerning the vendor concerned. This information is in the public domain so I see no legal reason why I might not use it for illustration.
- The results show the performance of iops vs latency.
- Points to note are that the latency doubles from 10% to 50% load, overall latency is reported as the average which disguises the considerably higher read latency.
- The configuration of the storage was all 15k 147 GB disks with mirrored volumes, no raid 5 or 6 the usually favoured choice of raid for all vendors. So does this actually help you - probably not!

As an aside here you should be very wary of setting long frequencies for gathering perfmon data.

Typically in a storage information gathering exercise you will log perfmon counters for analysis of trends - the longer the sample interval the more the data values will be distorted, or smoothed.

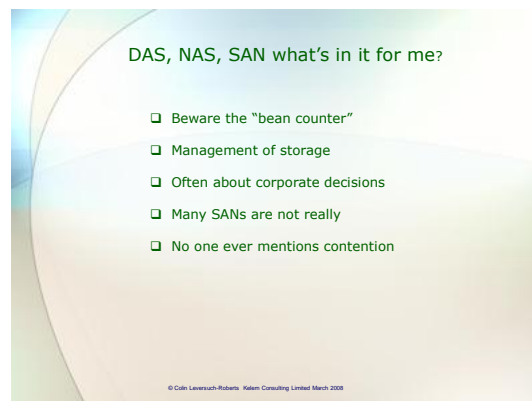
As an example I recount an interesting experience with some tests run against two vendor's SANs.

In my SQL testing SAN A outperformed SAN B by some considerable margin. Vendor B disputed my findings and produced a set of Iometer tests which convincingly showed that SAN B did in fact outperform SAN A.

Now of interest here was that the Iometer tests were, allegedly, set by DBAs for a large SQL Server implementation, the fact that the type of applications involved were different and the fact that my tests did not relate to either application were deemed immaterial.

The outcome of course was that I proved SAN B consistently performed worse than SAN A for a series of pieces of application code, backups, restores, index rebuilds and so on.

By all means make use of performance tools, but make sure your real decisions are based around actual SQL Server tests because that's what your users will be running, not Iometer, SQLIO or SQLStress.



DAS, NAS, SAN - what's in it for me ?

Essentially the choices here are defined by the scale and management of storage. What typically tends to happen is that once a company has bought into SAN technology everything gets put onto it - it doesn't actually matter usually if this is a good or bad thing from the DBA view.

- Beware the "bean counter" - I need to try to be diplomatic here, or just grumpy, many choices and decisions about storage are made at levels within companies between sales people and managers who are all trying to prove they can save money and deploy the latest must have. SAN storage tends to always be quoted at raw or raid 5/6 capacity using 1 TB or 1.5 TB sata disks.
- The upshot is that your perceived use of storage makes little reference to the reality, for example the humble transaction log; T logs should always be on raid 1, so you might have a 5GB log sat on a 147GB disk pair, or 300GB raw storage.
- The storage guys will of course add all your log files together and figure out that your total requirement is say 500GB , well your logs may be on 30 pairs of disks, they'll just allocate you 500 GB of a big raid 5 or 6 array, 60% of one 1TB disk.
- Let's say your SAN storage has 256 disk slots, with big raid 6 arrays and 1 TB disks you'll get about 220 TB of raw storage, let's say this comes in at £220,000 , £1k per TB.

- You need dedicated spindles, we'll stay at our 30 pairs, this removes 60 spindles from the mix, our remaining storage will be about 180 TB, we're now up to £1.22k per TB and so it goes on. If you want 10k spindle speed then 300 GB / 450 GB is the maximum capacity - oops we've over halved our capacity, storage is now £2.71k per TB. (450GB disks) £4.06k per TB for 300GB disks.

When I started with SQL Server 6 I could only use directly attached storage, I had 4 storage arrays attached to my server running a 3GB database.

Sometime around SQL 7.0 fibre channel arrived and I was told, incorrectly, that I had to switch to fibre channel to achieve clustered servers. I deployed an olap / datawarehouse solution in SQL 2000 to an early SAN, dedicated drives though.

NAS and SAN represent shared storage, in general terms NAS is a pretty poor solution for SQL Server, early use of NAS placed this on the general ethernet network - performance was/is miserable because you only have 1GB bandwidth, even with a dedicated network, which is what you should use, the bandwidth was 1GB, a third the performance of SCSI.

Things have moved on and with 10GB Ethernet iscsi can be a viable solution, I don't intend to specify vendors but there is at least one very good iscsi solution. A point to note here is that a 10GB Ethernet card may require 1GB of ram on the server, with dual cards that means 2 GB ram just for the HBAs.

Ethernet can be bad for storage because the order of packets can get mixed up and ethernet doesn't mind too much about a dropped packet.

Traditionally enterprise storage uses Fibre Channel because this was the first really reliable networked storage, fibre channel can also work over very long distances.

Networked storage - very important that this distinction is understood, a SAN isn't just a big storage box, it represents pools of storage in different locations which is managed and allocated as a whole and NAS would be part of the SAN which may consist of many types of storage.

Often DAS is still a faster solution for SQL Server, there is in fact no real limit to the scalability of DAS as you may use SCSI, SAS and FC for directly attached storage. Your choice of server will limit your ultimate scalability as you need increasing numbers of HBA's, however you can still multipath DAS.

The hidden issue with a SAN is contention, this is one factor which is often overlooked, in fact I might say it is ignored - this is a whole subject area in its own right and I'll return to this a little later. I've blogged briefly about contention (best viewed on the grumpyolddbaweb site)

<http://www.grumpyolddbaweb.co.uk/infrastructure/TrackingSANContention.htm>

One aspect often overlooked is the speed of light, it takes about 1 ms for light (fibre channel) to cover 124 miles, if your sites are 5,000 miles apart you have a 50 ms latency (one way) 100 ms round trip.

In scaling up or scaling out we'll discuss if a SAN is actually a SAN.

Raid – there can only be one

- There is only one raid level to use in a database which is anything other than read only and that's raid 1 (raid 10) (mirrored volumes)

```

graph TD
    VD[Virtual Disk] --- RS[Raid 10 Stripes]
    RS --- R1M1[Raid 1 Mirror]
    RS --- R1M2[Raid 1 Mirror]
    RS --- R1M3[Raid 1 Mirror]
    RS --- R1M4[Raid 1 Mirror]
    R1M1 --- R1M1a[Raid 1 Mirror]
    R1M1 --- R1M1b[Raid 1 Mirror]
    R1M2 --- R1M2a[Raid 1 Mirror]
    R1M2 --- R1M2b[Raid 1 Mirror]
    R1M3 --- R1M3a[Raid 1 Mirror]
    R1M3 --- R1M3b[Raid 1 Mirror]
    R1M4 --- R1M4a[Raid 1 Mirror]
    R1M4 --- R1M4b[Raid 1 Mirror]
  
```

- I don't intend to discuss raid in any detail as it's a well covered subject.

© Colin Levenson-Roberts Kalam Consulting Limited March 2008 12

RAID - " There can only be One "

There is only one raid level to use in a database which is anything other than read only and that's raid 1 (raid 10) , mirrored volumes.

There's a lot of hype about raid levels, I've seen " fast raid 5 " , in advertising - fast ? Raid 6, often called something vague so you don't know it's Raid 6, which allows a slightly higher level of redundancy when building large arrays - even worse than Raid 5! The problem arises with the parity writes when saving data, raid 5 requires 4 io operations to complete 1 write io, so that's a 300% degradation on writes, Raid 6 has two parity disks btw.

Some vendors may claim that caching will give Raid 5 or Raid 6 the performance of Raid 10, just think hard about this claim, if the caching can make writes 4 times quicker then it must surely improve the performance of Raid 10 by the same amount.

I don't intend to discuss raid in any detail as it's a well covered subject.

Now to put all this about disks, raid, bus speed and interface into context let's consider these explanations:

Should you go to a football match (I never have, but I attend rock concerts in stadiums from time to time) consider the approach route to the stadium as your BUS – the wider the approach the more people you can get down it – same with the BUS speed. Now consider your disk arrays as the turnstiles – the faster they turn the quicker people get into the stadium, the more turnstiles the more people you can get through. To get more people through quicker you get more turnstiles and/or faster turnstiles. Pretty simple.

Now Raid 5 – if you have raid 5 each person entering the stadium (writes) has to enter through 4 different turnstiles before they get in – poor explanation but I think you get the idea.

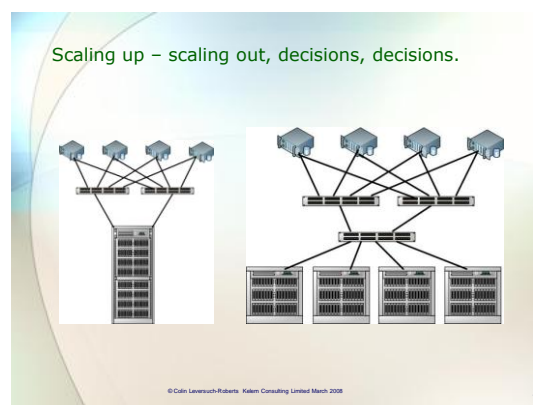
Raid 6 – would require 5 or 6, sorry not too sure how many extra io are involved with Raid 6

Raid 6 is used to gain capacity within an array, make use of maximum disks – typically used in arrays of over 10 disks, Raid 6 usually will not have a hot spare as there is a second parity stripe.

Raid 5 is best limited to 10 disks or less.

A quick word on formatting and partition alignment. Pre Windows 2008 you need to align disk partitions, the performance degradation can be around 10 - 15%.

Block size - to date I have not been able to measure any performance differences when using different block sizes to format an array.



Scaling up, scaling out, decisions, decisions.

A Storage Area Network should be agile, it should allow scaling in all and every direction.

That said how do we actually implement such an architecture?

I mentioned earlier about a SAN maybe not really being a SAN, well it's all in the definition I guess.

Most storage sold as a SAN is really a monolithic block of storage connected via a network, this type of storage is probably the most popular deployment, it usually has a limited capacity, if you want to have more storage you'll be encouraged to buy a new bigger unit. Connectivity is usually limited, e.g. it may only have 1GB ports, you have to buy the next model up to get 4GB and the next next model to get 8GB or more ports etc. etc.

Typically you may have to pay an additional license fee based upon capacity, back to bean counters and raid 5.

This storage scales up to the limit of the unit, so is not truly scalable.

What you really want is networked storage, this is more agile. This is scale out and is technically infinite in capacity

The diagrams are for illustration only, they are not network designs!

Bandwidth is also a key element in scaling, I assume that everyone here is familiar with teaming or multi path ?

Subject to the various bus speeds you can multiply up HBAs to increase bandwidth or data throughput (and to provide redundancy) The more paths you use the more switches you need and the greater the complexity, however this is an element of scale out.

Is the future SSD?

- ❑ Expensive - £1k for 64GB
- ❑ Questionable deployment practicalities
- ❑ Good for random io
- ❑ Could reduce number of disks

© Colin Llewellyn-Roberts, Nalim Consulting Limited March 2008

14

Is the future SSD?

Well it's a very interesting question.

Enterprise SSD are expensive - typically £1k for a 64GB SLC disk.

The real gain with SSD is the highly reduced seek time or latency, typically tenths of milliseconds.

Largely continuous throughput can still not match current 15k disks.

Where could you use SSD with SQL Server ?

Possible places are the first file for Transaction Log

SQL Server uses transaction log files sequentially, unlike data files. Allocating the first transaction log file as a SSD may well give significant gains.

Tempdb may be a contender depending on how big your tempdb is.

For a small database then SSD are viable proposition if you are unable to ease any more performance, you could use the lower latency to reduce the number of physical disks you need.

There is a case study showing successfully running a SQL Server 2005 database on SSD only. The database is/was only 80GB and I cannot give a link to this as I get this under subscription (it's a storage site) and I'm totally unsure on the legalities of doing so, sorry. You can find it under www.searchstorage.co.uk

It has been suggested that the optimiser in SQL Server actually has parameters built in which relate to the speed of disk access and that by using SSD you could get strange side effects - I have no verification of this and I doubt seriously that if I asked I'd get an answer from Microsoft anyway.

Technically you should be able to bypass SSD with small databases by using x64, a DL580 series server can hold 256Gb of ram so in the case of our 80GB database the entire thing should be in cache anyway!


For databases of a TB or more the use of SSD would be tricky, maybe selective filegroups on SSD - but that could prove a management nightmare.

I have installed a (cheap) SSD in my 5 year old laptop - it's much quicker - sql queries run 6 to 10 times quicker, I have another set of SSD in an array and have carried out some tests - for sequential io they were slower than a comparable sata raid, for random io they were considerably faster.

Operational aspects

- ❑ Fragmentation
- ❑ Virtualisation
- ❑ Sharing
- ❑ Latency
- ❑ Raid rebuilding
- ❑ Thin provisioning

Shared lun contention demo



© Colin Laversuch-Roberts, Kalem Consulting Limited March 2008

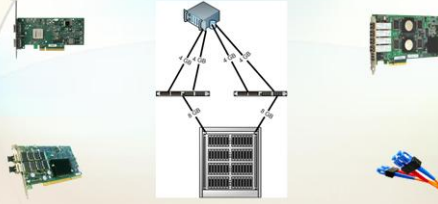
15

Shared lun contention demo



© Colin Laversuch-Roberts, Kalem Consulting Limited March 2008

16



Multi path 4 x redundant HBAs

© Colin Laversuch-Roberts, Kalem Consulting Limited March 2008

17

Operational aspects, fragmentation, virtualisation, Sharing, latency and rebuilds.

A brief few comments on the operational aspects of storage.

Fragmentation - said to be irrelevant when using a SAN. Difficult to test, not strictly true depending upon how you define fragmentation. Certainly has impact on conventional DAS or internal arrays.

Virtualisation - many edged sword - there are virtualised SANs, these disguise the actual physical storage and present virtual luns.

This brings us to shared resource.

Shared luns or non dedicated physical disks is the biggest cause of contention within the actual storage layer - now anyone who lives in london and use public transport will have seen contention in action, maybe you haven't made the connection but you should!

I use the Jubilee line currently, most of the stations have the lines parallel with common access to both platforms from a central stairway/escalator bank.

When a single train arrives we have usually pretty quick egress using two escalators, however if two trains arrive at the same time we have contention as two streams of people are attempting to use the two escalators, so progress is often a series of stop and starts - welcome to shared luns. At peak times when people are streaming down the escalators and both trains arrive at the platform your progress may come to a total stop, there will be collisions, stops, starts - this is read and writes on shared luns - using my live demo again !!

(I would have used video from the underground but taking photos or video will get you arrested these days!)

Now this is where big disks are bad - there may be say 6 platters inside a disk - but the heads all move together if it's pulling one piece of information it will be be in the wrong place to make a write, if we divide the disk into luns we will have big problems.

Virtualising the storage usually pools the defined luns into a "lump" from which you allocate your luns - what happens underneath goodness knows.

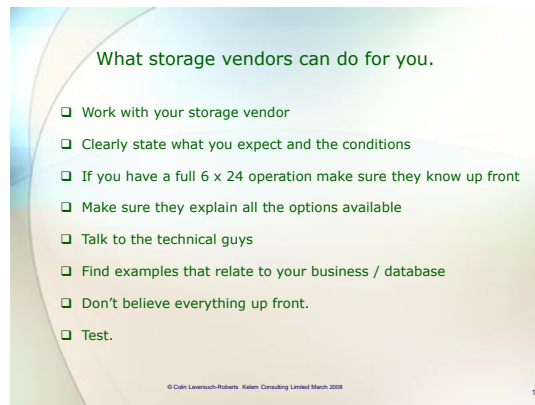
Now I've talked about latency, if there's a lack of monitoring this is probably one of the few perfmon counters which will at least give you a handle on performance problems - this is another subject unto it's self - watch for latency or increased latency without increased load. Note that latency may occur due to the fabric, not just the storage - this is one of the issues of the monolithic manner of many SAN implementations; 60 servers using 4GB HBAs but only 4 4GB ports on the storage, or maybe 4 x 4GB, 4 x 2GB. This is known as fan in and fan out and relates to how bandwidth is throttled or expanded.

This is why you may need scale out rather than scale up for your storage.

Lastly raid rebuilds, should a disk fail the raid has to rebuild when the new disk is added, Raid 5 and Raid 6 are particularly bad for this and the degradation of service during the rebuild may be sufficient to take your database down, the larger the disks the longer this will take; there was an excellent article, I think in Storage Magazine, which pointed out that when using a large array of 1.5TB disks the rebuild time may be several days; there are calculations available for this.

Raid 10 doesn't really suffer this way at all. Talking of which the risk % for array failure for a raid 5/6 increases as you add disks, for raid 10 it remains constant.

Thin provisioning largely doesn't with ntfs !



What Storage Vendors can do for you

It would be wrong to try to pick out individual vendors and features as these often vary from one product to another.

You're most likely to find these features on offer

- Multi Pathing
- Snapshot Backups
- Volume Copy

All vendors storage can do stuff you cannot do with conventional DAS, it's usually a matter of price.

Now Snapshots are a well publicised feature, however most of these require a total quiesce of the database, this may not suit your application and if your application uses more than one database you may not get a restore in a consistent state.

As with all such offerings it is very very important that you actually test that it works for your environment .. I'm not going to regale you with stories of best laid plans that work well in tests but not in reality; but I will mention one famous incident:

A Hospital installed a gas powered generator to provide emergency cover should mains power fail, all tests worked seamlessly with no problems until there was a real power failure. Sadly the generator required mains power to start, this had always been present during testing!



References:

1. [Serial Attached SCSI - wikipedia](#)
2. [SNIA Europe](#)
3. [Device Bandwidth - wikipedia](#)

[Grumpy Old DBA Web site](#)

[Grumpy Old DBA Blog](#)