Thursday, October 25. 2012

Disks

Just to get back to my article a few days ago: I really think that we won't tapes disappearing whatever some media pundits write all the time, especially not the areas where tape drives like the T10000 are used for. In fact tapes will get more important, because of a wall that is imminent. Yeah ... you may think "Sure, those limits were always broken in the past". But here we speak of a limit that is imposed by the older brother of computer science, physics. However this article isn't about tape and why i think we need them more than ever. I just want to show with this blog entry that the future in magnetic storage isn't that easy as you might think due to several years of steady increases in disks.

Just a few words about this blog entry: Since i'm an avid supporter of ZFS, i'm reading a lot about data storage and try to know as much as possible. It's a really interesting topic. When you think that a hard drive is a relatively simple device, just forget this assumption.

For this article I simplified stuff for this articles, at first to keep this article relatively short. Additional I used the same simplification that helped me to understand this topic. However: Perhaps i forgot something or i'm just wrong with what i've learned. It's just my summarization that was inspired by a internal presentation i recently read. On the other side I will reread the text in the next days and may make some change when i'm detecting that i wrote something incorrectly or unclear.

Rotating rust

When you use rotating rust devices to store data, you are doing so by magnetizing matter on a platter. It's pretty much that simple. But the devil is in the detail.

At first it's not a uniform piece of matter like a single crystal, you write to an incredibly high number of incredibly small parts of it. This small parts are called grains. This grains are random in location, density and volume. And it's not that way, that with current technology 1 grain is a bit, a bit consists out of several grains.

You may think, okay ... let's just make the grains smaller, so we have more density, but you don't have look farther than the wikipedia and look at "superparamagnetic limit" to know that you have a problem.

With current technology you can't increase the density infinitely. At some point the magnetized grains get so small, that they loose the information you've store just by thermal effects. Perhaps it doesn't happen right after writing, but hard disks are designed for a retention time of 10 years (writing something on the disk, putting it without power into a safe for 10 years, reading it). And that's a long time for bits to flip or for grains flipping their polarity.

By using perpendicular recording the problem was circumvented and purchased the industry some time. Instead making the magnetized areas smaller, they were differently oriented. Instead of longitudinal, you've stored the data perpendicular. Or to say it differently, the read/write head sees the small side, not the wide side of the area when you imagine the grain a rectangular box. However essentially the size of the magnetized area was the same.

Interestingly this lead to other problems being more important like the high fly write. In order to magnetize such perpendicular items you need an immense field strength at the right location. But when the write head is flying to high out of whatever reason (perhaps shock), it may not magnetize the grain but just the air above it. But the drive electronics thinks it has written. Rust can't compute and thus the media can't tell you that it hasn't received the data. Only way would be to make read-after-write. This is the reason why disk started to implement flight-height monitoring. But that's a different story.

The trilemma

Searching for alternatives is significantly harder than you might think. When you are searching for new technology have a problem called magnetic storage trilemma - or have to find a technology mediating between readability, writeability and stability. You want to have material that is able to keep the information stored by magnetizing it as long as possible with as less matter needed per bit. A material with high coercivity would be fine for that. However a material like that is hard to magnetize, or to speak in data storage terms. It's hard to write to it with acceptable field strengths. And when the magnetic field stored on the media isn't strong enough to be read, you have another problem.

To make it short: Current technology hits a brick wall around 1 TB/square inch. Out of this reason companies are researching for quite some time for alternatives. I want to describe a few of them and want to describe their challenges

as well.

Shingled recording.
One example Shingled Write Recording. This has the advantage to increase density without totally changing the process to manufacture a hard disk. The idea is quite simple. When you write to the platter you, write a quite wide track on the disk, when you write the next one, you move the write head only a fraction of the track size to the side. The data is stored in the area that isn't overwritten in this second round. To get it clearer: Just look at the next roof with roof shingles or roof tiles. The data is on the part visible. And as everyone knows who had to purchase roof tiles. There are a lot of more rows of roof tiles on the roof as when you would simply divide the length of the side of the roof by the length of the roof tile

There is just a problem you can't simple remove a tile. You have to move the tile above it to the top in order to be able to take out the tile. And it's the same with shingled write recording. You can't simply overwrite the data a location because of the shingled method. So you have a hard disk that has a quite unusual access pattern. You can randomly read, you can write only sequentially.

So essentially you've kept the manufacturing process, but either you have to make the drive controller much more intelligent (something in the area like the SSD) or you have to change the filesystems. Copy-on-write-filesystems like ZFS may have a significant advantage in using them, as totally writing sequentially is just a corner case of their usual operation.

But you would have a IOPS inflation problem here: Assume you have some database blocks in a track, you want to change a block, you can't overwrite the old block because of the shingled recording, you have to write it at a new location and a formerly sequentially read pattern is now a random one. Furthermore you have to do garbage collection, because when you have sequentially written to the end of the disk, you have to start at the beginning but that has to be freed from any data that was initially stored. And then the issues with a garbage collection in the background are starting to appear.

The big advantage of shingled recording is twofold: At first you get more tracks per inch. That's obvious if each track is less wide, you get more of them on the same part of the rotating rust. Thus have more capacity alone by that. You can use much stronger magnetic fields to write, thus using a material with a much higher coercivity, thus using much smaller areas representing a bit of information, thus increasing the density and thus increasing the capacity of a drive with a given form factor, too. So you get a higher areal density and thus larger capacities. And that's a good thing.

Energy assisted magnetic recording methods
On the other side you will find technologies like heat assisted magnetic recording (HAMR) or microwave assisted magneting recording (MAMR), they are summarized under energy assisted magnetic recording. Both are working on the idea, that it gets easier to magnetize them when you apply a form of additional energy to the part of the media you want to change.

For example heat assisted magnetic recording is based on the fact that materials have a high coercivity at room temperature, but have a much lower coercivity when influenced by the assisting form of energy. So with heat assisted recording, you heat a small point on the media above the curie temperature , the media looses coercivity at this location and in this moment you are able to change the polarity in exactly this location with normal field strengths while leaving unheated parts unaffected. When the point cools down, it retains the new polarity and the coercivity increases again. So you are able to magnetize materials that wouldn't be magnetizable at room temperature.

MAMR ... well i have to admit that i don't understand MAMR so far completely, but it looks like that when using MAMR the magnetic field isn't capable to change polarization alone. However the energy of the microwave leads to an increasing precession in the direction to up the point that the polarisation flips.

In my simple thoughts the the microwave energy is like sledgehammer used with gentle, but very deceive force to tell something that it should look into the direction you tell it and you tell it by the magnetic field, while the thermal energy in HAMR is more to induce an amnesia at the point with a sledgehammer, so the magnetic field has an easier job to tell the focussed point which magnetic polarization it has

I've read a lot of documents that people consider HAMR respectively MAMR as the next step forward. At the moment MAMR is considered as an easier to archive technology because the changes to the media and the head are not that big. However a lot of research has to be done. HAMR is something different ... you have to combine a new media with a head that combines laser and the write head.

However i see problems here as well. With HAMR when you laser fails to heat the media above the curie temperature, you can apply your magnetic field ... it will change nothing. So you have to correctly apply the thermal energy and the magnetic field or you have a phantom write. A lot of assumptions so i guess, that we end up with read after write to check that everything went well, but that would be problematic for the write latency because writing takes twice as long as before.

Futhermore in my simple thoughts I would assume that you can heat and cool down a material just a finite number of times before changing the structure of the material in a way that hinders you to store data securely. I would assume that you have some kind of wear on the material, thus you have to introduce some kind of wear-leveling into the rotating rust hard disks.

Media
A completely different approach is the "bit patterned media" stuff. With this technology the hard disk manufacture creates a pattern on the media that represents islands of magnetic substance with known location, shape and volume. Each of this islands represent a bit. The rest of the media is non-magnetic.

This is a contrast to the conventional media with it's randomly distributed, randomly sized and randomly shaped magnetic grains. When using such a media you have to magnetize several grains to store a bit in order to ensure the safe storage of data.

With this controlled pattern you can reach much higher densities than with the conventional data.

But that has it's disadvantages as well, you have to exactly time the fly-by of such an island under the head with the magnetic field to magnetize the island and not the space between the islands. Some documents suggest that such a method would need read-after-write as well, just to be sure that you've really written data and did not tried to magnetize non-magnetic matter.

However the main disadvantage of this technology has nothing to do with the disk in itself , it's the way the media has to be manufactured, this is significantly different to the way disks are manufactured nowadays, thus leading significant investments on the manufacturers side and those have to refinanced by the customers buying such disks. However there is a lot of research underway in order to reduce the needed changes in the manufacturing process.

2D
So far I wrote about writing things, but when you look at disks there are changes imminent as well, how data is read from the rotating rust. At the moment you read the magnetic fields into a waveform that is interpreted by statistical algorithms in the most likely digital pattern. However with ever increasing density of the tracks you run into the problem that each track can interfere with the neighboring tracks on a multitude of ways. When you have just the single waveform of your track, this interference is disastrous and your error-checking/correcting codes have to kick in.

The idea of 2D reading is to have information about the adjacent tracks in order to cancel out the interference of those tracks by mathematical means.

While this is a great thing for increasing the areal density, it has an obvious disadvantage. You have read the adjacent tracks in order to know what's there.There are two possibilities: You design a quite complex read head that is able to read multiple tracks at once, or you simply read the adjacent tracks as well sequentially. However in the second case the latency of the reading is increased by the need for additional rotations of the rotating rust.

Combining
In the discussion about future disk technologies there are a lot of technologies available and when you read into a literature there is a lot of discussion about combining those technologies, like shingled writing with 2D readout, calling this TDMR or two dimensional magnetic recording. This combined method is going even further than both technologies separated. This method is working with the knowledge that you don't just read a single track but have more broader information about the distribution of magnetized grains, the way they are polarized, so it can use this information to reduce the number of grains needed to store data with the remote and theoretical target of storing one bit in a single grain.

But the combination is not just limited to shingled recording. With other write mechanisms technologies changes in reading the data get a more important step in order to cancel out side reading (read heads aren't as perfect as in the theoretical model, so they start to detect the tracks at the side as well, when the tracks get really dense, however it looks like that this problem is more prevalent with some patterns than with others) or to cancel out inter-track interference. With one dimensial reading you don't know if it's a real signal or just side reading. So there is for example some

discussion about integrating 2D reading with bit patternd media.

Hybrid
Given that most of the new storage methods have some kind of latency disadvantages either on the write or read side (or both) i assume we will see much more hybrid approaches in the future. Either by application (for example Smart Flash Cache/Smart Flash Log in Exadata), by operating system (ZFS Hybrid Storage Pool) or by the drive electronics (you may remember that there is a kind of drive having stationary dust and rotating rust in one drive (of course the controller has to hide it behind one drive interface in the sense of the LUN, to that read caching and write caching on flash are transparent). Such hybrid approaches would hide the added latencies due new technology.

Why i'm writing all this
I'm pretty sure, that the brick wall will be broken at some time, but when you look at the methods, i looks like none is really easy, all have their own set of problems. No insurmountable obstacles, however i really think that the times of seemingly easily ever increasing capacity on hard disks are over. And given the storage technologies manufacturers and the scientific community is researching on, the behavior of future hard disks may significantly different than today leading to completely different performance characteristics.

At the end this differences may lead to an increased usage of technologies that are around for decades now, but somewhat unknown outside large installations - for example hierarchical storage management. A few years ago you had two major storage systems: Disk and tape, SSD and optical drives were somewhat exotic. Now we have SSD, disk and tape as media. In the future we have SSD, disk with random read/random write behavior, disks with random read/sequential write behavior and tapes. It get more and more problematic to manage that all manually and in this light hybrid approaches like the hybrid storage pool or outright hierarchical storage management systems like SamFS may come quite handy. And at the end it's one reason why i'm keep saying that you shouldn't throw away the concept of a rotating rust only hybrid storage pool with ZFS too fast. Just think about a hybrid storage pool with L2ARC and separated ZIL on normal disks and the storage pool on disks using this ultra-high capacity magnetic storage technologies.

Will be even more interesting times for us to explain, why this single large high capacity hard disk is even more not enough for the central database. And even harder to explain to the people in purchasing

Posted by Joerg Moellenkamp in English, Technology, The IT Business at 22:24