Thursday, September 3. 2009

Some perspective to this DIY storage server mentioned at Storagemojo

I've received yesterday some mails/tweets with hints to a "Thumper for poor" DIY chassis. Those mails asked me for an opinion towards this piece of hardware and if it's a competition to our X4500/X4540. Those questions arised after Robin Harris wrote his article "Cloud storage for $100 a terabyte", which referred to the company Backblaze, which constructed a storage server on its own and described it on their blog in the article "Petabytes on a budget: How to build cheap cloud storage". Sorry, that this article took so long and there may be a higher rate of typos, as my sinusitis came back with a vengeance ... right in the second week of my vacation. But now this rather long article is ready

At first: No, it isn't a system comparable to an X4540 ... even without the considerations of DIY versus Tier-1 vendor. I have a rather long opinion about it, but let's say one thing at first: I see several problems, but i think it fits their need, so it's an optimal design for them and they designed it to be the optimum for them. I assume, many problems are addressed in the application logic. The nice thing at custom-build is the fact, that you can build a system exactly for your needs. And the Backblaze system is a system reduced to the minimum.

This device is that cheap because it cuts several corners. That's okay for them. But for general purpose this creates problems. I want to share my concerns just to show you, that you can't compare this to a X4540 device.

And even more important: I have to deny the conclusions of the Backblaze people. This isn't a good design, even when you just need cheap storage, when you don't own a middleware that does a lot of stuff that ZFS would do in the filesystem for example in the hardware. On the other side it supports my arguments in regard of the waning importance of RAID controllers. The more intelligent your application is, the less intelligent your storage needs to be.

So ... what are my objections to this DIY device:

The DIY Thumper has no power-distribution grid. So when one PSU fails, all devices connected at this power supply will fail. In the case PSU2 fails, the system board is away, thus the machine fails. Game over ... until power comes back.
Connected to the last problem: Given the disk layout, the power-distribution isn't correct. They use it with RAID6, but RAID6 just protects you against 2 failures. I don't see a sensible layout in three RAID6 groups, that would allow the system to loose 25 disks at once. A more reasonable RAID Level would be RAID10, but there you have 5 disks without a partner in the other PSU failure domain.
I don't know if i consider a foam sleeve around the disks and some nylon screws as enough vibration dampening, especially when your hard disks. I'm looking forward to the next article they announced which was announced to cover this topic. It will be even more interesting to hear more about it in the future because of the performance and the longevity of the disks in such an environment. Just an example for the real world: Once we found out that disks near to a fan were a tad slower than the ones far away from the fan. This led to changes to the vibration handling in that system.
This baby cries for ZFS. So much capacity, no battery backup RAID controller, only 10^14 disks. But i see the reason, why this choice wasn't feasible for them: Since a few weeks ago, the OpenSolaris SATA framework hadn't support for port multiplier. This was introduced with the putback of PSARC/2009/394 to OpenSolaris. But now it's integrated. And given, that this baby just speaks HTTPS to the outside and the software relies on Tomcat, it should be a piece of cake to move to Opensolaris and ZFS now.
This design isn't really performance oriented. As they use Port multiplier to couple their disks to cheap SATA PCIe/PCI controller, one 3 GBit/s interface has to feed 5 disks. One ST31500341AS delivers round about 120 MByte/s (saw several benchmarks suggesting such a value). Five of them deliver 600 MByte/s, a little bit less than 6 GBit/s. So each SATA channel is oversubscribed by a factor of two.
Even more important, three of the connections to the port extenders are coupled to a standard PCI-Port. One PCI-Conventional 3.0 port (didn't find an information what the board provides, thus i assumed the fastest, source is the german wikipedia page about PCI) is capable to deliver round-about 4 Gigabit/second (to be exact 4,226 GBit/s). Thus you connected 18 GBit/s worth of hard disks at 4 GBit/s worth of connectivity.
I have similar objections for the PCIe connection for SATA-cards. Those ports are PCIe at 1x. One PCIe 1x port has a theoretical throughput of 250 MByte/s. So such a port would be fully loaded by just two hard disks. But this baby connects ten disks to a single lane of PCIe.
Of course those hard disks doesn't run at max speed all the time, i assume the load pattern will be very random in the special use case of Backblaze. But this leads to a high mechanical load to the disks and to some additional objections. Based on the manual of the hard disk, i see two problems here:

The ST31500341AS is a desktop disk. They do not even one of this nearline disks like we use in the X4500/X4540. When you look in the disk manual, all reliability calculations were done on the basis of 2400 hours of operation per year. But a year has 8760 hours. When you don't believe me this 2400 hours, just look at page 24 of the manual.
The reliability considerations of Seagate assume a desktop usage pattern, not a server usage pattern.
Seagate writes in their manual itself: "The AFR and MTBF will be degraded if used in an enterprise application". But given the long credits list at their end, i assume they've read the manual and considered this in their choice of hard disks.
There is another important point about the reliability of the disks: The AFR and the MTBF for the 7200.11 is valid for a surrounding temperature of 25 degrees celcius. Running it above this temperature reduces the MTBF and increases the AFR. Other harddisks build with enterprise usage in mind use another normal temperature vastly higher as the foundation of this calculations.
But due to the usage of RAID-6 those disks will see a high throughput in any case. RAID6 relies on a READ/MODIFY/WRITE cycle due to the nature of RAID6. So you read/write vastly more than just writing the modified data to disc. This may even interfere with the sparse throughput of the system. We've introduced RAIDZ, RAIDZ2 and RAIDZ3 to circumvent this kind of problems
No battery backup for the caches, but RAID6 ... well ... "Warning ... write holes ahead"
This system uses a Desktop Board, the DG43NB, thus system resources are a little bit sparse on this board. Just 1 processor and just 4 GB of RAM. I find the later one a little bit problematic. For general purpose a lot of more memory would be feasible. There are good reasons to have 32 GB or 64 GB in a X4540. Without a large amount of cache, you aren't able to shave off a little bit of the IOPS load to get back to a moderate load, thus the choice of a desktop disks gets even more problematic here.

I think, Robin Harris is correct with his comment, that this system is a DC-3. It flies, it can transport goods and passengers from A to B in a reasonable, but not fast speed but don't forget your parachutes It's the same with this storage, this hw needs the parachute in form of the software in front of the device.

But, and this is one of the key take aways for you ... even when other systems are more expensive, they are not overpriced. At first don't compare the mentioned list prices with the street prices for components. Second: Of course you can save an dollar at one or the other place, but: The seagate hard disk costs you 100 Euro at a big german computer online-shop, the HUA721010KLA330 (aka Hitachi Ultrastar A7K1000 1TB) costs you roundabout 200 Euro after a search at Google. Just using other (in my opinion correct for general purpose) disks, would double the price despite offering less storage. And even this price isn't indicative, as most often there are special agreements between drive manufacturers and system manufactures because of quality standards, quality management and conditions.

The technical differences of the UltraStar: 1 errors in 10^15, qualified for 24/7 operations by the manufacturer, qualified for a enterprise work pattern (and even here only a lighter one) and 1.2 Million Hours MTBF normalized on 40 degrees (AFAIK) instead of 0.7 million Hours at 25 degrees.

Quality costs. Period. The same for a desktop board in the DIY-"Thumper" instead of a custom build board for optimal performance (a SATA controller for each disk or using 8x lane PCIe for 8 disks instead of 1x lane PCIe for 10 disk e.g.). I'm pretty sure Sun could build an equally priced system, when you take the bare metal of the X4500 chassis and rip out all the specialities of the X4500/X4540 systems. But such a system with so many corners left wouldn't a be a system you expect from Sun. And yes, the X4540 has less capacity at the moment, but i think it's not far too fetched, that the X4540 gets 2TB drives as soon as they reached the same quality standards and qualification as the current drives givinh the X4540 a capacity of 96 TB.

To close this article: It's about making decision. Application and hardware has to be seen as one. When your application is capable to overcome the limitations and problems of such ultra-cheap storage (and the software of Backblaze seems to have this capabilities), such a DIY thing may be a good solution for you. If you have to run normal applications without this capablities, the general-purpose system looks as a much better road in my opinion.

Posted by Joerg Moellenkamp in English, Oracle, Solaris, Technology, The IT Business at 15:22

Interesting comments. I have a better idea now of why Sun charges as much as they do for disk drives for my x2270 ($600+/TB). Presumably they are selling me an enterprise-level drive. Still, I would like to be able to buy the empty disk carrier so I could put my own (desktop-quality) drive into it. Alas, there is no part number for that in the catalog.

Thanks for posting.

-cwl
Anonymous on Sep 3 2009, 19:25

There are other niceties: Did you ever tried to get a firmware update for a desktop harddisk (besides of big bugs like the one at Seagate recently) ? Or the rather long qualification process ....
Anonymous on Sep 3 2009, 22:15

"If this expensive equipment fails 1% less often over the course of a year that's 90 hours less downtime." - Me (I know it's not exact, but it gets the point across)

Precisely why I educate my clients in the importance of spending the right amount of money on equipment. Just becasue Cisco is too expensive, it doesn't mean we should use netgear.
Anonymous on Sep 3 2009, 22:44

Stop charging 3x the amount for the drives and we'll consider Sun. We tried to get the J4200 and it wasn't possible to buy the array with the trays and no drives.
We bought WD RE4 drives and another manufacturers drive array because we couldn't justify paying such a ridiculous premium on drives.

We bought the server from sun with minimal disks and OEM drive carriers.
Anonymous on Sep 3 2009, 23:25

We buy WD enterprise drives and have been for a few years. It's best to buy OEM drive carriers and enterprise drives on your own.
Anonymous on Sep 3 2009, 23:27

They have emulated a tape backup solution with their box. It's cheap, it does mostly writes, and a read once in a while.

Why would you want to use anything but a simple and cost effective hardware solution in this case? Get a desktop m/b, a few gigs of ram, and a bunch of home-type disks. With no fancy stuff needed, only cheep, raw, storage. A disk fails? Replace it. A power supply goes... You'll be back online with that box when the part is replaced. I don't think their service is geared towards the mission critical situations, just effective, cost-wise, backup solutions.
Anonymous on Sep 3 2009, 23:43

Of course, this is the reason why i wrote that this is an optimal hw for their needs. This is the advantage when you can build your own hardware based purely on your own needs and put many of the tasks from the hardware into the software. Then such an reduced-to-the-max approach is a sensible one. Otherwise oher concepts seem to be better ways to go.
Anonymous on Sep 3 2009, 23:49

It would be really nice if Sun offered the bare empty X4500 chassis and let people build their own with their own choice of disks and motherboard, if people don't need all the super speed and reliability of the X4500 but still want the density.
Anonymous on Sep 4 2009, 02:28

Fantastic! Great post highlighting -some- (I'm sure there are loads more) of the issues. Some people around here that seem to think the backblaze is the greatest thing. I'm busily trying to kill the "hype".
Anonymous on Sep 4 2009, 03:10

And then when one of the disks fail, we get a support call logged and after spending the time and resources diagnosing the fault Sun Support has to say "I'm sorry, your third party disk has failed, we can't help you". Real good for customer satisfaction.

Yes, I know you wouldn't make that support call, but many would.

Alan.
Anonymous on Sep 4 2009, 03:49

I think it depends on useage

for home - i rolled my own server with 3 zfs 1tb mirror pairs +netatalk and iscsi

But for enterprise we use thumpers and EMC sans

If you look at the graph of the petabyte costs per vendor .. sun are about right .

The bottom line for our clients is EMC is too expensive for most of them .. so we use thumpers - happly

EMC just costs too much - far to much
Anonymous on Sep 4 2009, 09:19

88 Comments on Hacker News till now, not bad

http://news.ycombinator.com/item?id=803136
Anonymous on Sep 4 2009, 11:36

Well of course you get more with Sun. But is it worth 10 times the price? I think no. In the end we compare price and reliability. The backblaze solution is working 99.9(99?)%. Is the remainding few points really worth paying 10 times more?
Anonymous on Sep 4 2009, 11:52

Well, this hw doesn't have an availability not even near of 99,9% (don't talk about 99.999%) . Five nines translates into 26,3 seconds outage per month. I don't think you can even swap a disk in this time. Calculate into the equation that there is no redundant power supply (there are two power supplies in it, but they are not redundant). All in all this system should have a really mediocre availability.

But there is a reason why this works for them: Their application logic keeps the data available. And perhaps they can reach 99.9% for the complete solution.

But this is not the general case. Most often we talk about applications that need a better availability, because they assume an datastorage with a reasonable availability.

As i wrote before ... it's about your application: The more intelligence you have in your application, the less less intelligence you need in your storage to keep data available.

By the way: The price calculations are somewhat skewed. To keep a decent availability they have to keep the data threefold from my opinion. Thus you get 333 TB from your 1 PB. Now use a system that is more available by the single unit. Now it may be sufficient to keep the data on just two systems to get the same availability. So you get 500 TB out of your 1 PB. Put into equation, that they should use other drives (doubling the costs) and that you have to factor in the price of the development, testing et. al. then the price isn't that different.

I don't included this thought into the article, because you can argue about the interpretations too much and i just know the stuff from the blog about Backblaze.

At the end: It's a perfect hw for their solution, but don't take for granted that this a solution sufficient even for similar cases. And for general-purpose with standard application: Forget it ... without ZFS i wouldn't even think about using this 10^14 disks ...
Anonymous on Sep 4 2009, 12:21

It would be nice, if SUN would build a system nearer to the backblaze specs for backup purposes.
It will not happen, since accountants will not understand the difference between enterprise 24/7 online storage and a backup device, which may have a service downtime of a few days. If the software supports power switching the disks, even the desktop drives will be ok.

There are a lot of applications, which can be done on cheap storage, but you have to explain a buyer, that a barn is not a fully automated high rack storage area.

You can store things in both and the barn wins in price.
Anonymous on Sep 4 2009, 12:56

Sorry, but desktop drives wouldn't be okay. Given that you use 1 PT not alone. You have to work with rather short idle time outs. You should keep into consideration, that an desktop drives is designed for 10.000 start/stops per year.

This leads to a start-stop every 52 minutes. Further you should keep in mind, that you end with waking up many disks due to striping and such nasty challenges like partial stripe writes. And then consider that you don't have one user, you have several. Perhaps thousands.

Thus to get the disk sleeping for at least 6360 hours (8760 - the specified 2400 power-on hours for the 7200.11) in a year you have to set rather short sleep time outs, but that will bite large chunks of your budget of 10.000 start/stops.

The problem: Those disk will still work reasonable well when used out of specification ... at the beginning ... but the problem will haunt you, when your disks get a little bit older. Of course you could preventively substitute your disks, but that affirms the old saying "Buying cheap, is buying two times"

Better to buy a disk, that is capable to run 24/7. right from start.
Anonymous on Sep 4 2009, 13:47

Given the rise of the SOHO NAS, HDs manufacturers are slowly introducing intermediate-level HDs, like the WD Caviar RE3 built for 24/7 with a MTBF of 1.2 million hours ... at a decent price.

What's missing now is an intermadiate-level storage solution. Not everyone needs 99.999%. Lots of IT departements are ready to accept to have a little higher failure rate for a 5x price decrease.

Sun can ignore this market, or embrace it while there's still room for a big player.
Anonymous on Sep 4 2009, 14:31

But what if the power supply goes during a write (highly likely)? Don't the absence of cache and battery on the controllers matched with different controller I/O speeds make data corruption a real possibility? If you lose a power supply, you may be down much longer than you think. I would assume they are backing up or replicating with their application solution like is mentioned, but, even then, bringing back that much data can take a while.

Seems like this is right on the money. A good solution for their needs and they should also be applauded for sharing, but for others there are some failure and performance points that need to be considered.
Anonymous on Sep 4 2009, 14:40

The argument with start/stop cycles is ok, but only, if the system is used as an enterprise server. For backups, there are a low number of accessors at a time. And a private home page server has a lot of pages with nearly no access. Small companies will have access patterns, that allow the drives to sleep the whole night.

If the system is not used for storing seldomly used files, using enterprise hard drives is recommend of course.

I think, there is a market for this systems as archive storage. How much storage will you need, to justify the expenses for two tape drives and a robot ? I think this will be in the 100 TB range, which is a lot for backups and archive of a 100 to 500 people shop.
Anonymous on Sep 4 2009, 18:01

10.000 start/stop-cycles?

Aren't desktop drives usually the ones with more start/stop cycles that a typical server disk?

And only 2400 power-on-hours - that would be great to have for energy consumption considerations, but I really don't see a 24/7 8760 day/Year application as that big of a problem for desktop disks.

The fact that it is not specified is no reason that it doesn't actually work and perform well.

I really don't see the desktop disks failing that often that you'd have to continuously swap-in new ones.
Anonymous on Sep 4 2009, 22:50

Sorry, but please look at page 23 (2.11 2.11.1 Annualized Failure Rate (AFR) and Mean Time Between Failures (MTBF) of the Seagate 2007.11 product line at http://www.seagate.com/staticfiles/support/disc/manuals/desktop/Barracuda%207200.11/100507013e.pdf.

This document specifies "2400 Power-on hours". I assume that has something to do with the bearings and wear on electronic components when powered.

Furthermore this document specifies that "10.000 Start/Stop cycles". Obviously you are right, that desktop drives are speced for more start and stops. When you look in the specification for the Cheetah 15k. (http://www.seagate.com/staticfiles/support/disc/manuals/enterprise/cheetah/15K.6/FC/100465943a.pdf) you will see that all this reliability calculations are done on the basis of whooping 250 start stops per year, but different to the desktop it's speced for 8,760 power-on hours per year, so you don't have to shut it down just to keep it in the power-on hours envelope.

When a device has a annualized failure rate of 0,37% at 2400 hours, then it rises to 1,25% when used out of spec in regard of the power-on-hours. Mix a non-desktop load to it, perhaps a few degrees away from the 25 degrees specified ambient temperature you have all the factors to shorten the life expectancy of your drives. Now take into consideration, that those people doesn't have just one drive they seem to have hundreds of them ....

The problem: Even a desktop drive can be used in enterprise load ... for a while ... like driving a motor constantly in the red area ... it will work for a while ... but you significantly shorten the life of the device. And there we get to another problems: When discs are buyed at the same time, used at a similar load, you could expect that the disks are dying at a similar time. Now you've additionally accelerated the wear on the disks by out-of-spec usage ...

And by the way,there are lot's of anecdotes of people using SATA disks in their ghetto RAID where they used FC drives before by the choice of the financial controller, put the same database load on it, and saw the drives dying like flies after a while.

I can just assume that Backblaze took this into their consideration and plan to substitute disks early to participate on the increased data density of disks so they don't have to buy new racks and systems, just new disks.
Anonymous on Sep 5 2009, 08:00

c0t0d0s0 is perfectly right:
Backblaze is selling BACKUP services.

BACKUP is the last line of defense. In case your data is lost you will want to recover from your BACKUP.

But this backblaze stuff is no BACKUP solution. BACKUP is done on TAPE not on DISKS. This is why the industry inventend tapechangers for backup purposes. Because a tape is much more reliable than a drive. Anyone heared about silent bit corruption ? No ? RTFM.

I would never backup my data in a datacenter like the one backblaze runs. Never ! Oh theres one exception: in case i never need my data back i would trust backblaze. But on the other hand - in this case it would be the cheapest solution to backup all data to /dev/null. Cheap and fast.
Anonymous on Sep 11 2009, 13:06

I think Micheal is off the mark.

1. Tape media fails. It does not offer fast random access, like its magnetic storage HDD counterparts. We stored data in the 1980s on three different tapes and put them in different locations. When we needed to recover an old program, all three failed.

2. I am not sure what BB does exactly, but I would imagine if there were conservative, is to have a copy offsite too. And weekly, run a hash algorithm on each file stored and isolate any degradation. I believe the SW mitigates any issues here.

3. Bit errors will occur. They can happen in RAM before the write and in numerous other ways. Again, application logic using big hashes, can mitigate much of this and provide the highest reliability.
Anonymous on Sep 24 2009, 07:54

Hey, you have a great blog here! I'm definitely going to bookmark you! Thank you for your info.And this is Home Improvement site/blog. It pretty much covers Home Improvement related stuff.
Anonymous on Sep 27 2009, 13:22

Jones,

1st: Yes Tapes can fail. But it is more unlikely than loss of data on Hard Drives. Did you ever hear of "silent data corruption" ? No ?

Then read this please:

2nd: i think they don't have an offsite copy.

3rd: Yes bit errors will occur. Using application logic to avoid bit corruption is only one part of the problem. But using CRAP (like backblaze does) makes all worse.
Anonymous on Oct 13 2009, 10:15

Link to silent data corruption:

http://raidinc.com/pdf/Silent%20Data%20Corruption%20Whitepaper.pdf
Anonymous on Oct 13 2009, 10:27

I read about the backblaze, and found this blog 20 minutes later.
96 TB is definitely too much for me, but I'm thinking about building 2 storages 1 for backup, and the other one for my esx farm. Using NFS.

I came to the conclusion that I could easily upgrade the Backblaze design after reading your post, by modifying 3 major elements:

SATA controller: 3aware 9650SE-12ML - 8-Channel
File System: ZFS with OpenSolaris 0906 - RAID-Z2
Hard drive: 8 or 12 Disk (still wondering between the Seagate Barracuda ES.2 ST31000340NS and the Western Digital Caviar Black WD1001FALS)

My major problem is to find the enclosure to install these 8 or 12 HD !

What do you think ?
Anonymous on Dec 10 2009, 07:53