Monday, August 3. 2009

The waning importance of storage array controllers

I found an interesting text this morning. In "SSDs, pNFS Will Test RAID Controller Design" Henry Newman speculates about the future of RAID controller and the difficulties their vendors will have in the light of the advent of SSD. My opinion is somehow divided: Yes, RAID controller vendors will have a difficult time in the future. But: No, this won't be the fault of SSD.

Interestingly the storage market behaves in waves (like many other markets). The pendulum moves between RAID-Controller and JBOD for years now, at the moment the pendulum moves to JBOD, it was on the RAID-Controller side fore many years, albeit the next few years may destroy the pendulum, as the shortfalls get more and more visible and other technologies are available.

The most imminent thread to dedicated RAID controller has it's foundation in the past: Hardware RAID was invented many years ago, when the main CPUs hadn't a large amount of power to compute the necessary calculations for RAID5 for example. So it was obvious to offload all these calculations into an external device, the RAID-Controller was born.

But this method isn't without shortfalls like the read-modify-write cycle when you modify a stripe or the point that a HW raid controller gives some protection, but just in case you have to recover. On the other hand that HW raid controller hides the redundancies for more intelligent mechanisms: For example when you use HW RAID you just see one copy with one checksum from the more intelligent mechanism. The system can detect the error, but has no redundancy to correct it. A more intelligent mechanism may use the redundancies of RAID to correct the data, even when it was corrupted on wire. A more intelligent mechanism than a HW raid controller may be aware of the placement of data on the disk and doesn't have to put millions of zeros or useless already deleted data in sync. One of those more intelligent mechanisms would be ZFS, but I'm sure that the future will bring us other, similar technologies as other operating environments have pretty much the same problems with their storage.

I really think, that all storage will look really similar to the S7000 in the future. When we talk about increasing requirements to the storage systems by new storage media we get to a point where some embedded CPU aren't enough and we end with systems that look really similar to an x86 server. Maybe they will be better hidden than at our S7000 series, but it will be similar. And when we already are at this point, many storage companies will come to the conclusion that it may be a good choice to trow away their operating system they've used in the past and use something already available, a Linux, a BSD or, well, an OpenSolaris.

But then we get to a more important point: The people using this component could get to the idea, that storage arrays like we now them today with their centralized storage controller could be just a bottleneck. And when all this storage-stuff is done be general-purpose OS on general-purpose hardware you could come to the conclusion, that your servers could do the job as well and get rid of some of the shortfalls I've described above. To explain that dedicated storage controllers have still some advantages will be a tough job for the vendors of such components.

For a long time data services like replication were one of the advantages, but many of them are already available with OpenSolaris for example: You can already do replication (synchronous with the help of AVS and asynchronous natively) with ZFS, you can do compression, you can thin-provision, you will be able to do encryption and deduplication, you can do all this file-level and block level sharing protocols. So even the resort of having more data services is just a short-lived, almost non-existent last resort.

pNFS is the next new technology, that may be problematic for the future of dedicated storage controllers. Distributed in nature, there doesn't seem to be a niche for this controllers. There are just large amount of servers with a modest amount of hard disks per server but again ... just as a JBOD.

So, i've just explained, why dedicated storage controllers may be lose their drive(s), but why isn't it the fault of SSD as I wrote in the beginning. The reason lies in the nature of SSD: You simply put SSD not behind RAID controllers. You keep the distance between the CPU and the SSD as short as possible. Period. Any given local area storage network introduces latency. The higher the number of I/O operations per second, the more harmful is any additional latency. There is only one exception: In case of a cluster, you have to put everything you need to fail over a consistent version of your data to the surviving node in the local area storage network to enable a fail over. And when you don't put the SDD behind the RAID controllers, they can't be the bottleneck.

ZFS offers a technology called Hybrid Storage Pool (I'm pretty sure that we will see similar technologies in Linux and Windows at a point in the future) and with technology you don't have the need to put SSD behind a controller. It's in front of the controller and reduces the load to the controller. Many people still think about SSD as a substitute they hard disk, but we have to thing different, we can think different with HSP.

So: Yes, RAID controller are really a technology that may be of waning importance, but the reasons are different.

Posted by Joerg Moellenkamp in English, Oracle, Solaris at 20:19

"The pendulum moves between RAID-Controller and JBOD for years now, at the moment the pendulum moves to JBOD, it was on the RAID-Controller side fore many years, "

From what I have seen the last ten years, everything and its grandmother move their data to centralized block storage or NAS filers. I see data centers planned where none of the servers even have internal disks for OS nor swap. How do you back your claim that JBODS and internal disks are on the rise? Any statistics on that?

"pNFS is the next new technology, that may be problematic for the future of dedicated storage controllers."

You know, I really feel sympathy for Sun and its employees but this day dreaming stuff does not help.
Anonymous on Aug 4 2009, 12:15

You already said it: Last 10 years. 10 years is a really long time.

I´m talking about the next 5-10 years ... the requirements for storage are in flux at the moment. And by the way: I didn´t talked about centralized storage, i just talked about this dedicated RAID storage controllers.

The proof point for internal SSD is quite simple: Just look at the IOPS number of SSD and the latency introduced by SAN and Storage controllers. SSD have to be in the server. Otherwise you will loose 1/3 of the IOPS by latency alone for sync writes.

JBODs will be on the rise with filesystem technologies that can do the job of storage management better then dedicated storage devices. For Solaris ZFS, for Linux btrfs and i´m sure Windows will have similar tools in the future.

You keep the storage centralized, but use internal disks or SSD do hide the latency. By using technologies as hybrid storage pools this is a feasible architecture.

pNFS isn´t problematic because of the market share, you´ve took this wrong, you just can´t sell storage array controllers there, as you have a different model here. As it´s distributed in nature, you simple don´t need them Same for other parallel filesystems like pCIFS. pNFS is an upcoming storage technology not just at Sun, it´s coming from everywhere.

At the end we are in the same situation than we had with high-end servers a few years ago: The big storage boxes will be commoditized by smaller concepts. pNFS/pCIFS is to storage what Beowulf was to the highend number crunchers.
Anonymous on Aug 4 2009, 12:59

"And by the way: I didn´t talked about centralized storage, i just talked about this dedicated RAID storage controllers. "

OK, so you mean centralized JBOD storage is on the rise. I think I have seen some of them Storagetek or was it Storedge JBOD boxes bundled with Sun servers. I think they even had FC connection. But I fail to see that these systems are gaining any momentum in the data centre.

"SSD have to be in the server. Otherwise you will loose 1/3 of the IOPS by latency alone for sync writes."

Not correct. On centralized storage several servers will have LUNS on the same SSD thus increasing the utilization. Internal disks are a receipe for poor utilization both capacity wise and IOPS wise. This is one of the drivers for centralized storage in the first place.

"JBODs will be on the rise with filesystem technologies that can do the job of storage management better then dedicated storage devices"

Yeah, you probably know that there are very few data centres with Solaris only servers. So if all storage management and logic should be done from the server, you will have half a dozen different ways of doing it and systems that can't do any decent storage management at all. Again we have a reason for the increasing popularity of centralized storage management.

And then we have DR. With your model DR must be done from each individual server and on different systems which probably is not possible and certainly not feasible. A management nightmare compared to the relative ease of aynchronous mirroring of an enterprise storage cabinet. Another driver for centralized storage.

As for SAN IOPS. A modern SAN has the capacity for several hundred thousands of IOPS, more than any server or application currently can utilize. SAN technology has always been ahead of server and disks when it comes to performance. Things have changed now with SSD so that some back end controllers will have difficulty keeping up but that gap will be closed.

Hybrid storage. Isn't that really a solution for a problem that didn't exist before ZFS? From what I have heard the design of ZFS creates so much overhead that SSD for meta data is a must to avoid poor performance.
Anonymous on Aug 4 2009, 16:24

All the stuff you write, gives me the impression, that you look at SSD from a rather oldfashioned way. This limits the usage of SSD vastly. You think about it as a substitute for rotating rust, I think of it as a augmentation.

I can just assume, that you think this way, as your primary area of interest is the IBM world. It's just a guess, but the nick lparvirt looks like a hint in this directory. The IBM world misses a technology like the hybrid storage pool, and thus I just can assume that those concepts are not really known to you.

Yes, the L2ARC on SSD solves a challenge. It helps to solve the challenge of random writes on a system that is read sequentially later on. A challenge all COW filesystems have. Brtfs will have it in the same way. SSD can reduce this effect, but this was only a minor reason to introduce the hybrid storage pools. The main idea was to use SSD in normal system transparently for user and application in a way that uses the strength of rotating rust (big size at low price) and SSD (lower latency) to hide their respective weaknesses (rust: slow, ssd: low capacity for the buck). This is a basic idea. (I can just speculate about the source of the idea that we need it for metadata, but i have some ideas that have to be seen in conjunction with your nick lparvirt)
The L2ARC is not a metadata cache, albeit you can configure it this way. The L2ARC is a second level of the adaptive replacement cache of ZFS, thus it enables you to increase the size of the data (and not just metadata) cache. Everything contained in the cache doesn't have to be transmitted over the wires of the SAN and keeps it free for different stuff. Furthermore it reduces the latency of the read request: Reads are synchronous by nature. It's a difference if every single access has to go 2 microseconds locally or 10 microseconds remotely.

Of course it's possible that a single server doesn't have enough load to saturate a single SSD but you should take into consideration that storage utilization isn't a value in itself. Of course it's possible that you can use the IOPS not used by a server due to latency issues by using it from another machine. But you have to think about this: 8 millicseconds is a long time for an SSD but it's an incredible long time for a CPU. Placing the SSD in the central storage may increase the utilization of your storage, but it reduces the utilization of your servers, as it has to wait longer on data to process.

We don't talk about SSD at the costs of 10.000 € any longer, we talk about SSD that are available at 300€ (Intel X25-E). At this price an SSD in conjunction with hybrid storage pools can reduce the load to the SAN vastly and thus reduce the costs and complexity. Denying the value of in-server SSD would be like denying the value of using server memory as cache. The SSD in the server for caching purposes enables you to use larger disks to store your data without sacrifying performance and as the SSD shaves away a large amount of I/O operations in conjunction with the RAM caches it keeps the disks in an utilization pattern that doesn't kill the disks in a short time.

By the way: A L2ARC SSD is never underutilized in regard of the capacity as long as the dataset of the system is larger than the SSD. And even when it's smaller you could drive the utilization higher by using a part of it for swap space for instance (which would load the SSD on the IOPS side, thus driving utilization on the IOPS side) Swapping isn't a performance-killer in itself. The head movements kill you. But SSD doesn't know this problem, thus it would be a valid alternative to use SSD to give a server a virtual memory with access characteristics that doesn't trash the system.SSD in the server can even help you to shave away the latencies introduced by iSCSI oder FcoE and making such protocols a valid alternative as the network is just used for a vastly reduce amount of requests.

Yes, obviously your view was correct for a long time and I see the validity of your comments about the rising complexity of the management but you have to take into consideration that the technology of storage is changing and even more important the models to think about storage organisation are changing, too. You have to take into consideration that the upcoming sizes of data sets doesn't allow dataservices without knowledge of the internal structure of the data and just looking at it as a horde of blocks instead. Recovering a 4 TB disk in the future needs more than just RAID calculcations, it needs knowledge what to recover first and knowledge where data really is on your disk to keep recovery times in reasonable limits.
Anonymous on Aug 4 2009, 19:41

OK. You want server side DR and storage management but at least you see the challenges with it. I think centralized management and DR is the way to go and the trend seems to be with me on this one. I see you have to make some theoretical sacrifice when it comes to speed but for now I don't think that is of any significance compared to the enormous benefits for the customer. I also think that the storage systems will be improved now that SSDs may saturate the pipe line from cache to disks in storage systems(how serious this is differs between manufacturers).

When it comes to SSDs on the server I am all for that as long as it doesn't mess with centralized DR ie. that you have to take special server side considerations when you do DR planning. So if you can increase read performance by putting a SSD on the server without interfering with centralized storage management and DR, I think it is a very good idea. I have read up on L2ARC etc. but I haven't been able to tell if the SSD only will hold persistant data. If you can rip out the SSD and not lose any data it is OK and should play well with centralized DR. Maybe you can answer that one for me.

On a general note. I see that SUN does not like SAN ie. centralized RAID storage but why make it so difficult for the many customers where EMC/HDS/HP SAN storage is used? I mean, if this is a way to convince the customer to scrap the storage systems and turn to JBODS it will not succeed. Every thought put into ZFS is focused on JBODS, when I ask Sun people about the best use of ZFS with EMC devices they go blank. Every Powerpath device has to be manually partitioned to be included in a zpool for example.

So to be a little harsh. It seems to me that one of the problems with SUN is that they don't make solutions which are optimal to the situation where customers _are_, but rather make solutions for situations where SUN think that customers _should be in_. This rhymes well with your thoughts on server side storage management.
Anonymous on Aug 5 2009, 15:39

At first: Albeit ZFS is directed to JBODs, it's not exclusive and many of my and others colleagues customers use it on large disk arrays.

Regarding L2ARC (Level 2 Adaptive Replacement Cache) suggests, it's a cache: So there is no data on the L2ARC that isn't on the pool as well. It's an extension to the ARC in the main memory. So you can keep your disks in the central storage, but keep the the SSD as close as possible to the CPU.

In regard of seperated ZIL it's a little bit more complex: When you don't need to do a failover your service to a different node, you can keep the SSD for the sZIL in the server as well and shave of several microseconds per write I/O, when you want to make a cluster failover, you obviously need to failover the sZIL as well. You could solve that by putting the SSD in the central storage or by sharing the the SSD just between the cluster nodes (when you want to shave off the microseconds introduced by larger SANs)

At the now to something forward looking: It looks like, you don't want central storage, you just want central administration of the storage. You don't want to manage replication on each server, you want a central instance to control it. But you don't need a central storage array for it, you just need intelligent management or a storage protocol that includes such behaviours. That's a large difference.

I see two possible directions: Larger and larger storage arrays to keep up with the increasing requirements or more intelligent management of distributed systems and a compute-cluster approach to storage. I assume the next 10 years will show us, where the industry will head to. And besides enterprise storage there are other developments of handling storage for example in HPC or in Web Services. Just look at systems like Hadoop, the processing and storage of data is solved completely different, but it's a valid and in some cases even vastly more effective way to store data.

In 5-10 years the way we look at storage will look different, and i'm not sure that large disk array controllers will have the same role than today. They will have their strongholds ... but thats strikingly similar to the stuff that happend to high end unix.
Anonymous on Aug 5 2009, 22:02

I want central storage because that makes central storage management possible. What you suggests(decentralized functionality and centralized control over a heterogeneous environment) does not exist and I have heard of no one that wants it either except Sun. And lets face it, Sun-servers(or Power servers for that matter) does not dictate the parameters for the storage infrastructure in the data centre, that time is long gone.

Btw. you haven't yet justified your claim that the trend has changed direction from raided storage towards JBODs. Willing to admit that it was pure fantasy or that you forgot to look outside the small world of Solaris/ZFS?
Anonymous on Aug 6 2009, 09:40

Of course HDS or EMC isn't willing to go this way, it would ruin their business modell.

And again: I didn't talked about about yesterday, i don't talk of today, but about not so far away future. Of course there is not such a centralized, heterogenous storage management.

By the way, even EMC goes in the direction of distributed architectures with vmax. NetApp is said to develop something similar for their next-gen filers. It would be just a logical step to participate on the compute and i/o power of the servers to get rid of the potential bottleneck with the storage array controller as a singleton. But this step would not come from a vendor that earns money by selling storage controllers, it will come from one of the OS vendors or from a server vendor. Time will tell.

I have some insight into some Linux shops (outside of my job), and many of those people are not fond of purchasing some big storage boxes. Willing to admit that there is a market besides high end enterprise?
Anonymous on Aug 6 2009, 10:34

Question: Isn't it true that an x86+SSD+JBOD hybrid pool system can serve as a centralized storage system (Both block and file interface)? Why "lparvirt" and alike think it is per server only? The argument is can an x86+SSD+JBOD (ZFS is the glue among them) outperform a traditional RAID controller+Specialized Raid subsystems(interfaces, interconnects, frontend/backend etc) and be more cost effective at the same time? In terms of SSD, in RAID subsystem, you create a "tier 0" on SSDs as individual LUNs to I/O savvy hosts, while in x86+SSD+JBOD system, the SSD can be benefit to all LUNs (in all tiers) as they are the Level 2 cache. For this reason, I think the SSD is a better fit for the same bucks in the latter system, any thoughts?
Anonymous on Aug 6 2009, 22:40