Feeding the Pelican: using archival hard drives for cold storage racks

Richard Black
Microsoft Research

Austin Donnelly
Microsoft Research

Dave Harper
Microsoft Research

Aaron Ogus
Microsoft

Antony Rowstron
Microsoft Research

Abstract

Microsoft's Pelican storage rack uses a new class of hard disk drive (HDD), known by vendors as archival class HDD. These HDDs are explicitly designed to store cooler and archival data, differing from existing HDDs by trading performance for cost. Our early Pelican experiences have helped some vendors define the particular characteristics of this class of drive. During the last twelve or so months we have gained considerable data on how these drives perform in Pelicans and in this paper we present data gathered from a test and a production environment. A key design choice for Pelican was to have only a small fraction of the HDDs concurrently spun up making Pelican a harsh environment to operate a HDD. We present data showing how the drives have been used, their power profile, their AFR, and conclude by discussing some issues for the future of these archive HDDs. As flash capacities increase eventually all HDDs will be archive class, so understanding their characteristics is of wide interest.

1 Introduction

Many cloud and large-scale data center operators have been exploring cloud storage optimized for colder and archival data, for example Amazon's Glacier Service [1], Facebook's Cold Data Storage [10], Google near-line storage [11] and Microsoft's Pelican [3]. In contrast to online storage [9], this storage trades data access latency and throughput for lower cost; access latencies of multiple seconds to minutes (to even hours) are normal with a per-file throughput that is often restricted.

One way to make these services cost-effective is by using custom designed racks. These designs achieve cost savings by right-provisioning resources at the rack-scale, such as power, cooling, network bandwidth, CPU, memory and disks. Sufficient resources are provided to only support these colder and archival workloads. Examples include Microsoft's Pelican [3] and Facebook's Open Compute Storage [12]. These systems provision insufficient rack-level power to allow all the HDDs to be concurrently spun up; in Pelican it is 8% of the HDDs, with the rest in standby. By implication, the rack cooling is then only provisioned to handle the heat generated from a subset of the HDDs spinning.

This creates a harsh environment for HDDs, where they are frequently spun up and down. Pelican was designed together with a new class of HDD: archival class HDDs. We briefed all the major HDD vendors, and worked closely with one vendor to help specify the performance we wanted from these new HDDs.

We have been running Pelicans for over a year, and in this paper we report on our experiences with this new archival class HDD. We have experimented in our test lab with drives from multiple vendors, and have also profiled drives in a production environment. Anecdotally, many storage experts are skeptical that an HDD can be spun up and down tens of thousands of times per year without significantly increasing failure rates. Has it impacted our disks so far?

We explore multiple metrics of interest. We start with the power profile of eight archive HDDs including both SMR and PMR variants, and note that drives which differ only in their firmware can behave very differently in their power draws. We then look at HDD workload characteristics that could impact failure rates, from the amount of data passing through the head, to the number of power-on-hours, the temperature profile for the drives and the number of spin ups. Finally, we look at the AFR and then discuss interesting options for the future design of these archival class HDDs.

Before exploring their characteristics, we give a short high-level overview of achival class HDDs and Pelican.

Archival class HDDs. An example of such a HDD is the Western Digital Ae model WD6001F4PZ[1] which

[1] `http://www.wdc.com/wdproducts/library/SpecSheet/ENG/2879-800045.pdf` accessed 15th February 2016.

is a 6TB drive operating at 5,400 RPM. The marketing literature states that *"The WD Ae hard drive is best suited for cold storage, backup and data archiving where data is stored on disk but rarely if almost never read again."* All the main vendors have equivalent HDDs, for example Seagate's ST8000AS0022[2].

Pelican uses these archive drives because both Pelican and the HDD vendors are optimising for lowest cost per GB of capacity. These HDDs have three basic characteristics that impact lifetime, usually specified per year: power-on-hours (POH), head wear in terms of data transferred across the head, and number of controlled spin downs. POH is defined as the number of hours when the HDD is active, i.e. not in idle_B (heads unloaded) or deeper power saving modes, including: idle_C (low RPM), standby, or powered off. Head wear is important as HDDs use a thermal actuator to move the head to a lower fly height while writing and reading [13]. At the lower height, the risk of damaging interactions between the head and the media is elevated. "TBs transferred" is used as a proxy for the length of time the heads have been at the lower fly height.

Pelican. A Pelican storage rack has 1,152 HDDs and two servers. Each HDD is connected to a SATA 4-port multiplier, which is connected to a 4-port SATA HBA. Pelican uses PCIe to connect the 72 HBAs to the servers, such that each HBA can be attached to either server. There is sufficient power and cooling provisioned to allow only a small fraction of the HDDs to be spinning and performing IO (active) while the other HDDs are spun down (standby).

A Pelican power domain contains 16 HDDs and has sufficient power to support two HDDs transitioning from standby to active, with the 14 other HDDs in standby. A Pelican cooling domain has 12 HDDs, each in a different power domain, and can provide sufficient heat dissipation to support one HDD transitioning from standby to active and 11 in standby. These domains represent constraints on which HDDs can be concurrently spinning, imposed by the physical rack design. The cooling domain means that at most 96 HDDs can be concurrently active in a Pelican, one per cooling domain. The power domain means that two per tray can be spinning, further restricting the choice about which HDDS can be concurrently active.

To handle this HDDs are placed into 48 sets of 24 HDDs, referred to as groups. Each blob stored in a Pelican is striped across 21 HDDs within a single group, using 18+3 erasure coding. In order to read or write to a group, the group needs to be spun up before the HDD can be accessed. Due to the power and cooling constraints at most 4 groups can be concurrently spun up. In a Pelican spin up is the new seek latency, and the software stack needs to schedule IOs to maximize throughput while minimizing impact on per IO latency. For full details please refer to Pelican [3].

[2] `http://www.seagate.com/www-content/product-content/hdd-fam/seagate-archive-hdd/en-us/docs/100795782a.pdf` accessed 2nd March 2016.

2 Studying Archival HDDs

The data shown in this section is gathered from two different deployment environments. The first is a Pelican operated in a test lab used to experiment with the hardware and evaluate HDD products. This test rack is populated with HDDs from multiple vendors and, in some cases, with multiple versions of the same HDD. Data presented from this environment will be labelled as *test*. The second data set is from a deployment in a production environment, which will be labelled as *production*.

We start by describing the different HDDs used in the test and production environments. We use eight different kinds of HDD from three different vendors, referred to as vendors A, B and C. We refer to the HDD kinds as A1, A2, B1, B2, B3v1, B3v2, B4 and C1. It should be noted that B3v1 and B3v2 use the same physical drive hardware and only differ in their firmware, B3v2 aims to achieve lower spin up latency and has a modified power draw profile. We worked closely with vendor B to develop their archival class HDD, hence they represent five of the HDDs used.

Table 1 summarizes the key features of each of the HDDs. Five of the HDDs use PMR [14], while three use SMR [6]. The labels HA and Auto denote Host-Aware or Autonomous SMR, respectively [6], where HA is an autonomous HDD-managed mode with a Host Aware Zone feature set [2]. This allows a host OS to determine it uses SMR and optimize the file system to enable better performance. C1 (we believe) is an autonomous drive-managed SMR HDD. We infer this from the power profile and delay between the platters spinning and being ready, and there is an additional power draw during spin down. However, the vendor has never confirmed it is an SMR HDD.

All B HDDs are ranked by release date, so B1 was the first and B4 the latest. B3v1 is an updated version of B2, with increased capacity and reduced spin up power draw, but this increases the spin up time. B4 is a larger capacity version of B3, and also has a reduced power draw when idle. B1 through B3 are experimental HDDs which are not generally available. We include them to help understand how these HDDs have evolved.

2.1 Spin up

In Pelican, in order to read data from a set of HDDs, they will normally need to be spun up. The spin up latency

Name	Technology	Spin up (s)	Capacity (TB)
A1	Auto SMR	10.1	8.0
A2	HA SMR	10.2	8.0
B1	PMR	7.9	4.6
B2	PMR	7.8	4.5
B3v1	PMR	9	4.9
B3v2	PMR	6	4.9
B4	PMR	6.4	6.1
C1	Auto SMR (?)	8.6	8.0

Table 1: Summary of HDDs, spin up time in seconds.

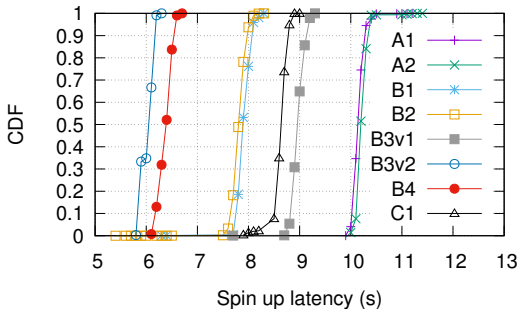


Figure 1: Spin up latency.

represents a lower bound on the time to first byte for any read request which accesses data on HDDs that are spun down.

Figure 1 shows a CDF of spin up latency for all eight kinds of HDD. We use data gathered from several samples of each kind of HDD. The spin up times are grouped into 100ms buckets and their CDF plotted. For any HDD kind the minimum number of spin ups is 342, and the maximum is 114,727. There are two key observations: *(i)* the distribution is tightly centered around the mean, *(ii)* there is about a factor of 1.8 between the fastest and slowest. To the first approximation the spin up latency is a function of the power draw and the required RPM for the platters.

We therefore ran an experiment where we measured the power draw for each HDD through a spin up, idle, and then spin down cycle. We measure the power relative to each HDD's standby power. Figure 2 shows the power draw for each of the HDDs versus time. At time zero, each HDD is sent a spin up command. The colored upwards-pointing arrows show the time when each HDD's spin up completion is received by the OS. Each HDD then remains idle until 18 seconds when the filesystems are unmounted and the HDD is sent a *standby immediate* command. The colored downwards-pointing arrows show when the *standby immediate* completes.

Figure 2 shows a number of interesting events. After A1 and A2 have completed their power draw for spin up, they run at active power levels for a further two seconds before completing the spin up to the host OS. Since they use SMR we speculate that they are loading internal state (e.g. remap tables) from the media. They also take longer to spin down, especially A2 which uses host-aware SMR.

The B results show the progress made from early versions of the HDD to the latest (e.g. B3v2 and B4). From Figure 1 we see a reduction in spin up time, and from Figure 2 we see the power draw reduced while in idle. Interestingly, B1 and B2 *complete* the spin down command but continue to draw elevated levels of power for a little over two seconds before going completely idle. We speculate that this may be due to the HDD electronics remaining in a higher-power mode until some timeout triggers, and this behavior does not occur after B2. It is also interesting that B3v1 and B3v2 only have different firmware revisions, yet they behave very differently in their power draws and thus spin up times. B3v1 limits its peak draw, but this results in a much longer spin up time: 9 seconds, compared to 6 seconds for B3v2.

2.2 Using Archival Class HDDs

We have B3v2 HDDs currently deployed in a production environment, and now look at the stress Pelican puts on them. We gathered detailed statistics from 7 May 2015 to 21 December 2015 (228 days) for a constant set of HDDs and scale to 365 days to aid comparison with HDD parameters which are quoted per year. There are four basic metrics that impact the failure rate of these HDDs: *(i)* the number of spin up spin down cycles, *(ii)* the volume of data transferred across the head, *(iii)* the number of power-on-hours, and *(iv)* temperature. We now look at these metrics for a Pelican.

Figure 3(a) shows a normalized distribution of HDDs versus spin ups per year. The median HDD does 70,800 spin ups and the maximum is 100,000. It should be noted that Pelicans do a "controlled" spin down, i.e. the spindle is stopped before power removal. This is as advised by the vendors and the number of cycles is within what they believe the HDD can handle.

HDD failure can also be induced by head wear. Figure 3(b) shows the distribution of TBs transferred per HDD per year. The strictest limit across all vendors is 60 TB/year for a HDD, and only 0.7% of our population were over. Obviously, this is highly dependent on the workload, and is an important consideration when designing a system. In Pelican we do not do any explicit wear-levelling, although a number of policies are implicitly load aware. We believe that wear-levelling may need to become a first class concern in HDD-based storage.

The next metric of interest is the power-on-hours. Figure 3(c) shows the distribution of power-on-hours for the HDD population. The median is only 165 hours, which is well below the strictest limit from all vendors of 3120

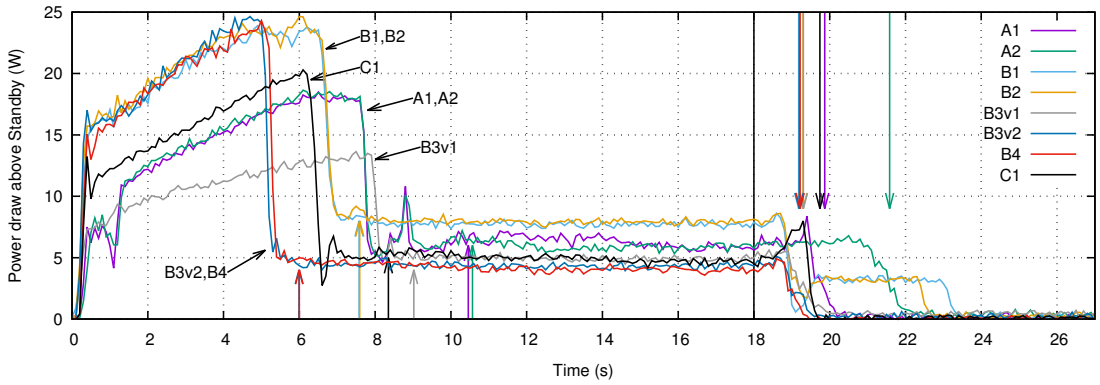
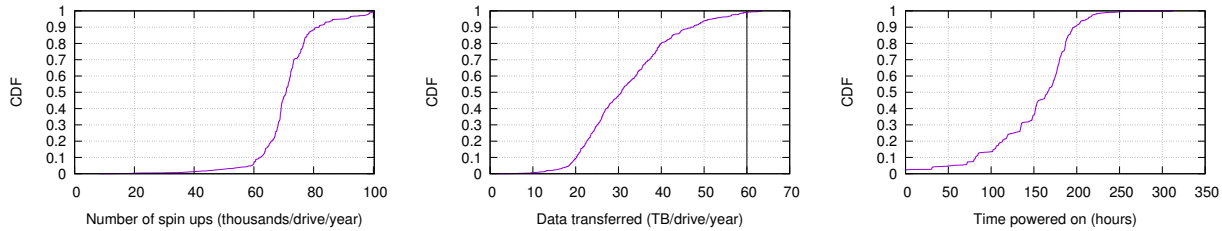


Figure 2: Power draw while spinning up, idle, and spinning down.



(a) Number of spin ups. (b) TBs transferred. (c) Power-on-hours.

Figure 3: CDFs of basic HDD metrics.

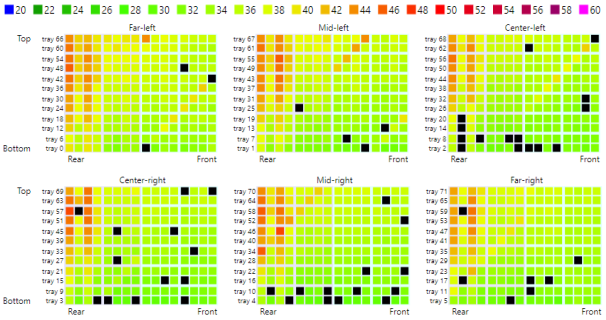


Figure 4: Temperature snapshot from a single rack. The squares are HDDs, colored by their temperature.

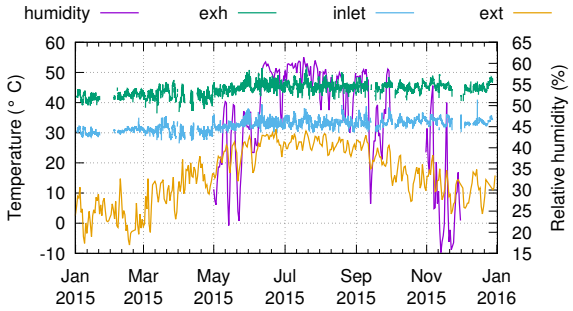


Figure 5: Humidity near rack (%), and temperatures: mean rack exhaust (exh) °C, inlet °C, and datacenter exterior (ext) °C.

hours (i.e. 130 days) per year.

The final metric is temperature. Figure 4 shows a snapshot of the temperatures in a single rack in the production environment on 15 February 2016. A Pelican rack is vertically cooled; cool air enters at the bottom front of the rack, and is exhausted out at the top rear [3]. For this single rack, we have shown as black squares all HDDs that have failed in the rack in the last 12 months. Figure 5 shows mean daily temperature outside the datacenter hosting this rack, as well as the mean hourly inlet and exhaust temperatures across this rack. This rack is in a direct evaporative cooled data center [7], so during the summer months the relative humidity rises, keeping the temperature controlled. Figure 5 also shows the humidity near the rack, but we only have data from May to November 2015. While the exterior temperature fluctuates due to the passing seasons, the inlet temperature is fairly consistent across the year, but the humidity increases with external temperature.

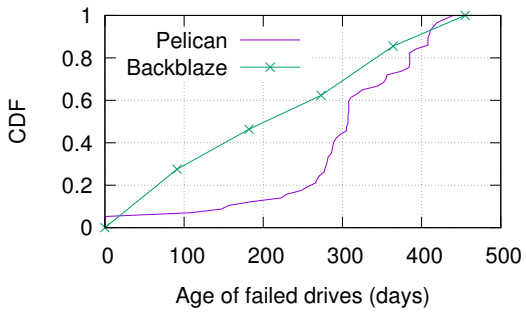


Figure 6: Age of failed HDDs.

2.3 Disk failures

Having looked at potential sources of disk failures, we now report on the actual failures observed. Over the whole deployment for 440 days, the average AFR (annualized failure rate) is 3.96%. Figure 6 shows the normalized cumulative disk failures versus age in days since they were deployed. It also compares our failure rates with those reported by Backblaze [4]. Unlike them, we do not see higher rates of failure for young HDDs.

Surprisingly, our failures are inversely correlated with temperature: the cooler disks are more likely to fail than the hotter disks. This is possibly because the cooler disks are closer to the air inlet vents, and thus see a higher relative humidity than the hotter disks. Correlations between humidity and elevated failure rates have recently been shown [8].

3 Discussion

We believe that our experiences with Pelican and archival class HDDs highlight a number of interesting challenges. The most obvious one is the issue of head wear, and limiting the volume of data transferred across the head. This highlights the need for a deeper understanding of the trade-offs and benefits of scrubbing and the correct strategy for using these low-cost HDDs. We also believe that to use these archival class HDDs we need to think about treating wear-levelling as a first class citizen, as we do for Flash. If we can address these issues this will increase the temperature of data that can be stored on archival class HDDs.

Given our experiences with B3v1 and B3v2, we would like to see vendors expose more control over their HDD performance – an aspiration shared by others [5]. We are increasingly seeing cross-layering being used in the data center, where more control of lower layers increases performance per dollar at a system level. The difference achieved between B3v1 and B3v2 demonstrates how changing the low-level parameters can significantly impact performance. This also extends to the HDDs reporting their current power usage. Pelican uses Power Up In Standby (PUIS) and floats pin 11 of the SATA power connector and yet we cannot always faithfully determine the current state of the drive. Therefore, Pelican also monitors power consumption at a tray-level (16 HDDs) in order to understand the disk states. It would be easier if we could ask the HDDs, and they faithfully tell us!

Acknowledgements

We would like to thank the other original members of the Pelican team, and in particular Shobana Balakrishnan, Adam Glass, Sergey Legtchenko and Eric Petersen.

References

[1] Amazon glacier. `http://aws.amazon.com/glacier/`, August 2012.

[2] AMERICAN NATIONAL STANDARDS INSTITUTE (ANSI). *T13/BSR INCITS 537-201x, Zoned Device ATA Command Set (ZAC), Working Draft Revision 04j*, Nov. 2015.

[3] BALAKRISHNAN, S., BLACK, R., DONNELLY, A., ENGLAND, P., GLASS, A., HARPER, D., LEGTCHENKO, S., OGUS, A., PETERSON, E., AND ROWSTRON, A. Pelican: A Building Block for Exascale Cold Data Storage. In *OSDI* (Oct. 2014).

[4] BEACH, B. How long do disk drives last? `http://www.backblaze.com/blog/how-long-do-disk-drives-last/`, Nov. 2013.

[5] BREWER, E., YING, L., GREENFIELD, L., CYPHER, R., AND T'SO, T. Disks for data centers. Tech. rep., Google, 2016.

[6] DUNN, M. E., AND FELDMAN, T. Shingled magnetic recording - models, standardization, and applications. In *SNIA Storage Developer Conference Tutorial* (2014), SNIA.

[7] GREENBERG, S., MILLS, E., , TSCHUDI, B., RUMSEY, P., AND MYATT, B. Best practices for data centers: Lessons learned from benchmarking 22 data centers. In *14th biennial ACEEE conference on Energy Efficiency in Buildings* (2006).

[8] MANOUSAKIS, I., SANKAR, S., MCKNIGHT, G., NGUYEN, T. D., AND BIANCHINI, R. Environmental conditions and disk reliability in free-cooled datacenters. In *FAST'16* (Feb. 2016), USENIX.

[9] MARCH, A. Storage pod 4.0: Direct wire drives - faster, simpler, and less expensive. `http://blog.backblaze.com/2014/03/19/backblaze-storage-pod-4/`, March 2014.

[10] MORGAN, T. P. Facebook loads up innovative cold storage datacenter. `http://www.enterprisetech.com/2013/10/25/facebook-loads-innovative-cold-storage-datacenter/`, October 2013.

[11] NEWSON, P. Whitepaper: Google cloud storage nearline. `https://cloud.google.com/files/GoogleCloudStorageNearline.pdf`, March 2015.

[12] OPEN COMPUTE STORAGE. `http://www.opencompute.org/projects/storage/`.

[13] Disk read-and-write head. `https://en.wikipedia.org/wiki/Disk_read-and-write_head`, Feb. 2016.

[14] Perpendicular recording. `https://en.wikipedia.org/wiki/Perpendicular_recording`, Feb. 2016.