MAXIMIZING DATA'S POTENTIAL

Lowering Costs

May 10, 2018

SEAGATE

A Bit About Your Speaker

Currently has the business management for Seagate enterprise cloud customers, and world wide enterprise distribution business (pricing, contracts and business management)

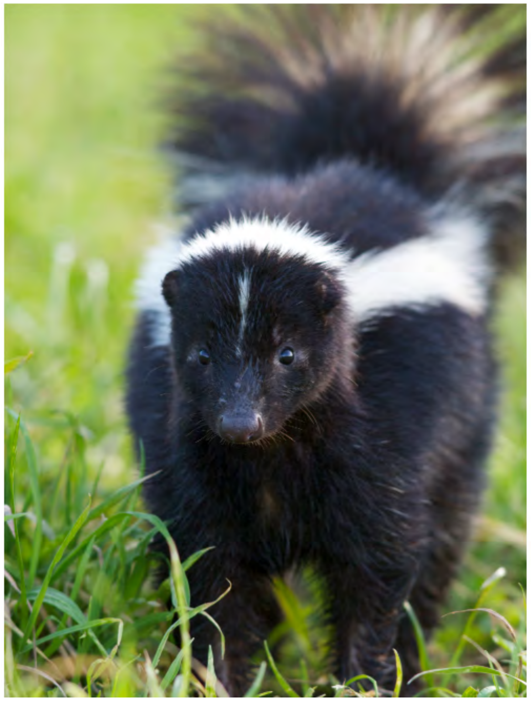
Backblaze is one of our prized customers and asked me to present

In recent history, I've had responsibilities for 80 member engineering groups customer co-design groups, with labs in Shanghai, Taipei, and Redmond, Washington

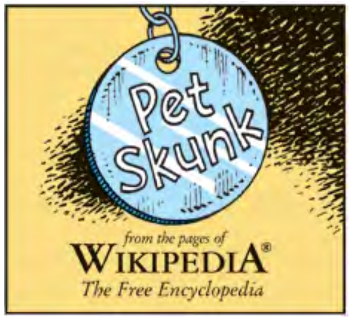
Also had over 15 years of supervising representatives to the industry standard committees for both T10 (SCSI), T13 (SATA), and the SATA-IO working group

Ted Deffenbaugh

HDDs: The Skunks Of The Industry



I had one of my HDDs co-workers come to me the other day and say, “Man, we don’t get enough credit, some times I think we’re treated like a skunk.”



I said that I always wanted a pet skunk as a kid growing up. They are legal in 17 states, and their owners like them quite a lot.

HDDs will return just as much love as a pet skunk and have the added benefit of being legal in all 50 states

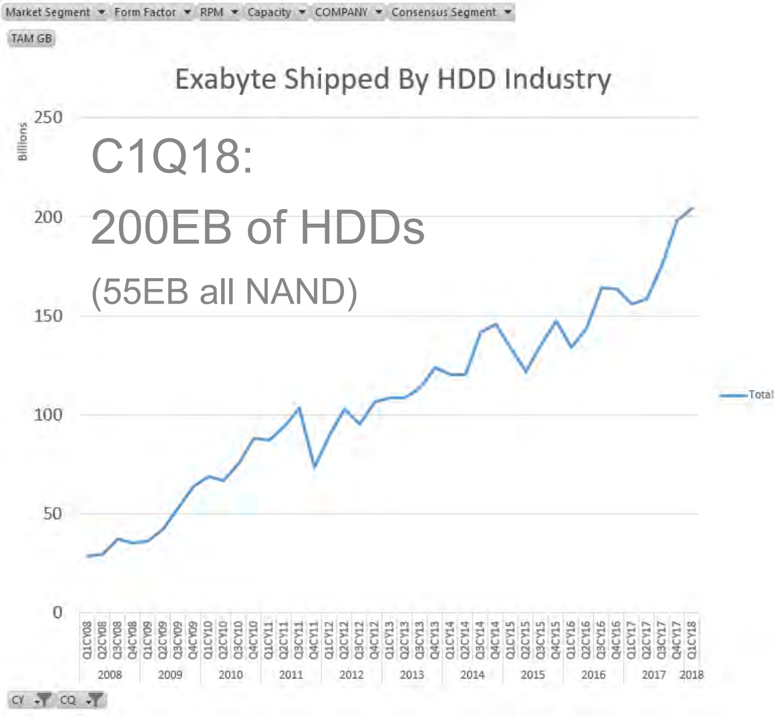
Thus I’m Nominating Skunks As The Official Mascot Of HDDs

Put A Skunk In You Data Center

A pet skunk is a skunk kept by humans for companionship and enjoyment.
I do a lot of things really, really well

The Great Migration: HDDs Into The Data Center

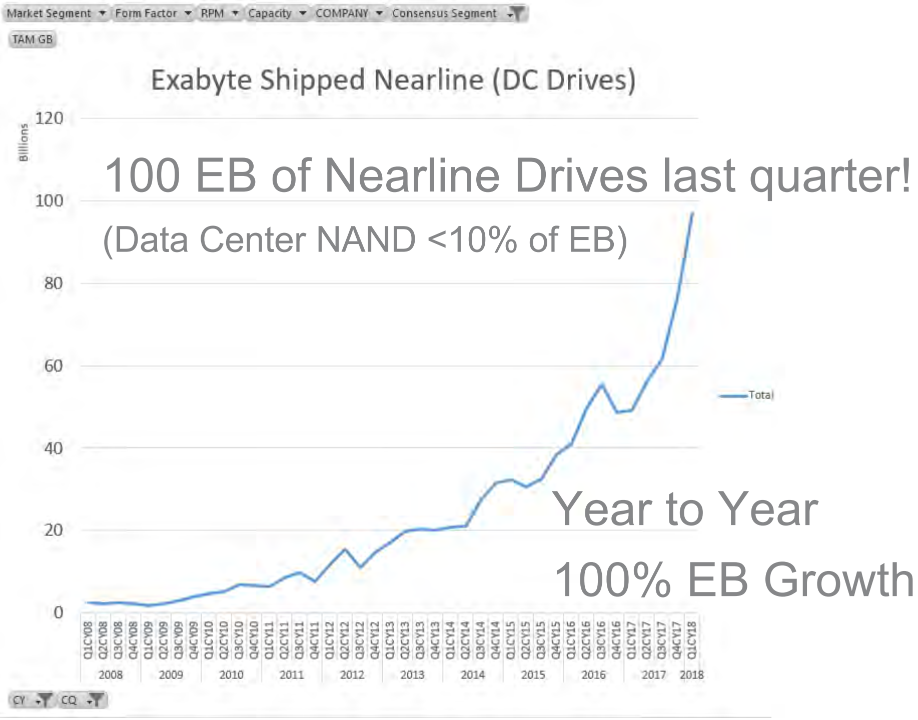
Total NAND is Growing ~40-50% Bit Growth Vs. DC Byte Demand Of 100%



C1Q18:

200EB of HDDs

(55EB all NAND)



100 EB of Nearline Drives last quarter!

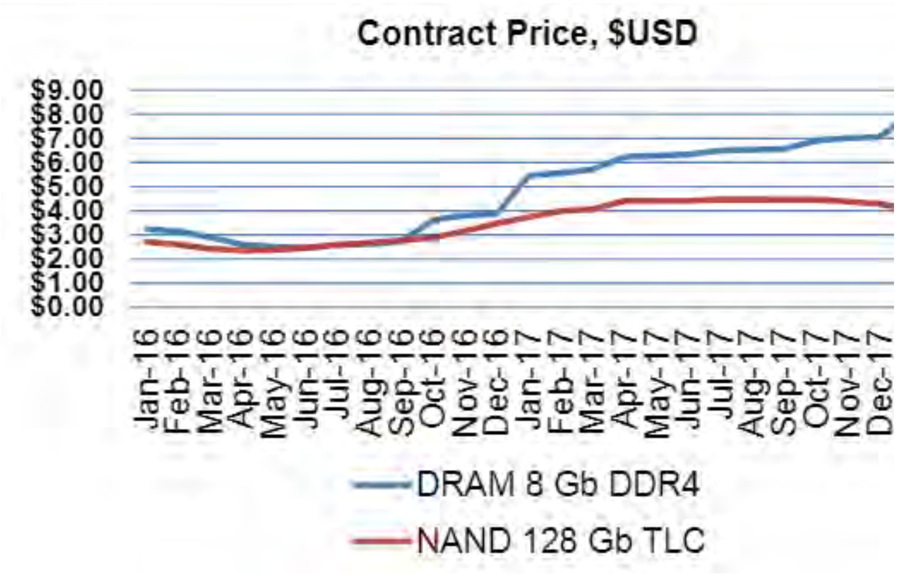
(Data Center NAND <10% of EB)

Year to Year

100% EB Growth

Now NAND is Great

However SSDs Are Small, Trendy, And Turn Out To Be Getting More Expensive

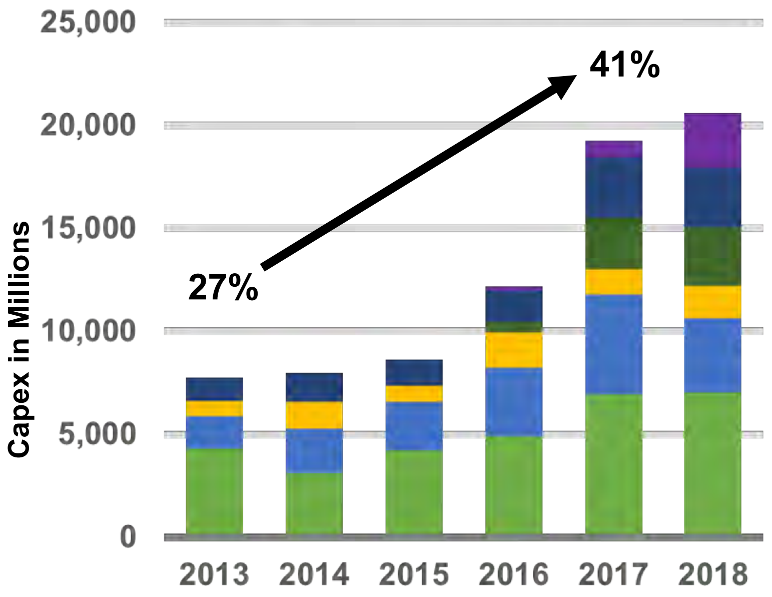


The Cost of 128Gb TLC doubled over 2 years

Capex Is Up 250% For NAND

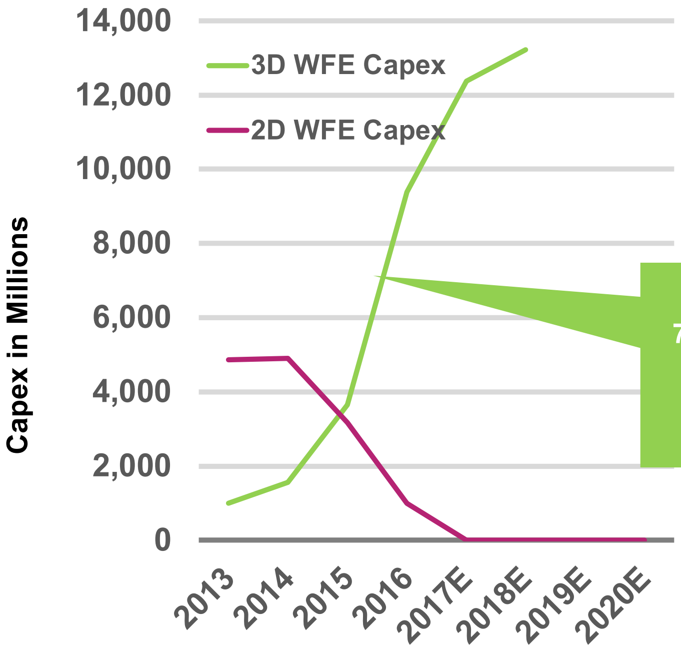
YES, THE CAN MAKE IT, BUT THE INVESTMENT HAS TURNED MASSIVE, WHICH DRIVES HIGHER PRICING

Seagate CapEx Model



$30B Industry
With $8B Capex

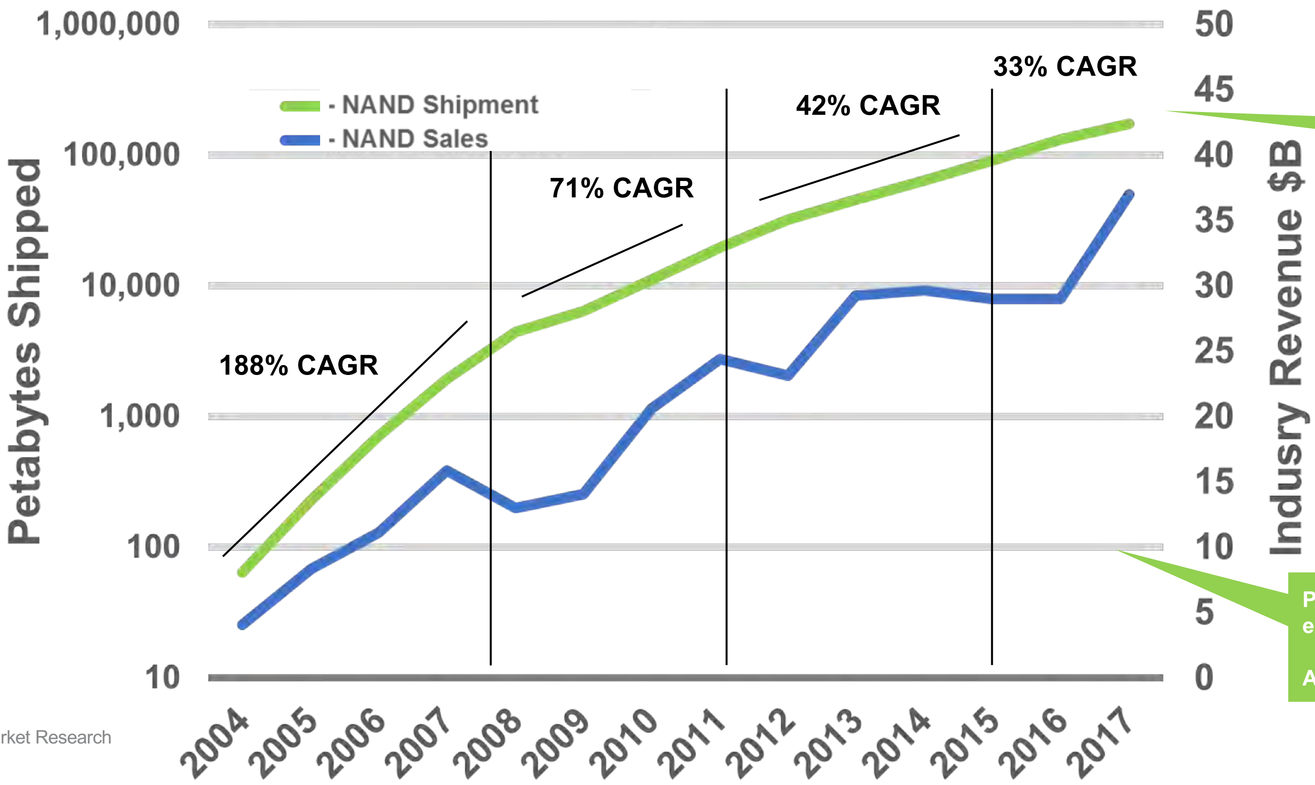
$??B Industry
With $20B Capex



70% of CapEx driven by 3D technology

Moore’s Law Is Dying, Thus NAND Bit Growth Slowing

THE EB SHIPPED CAGR IS SLOWING, BUT STEADY GROWTH[1]



170-180EB

Petabytes shipped on an exponential scale

A straight line = Same CAGR

Seagate Market Research

- 1 *Over the last 13 years*

The Debate Is Not If NAND Pricing Will Come Down

…it will, but the debate is how fast,
and my expectations is that cost declines will slow

Seagate believes in NAND, having invested $1.5B In Toshiba Memory Co, so we know it has a great future, but we need the skunk!

Convinced? HDDs are Much Easier Than Skunks

HOW DO YOU CARE FOR YOUR PET SKUNK?

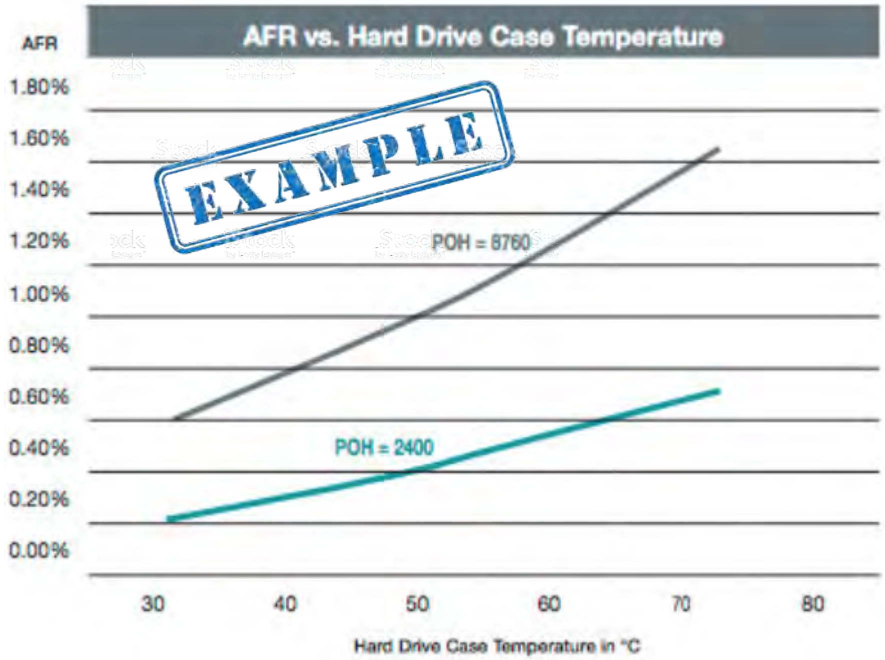
Spanking or hitting a skunk is not recommended, since it will cause him to become vengeful.

In a similar way, there is some feed and caring of your pet HDD to drive out costs

Keep Your Skunk (Or Hard Drive Cool)

LEAVE THE WINDOWS OPEN ON THE CAR

While we put more margin into a Data Center Drive, but being cool is just upside to your system reliability

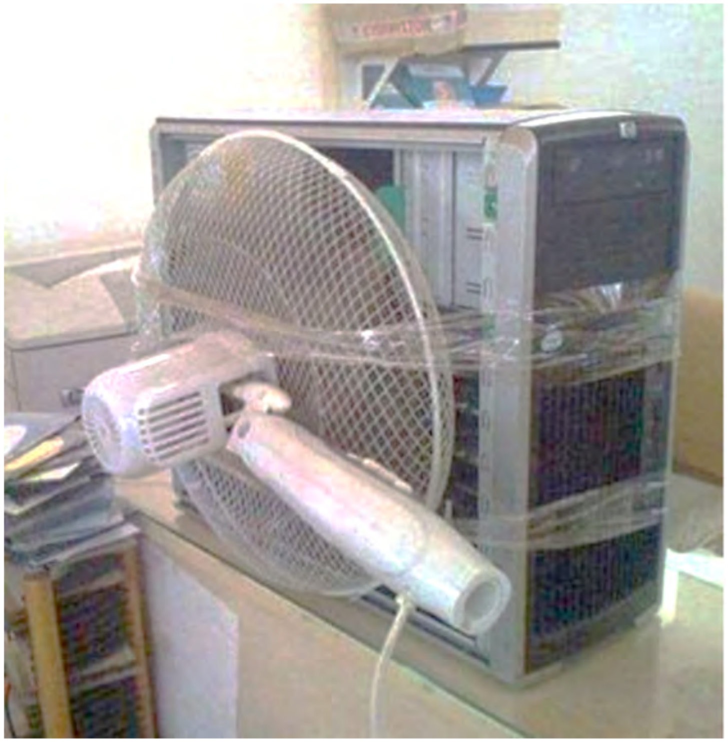


(The above was for surveillance class)

https://www.seagate.com/tech-insights/optimizing-video-surveillance-system-reliability-performance-master-ti/

A Super Reliable System?

COOLING AT ITS BEST TO HELP RELIABILITY?



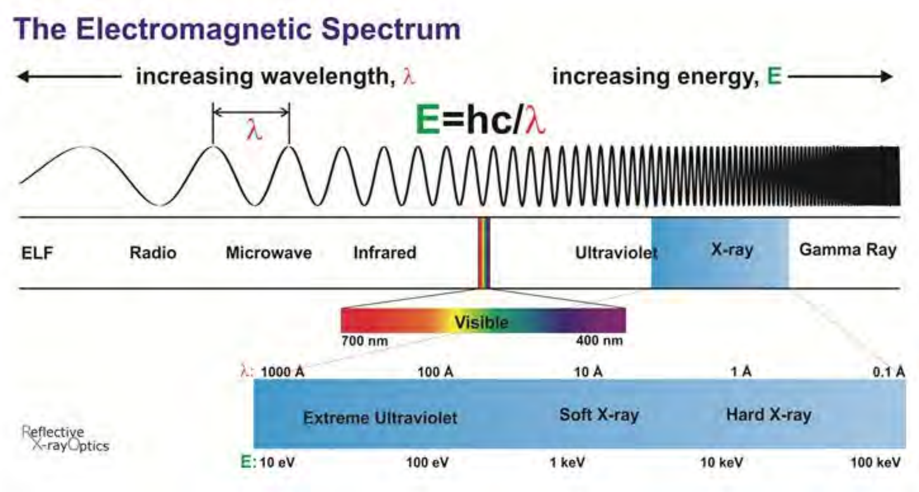
Obviously a job done by an amateur, real engineers always use duct tape

Vibe Is A Real Issue For Us

WHEN CARS TRAVEL SO CLOSE, YOU CAN'T HAVE THE ROAD A ROCKING!

We write the data in tracks that are in the range of 60nm wide

60nm

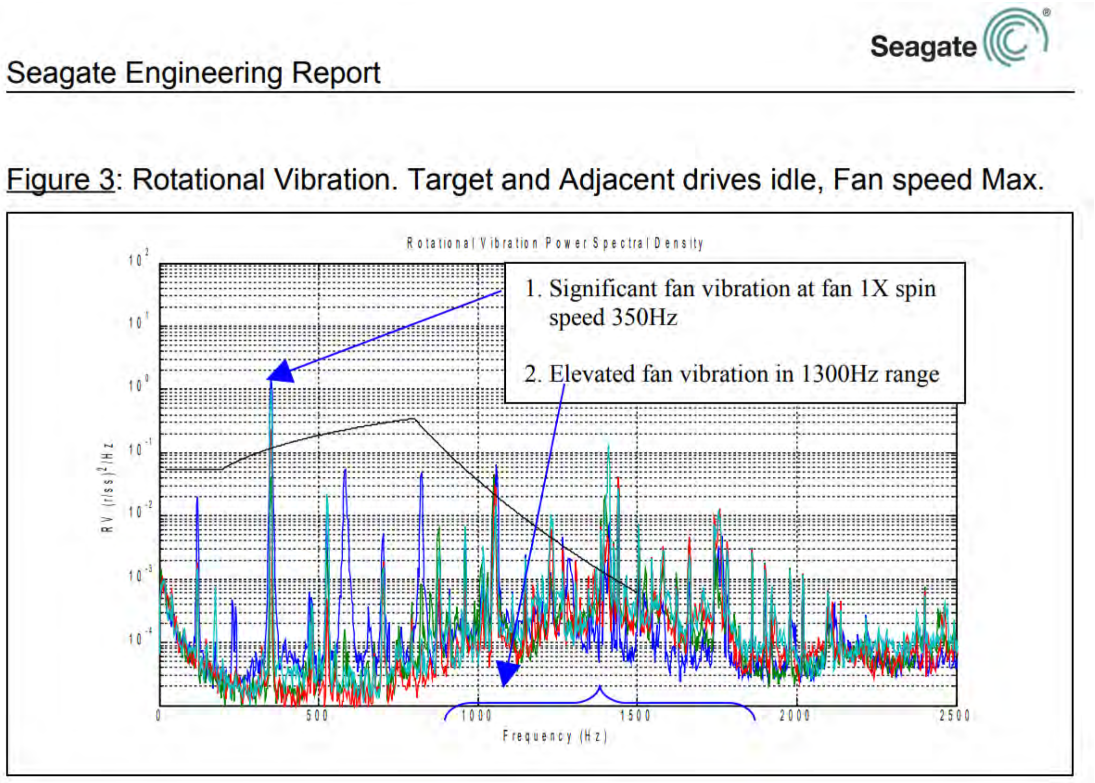


Which means that we are track width is pushing into the soft x-ray range!

(And we fly just 1nm off the disk)

Fans Are The #1 Reason For Vibe Issues

WE TEST WITH OUR MAJOR OEMS



Seagate

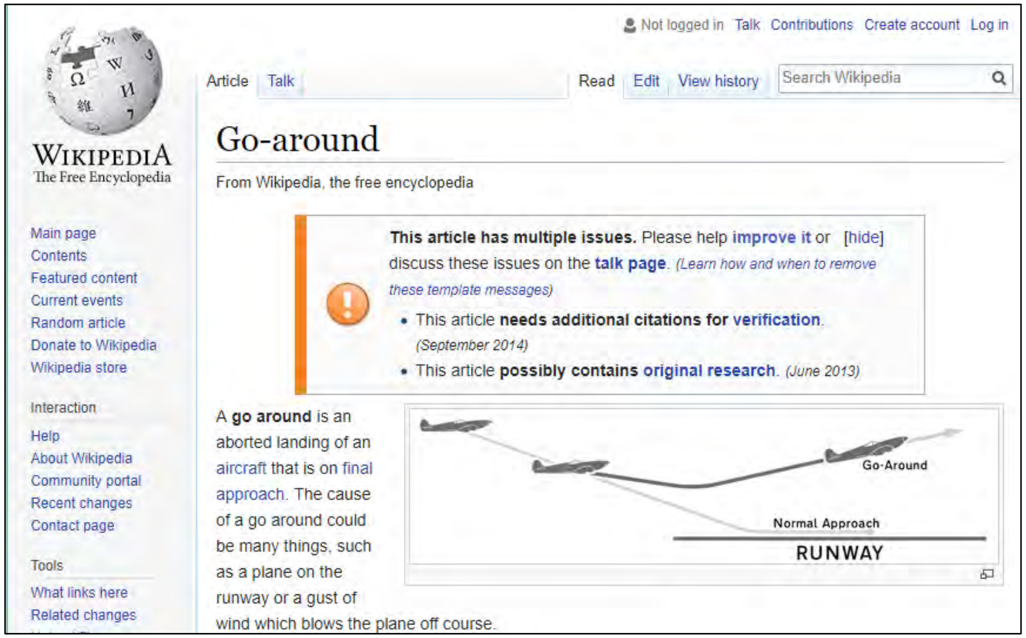
Seagate Engineering Report

Figure 3: Rotational Vibration. Target and Adjacent drives idle, Fan speed Max.

1. Significant fan vibration at fan 1X spin speed 350Hz
2. Elevated fan vibration in 1300Hz range

How Do We Handle Vibe? Do We Lose Data?

NO, WE ALWAYS PROTECT THE DATA



Go-around

From Wikipedia, the free encyclopedia

This article has multiple issues. Please help **improve it** or discuss these issues on the **talk page**. *(Learn how and when to remove these template messages)* [hide]

- This article **needs additional citations for verification**. *(September 2014)*
- This article **possibly contains original research**. *(June 2013)*

A **go around** is an aborted landing of an aircraft that is on final approach. The cause of a go around could be many things, such as a plane on the runway or a gust of wind which blows the plane off course.

We love control theory.

We have servo systems that say when it is safe to write, and if things don't look good (vibe), we handle it like a pilot and "go around"

While this preserves data, a "go around" (burned rev) takes 8ms or 8,000,000 nanoseconds

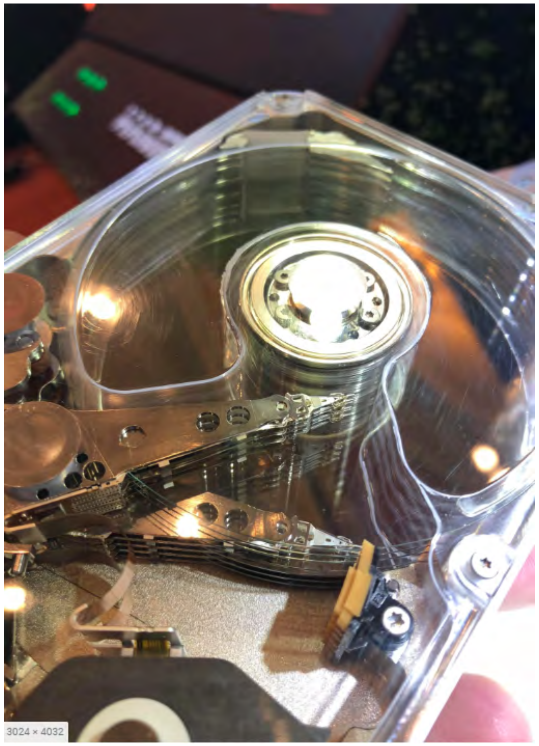
Anything Else To Keep My HDD In Fight Shape?

WHAT IS THE JUNK FOOD OF HDDS

Skunks need a wider variety of food than most pets. They tend to have a voracious appetite, making obesity a common problem.
EZ CHEEZ

The Burned Rev Problem

BACK TO BASICS



Seagate's new dual actuator drive moves data at approximately 5Gbits per second sequentially

This means that every nanosecond, we can move 5 bits, which is pretty fast

The problem is when we burn a rev, we put a 8 ms (8,000,000 nanoseconds) into the system

In the worst case, where we were transferring 5 bits at a time, we just slowed down our computation by 8 million!

What Operations Cause A Burned Rev Naturally?

THERE ARE ONLY FOUR OPERATIONS

Sequential Read: No problem, full speed ahead

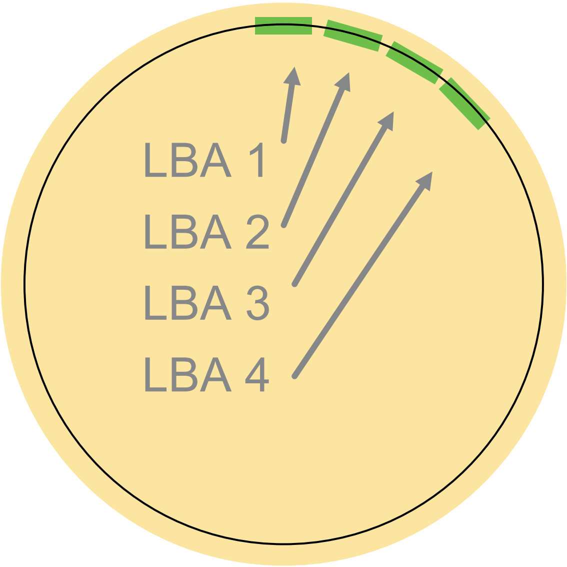
Sequential Writes: No problem, full speed ahead

Random Writes: Uh-oh, could there be a problem?

Random Reads: We found the problem!

Random Writes

WRITE CACHE SOLVES A LOT OF THE ISSUES



We write the data in circles
in units called LBAs
(Logical Block Addresses)

If you have a workload that writes

- LBA 4, then
- LBA 3,
- LBA 2
- LBA 1

Every time you write a block, the next block you want to write already past you

However, if you write LBA 4-1 into a buffer (or cache), the drive can write the data as 1-4 and not burn any revs!

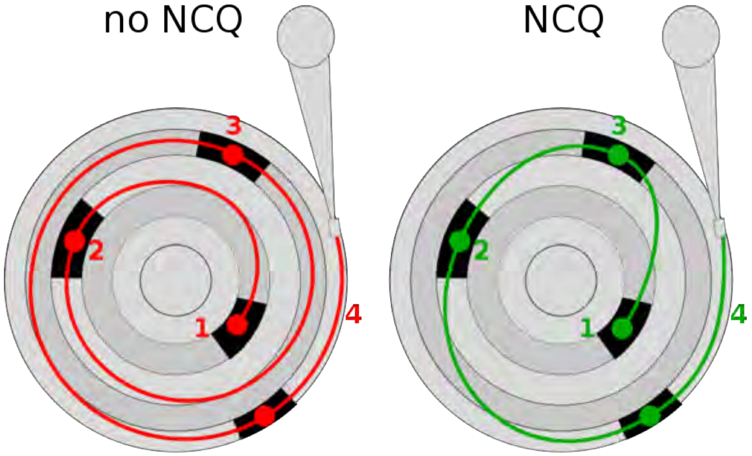
But I Thought Write Caching Is Bad?

HOPEFULLY, YOU'VE UPDATED YOUR COURSEWORK

- Classically write caching was thought of as bad because after data was committed to the disk, you destroyed the data in host memory
- If the disk said, "I got the data" but it was only in write cache waiting to be written, and came back later with a problem, you had probably destroyed the original data, and life was miserable
- Engineers at the time came up with a method of accepting the data, but keeping a flag open (or tag) that sophisticated systems would use
- However, as a lot of workloads have moved from ACID to BASE databases, with workloads like Hadoop or FlumeJava, we don't care if we have perfect data integrity
  - Thank you Dr. Eric Brewer
- Then in the modern datacenter, we erasure code everything, and we have multiple copies
- Moral: Run SQL on flash, and use hard drives and write cache for everything else

So What Is The Problem?

RANDOM READS, THE FINAL FRONTIER



- Really, only random reads are the item holding things back
- Ironically, we already know the solution, a version of pipeline/branching that we've worked on for years call "Tag Command Queueing" or its SATA brother NCQ
- Wikipedia has an article of course at https://en.wikipedia.org/wiki/Native_Command_Queuing

The Old Dogs Had A Few “Knew” Tricks

WHEN ALL YOU HAD WAS HDDS, YOU WORKED WHAT YOU HAD

The old SCSI spec had “unlimited” queues of up to 128 commands, but we had one sophisticated customer that complained they needed more to wring out more performance

Virtually all of the largest most sophisticated customers are not using this age old technique

However, all of them say that they want to

A Slide On Why They Have Not Used Queuing

WHY ARE WE REDISCOVERING THE ANCIENT ART OF QUEUING

- The explosive growth of data centers was founded by a new breed of engineers
  - Coincidental with the move to new architectures and non-SQL databases
- They threw out a lot of the old
  - Example: SATA was not considered a “real” interface, but now is the predominate interface
- Nobody stopped to implement queuing, and at the same time, NAND looked like it was going to take over the world with the price drops
  - I had many of the largest customers say “well these NAND guys look like they could catch you”
- As it is becoming obvious that HDDs will be a large part of the data center, we’ve found out that QoS (Quality of Service) tolerance was not built into the system
  - And queing HDDs can have tail-latencies for IO that their system does not handle well
- However, we are addressing this with a flurry of work that allows full employment for anybody working on storage stacks (see next page)

Back-Up: Some Of The Work On QoS

(OR THE FULL EMPLOYMENT ACT FOR STORAGE STACK SOFTWARE ENGINEERING)

- SATA-IO
  - ICC (Isochronous Command Completion) is described in SATA revision 3.3 for use with READ FPDMA QUEUED and WRITE FPDMA QUEUED commands

- T13: ACS-4 limiting command completion time
  - SCT Error Recovery Control command
  - Streaming feature set
  - Rebuild Assist feature set

- T10: SPC-4 limiting command completion time
  - Command Duration Limit A mode page
  - Command Duration Limit B mode page

- NVMe Working Group (has rotating media commands, but no HDD person has announce product)
  - IO Determinism
  - Directives (Streams)

- OpenCompute
  - Fail Fast (SAS and SATA)

The only things we have to fear is the fear of coding itself

Summary: Adopt A Hard Drive (And Maybe A Skunk)

WAY TO LOWER COSTS

- HDDs use in the data center is accelerating, and the trends looks like they are there to stay
- While costs of rotating media is great, there a couple of things that you need to do to preserve costs and performance
  - Keep your HDD cool
  - Keep vibe away from your hard drive not because of reliability, but because of performance
- Speaking of performance, hard drives can be very, very fast
  - Simply use write caching with the appropriate workload and erasure coding
  - However, their Achilles' heel is their random read performance
- However, we have ways of unlocking the random read performance of hard drives
  - The secrets of the ancient coders, coming to a data center near you

- Questions?