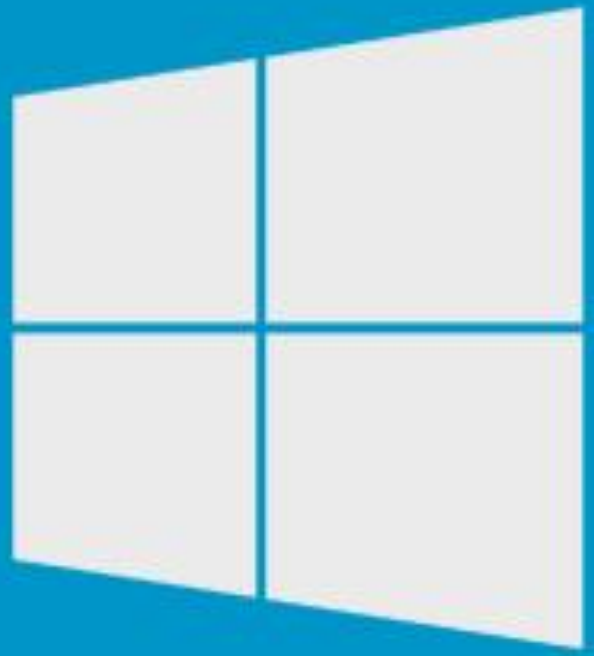
How to store a zettabyte on a budget



Microsoft Azure

Aaron Ogus
Partner Development Manager
Microsoft Azure Storage

Agenda

- My background.
- Why store your data in the cloud?
- How can you store efficiently?
- Let’s talk about storage media.
- How much does it cost to store data today and in the future on
  - HDD / SSD / Optical / Tape

Disclaimer

- I meet with lots of hardware suppliers, they all have confidential roadmaps that I cannot disclose. Everything here is publically released or I have permission to show.

- I am not disclosing how Microsoft stores cloud data, I am outlining alternative ways to build storage.

My background

- Started in Windows Azure Storage in late 2007
- Started with 4 x 750GB HDDs in a 1 U Server
- Optimizing design for the last 8 years
- Meet with Suppliers regularly:
  - OEM, ODM, HDD, SSD, CPU, Networks, Disk Controllers, NIC etc...
- COGs reduction 98%
  - 90% from technology improvement, 80% from storage design improvement.
  - .1*.2 = 0.02

Why Cloud Storage?

- Phones, tablets and laptops get lost or dropped.
- Cloud enabled devices need less local storage.
- In the cloud all devices can share data.
- Cloud storage is becoming cheap enough, all data can go there.
- Cloud scale compute needed to process large data sets

- What about Security?

Which is a cheaper way to store data?

4 Drive Server vs. 90 Drive Server

There is actually debate over this.

- Windows Search and Google, both thought that the 4 disk server was better.
- Why? – one type of server in fleet, can be repurposed.
- I was buying the server anyway, the slots for disk are “free”.

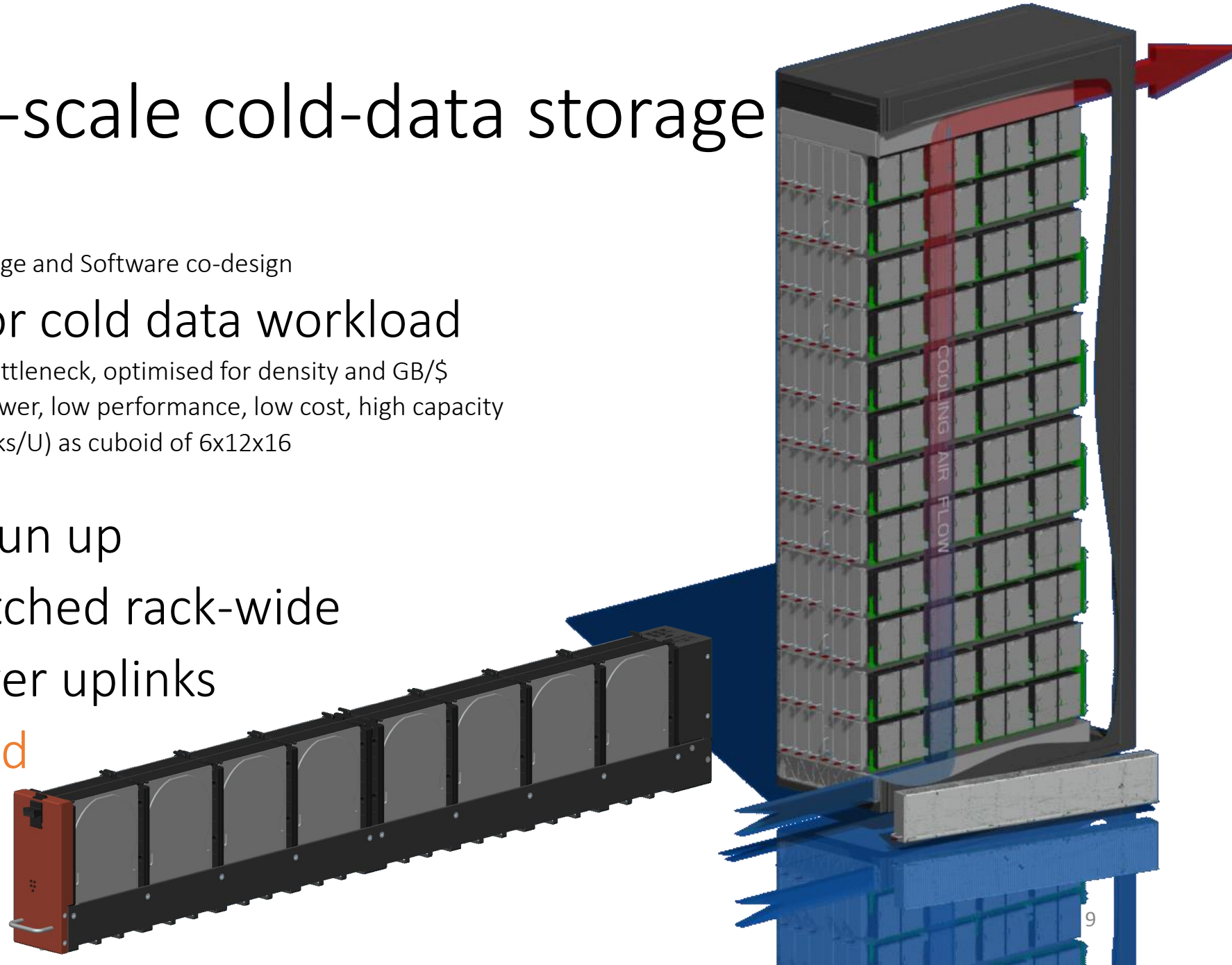
- Some very smart people believe in 4 disk servers for bulk storage.
- It is possible if you are buying enough servers anyway that the 4 disk solution is cheaper. However you have tied your ratio of compute and storage together.

Is there a cheaper way to store data on HDD?

- More drives per server is generally cheaper, lower slot tax.
- How far can you take it?

Pelican: rack-scale cold-data storage

- Converged design:
  - Power, Cooling, Mechanical, Storage and Software co-design
- Right-provisioned for cold data workload
  - All resources close to balanced bottleneck, optimised for density and GB/$
  - New “archive” SATA disks : low power, low performance, low cost, high capacity
  - 1,152 disks in 52U (density 22 disks/U) as cuboid of 6x12x16
  - About 3.5 kW per rack
- At most 8% disks spun up
- 2 Servers, PCIe stretched rack-wide
- No TOR switch; server uplinks
- Constraints managed in software



Selling what you buy

- When you buy datacenters, generators, networks and servers but sell ads, you might not know what things cost.
- When you buy datacenters, generators, networks and servers and you rent them, you will know what things cost or be out of business.

- Like inflation, Moore's Law hides' inefficiency.
  - I set a goal of 20% YoY improvement in my design, that's incredible!

- Anyone running a service and not renting out infrastructure might be very inefficient, and not know, or know and not care.

Why would you need to store a zettabyte

- Photos, by 2020 there will be 5 billion smartphones
- 1MB per picture, and 2 pictures per day = 20 PB / day = 7.3EB / year
- IoT / IoE?
- How about this?
- Full Resolution Video Clips and Movies
  - Easily need a zettabyte

- Enterprise Data Sets
- Some predict 25 zettabytes / year by 2020
- Some say over 400 ZB /year now.

Profile of Stored Consumer Data

- A very large percentage of data is accessed only in the first few days
- The longer data is stored the colder it becomes
- Expectations for recovery can be met with lower quality renderings
- High quality must be available in a reasonable time

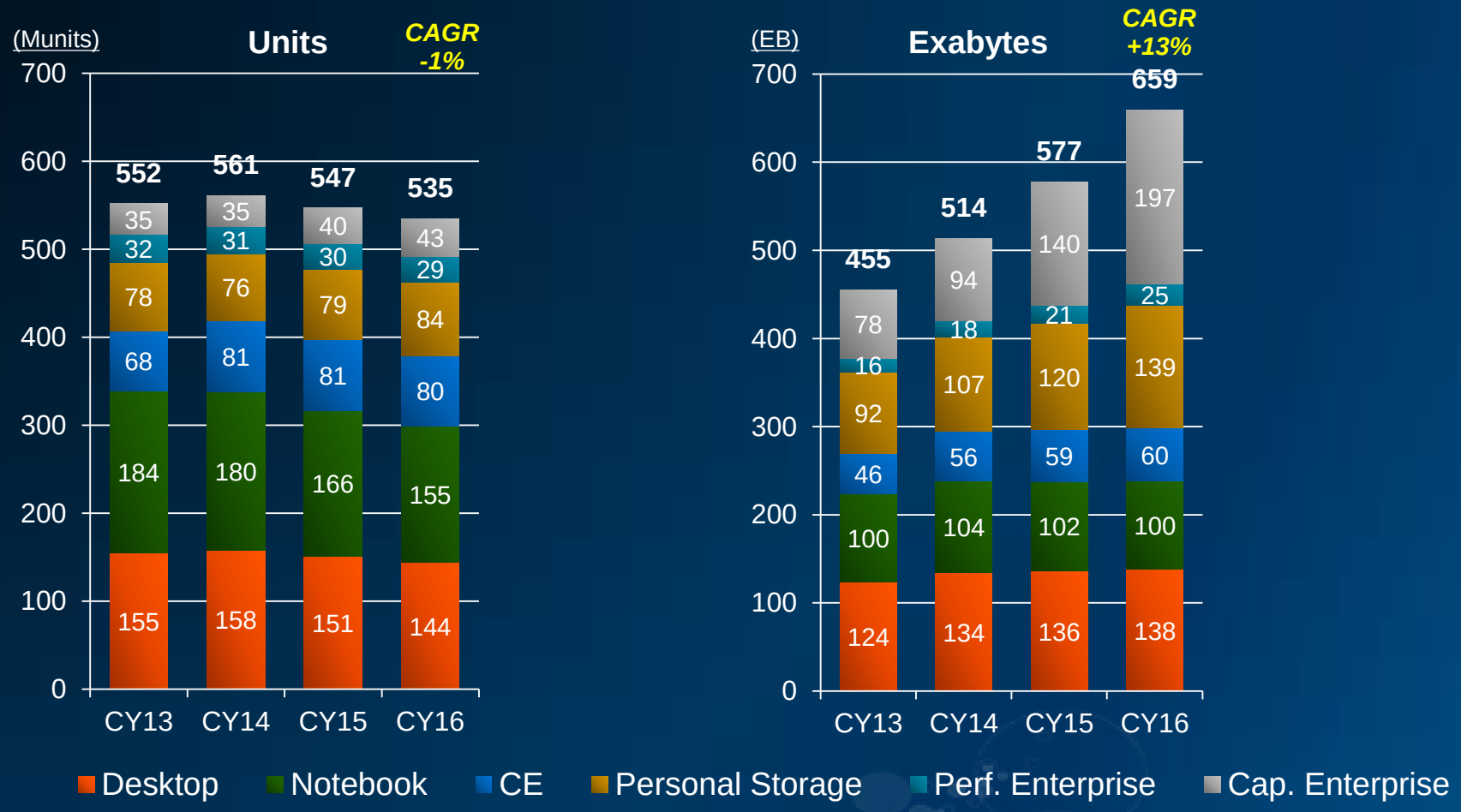
- Profile is changing, may become hotter
- Reprocessing may be needed from time to time
  - annotation, face recognition, site recognition etc...

Is there a zettabyte of media produced every year?

- Thanks to HGST for this data...

HDD Market Outlook

Exabyte growth driven by Capacity Enterprise HDD's



Trajectory

- Trajectory for capacity HDD doesn’t indicate an annual zettabyte of DC class drives by 2020. About 400EB-700EB/year by 2020.

- Demand changes everything though. If they come, we will build it.

Where is technology going?
Your own “dirty” crystal ball to the future

Year	1991	2008	Improvement	Annual
RAM (bytes)	4,000,000	16,000,000,000	4000	1.63
CPU (GHz)	33,000,000	16,000,000,000	485	1.44
Disk (bytes)	60,000,000	750,000,000,000	12,500	1.74
NIC (bps)	10,000,000	1,000,000,000	100	1.31

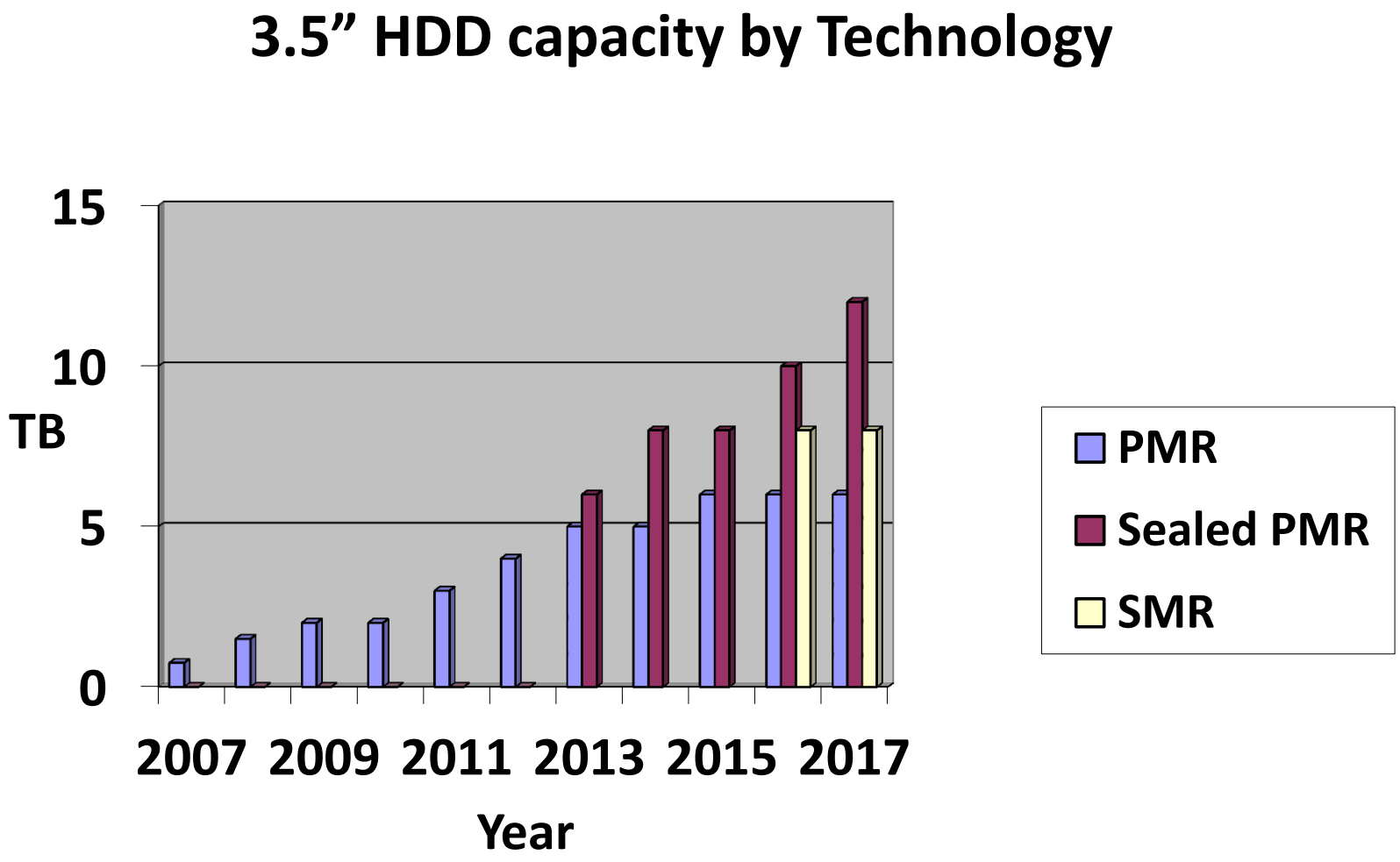
Year	2008	2015	Improvement	Annual
RAM (bytes)	16,000,000,000	128,000,000,000	8	1.35
CPU (GHz)	16,000,000,000	64,000,000,000	4	1.22
Disk (bytes)	750,000,000,000	8,000,000,000,000	11	1.40
NIC (bps)	1,000,000,000	40,000,000,000	40	1.69

Will exponential improvement continue?

- HDD hitting limits of PMR, top of S-curve, need a leap.
  - 1,2,3,4,6,8,10,12
  - 100%, 50%, 33%, 50%, 33%, 25%, 20%, ???
  - SMR: one time 15-30% bump
  - HAMR: hope for the future?

- Today, about $0.04/GB, by 2020 if we get HAMR $0.01-$0.02/GB

History and Future of HDD, 2014 and later estimated.



Best guesses from 2014... (actual devices + projections, not a roadmap)

Will exponential improvement continue?

- SSD
  - On 1y nm node, now going 3D, and backward in geometry.
  - Still 5-10x the cost of HDD
  - Expect 30% CAGR
- Today $0.50 / GB
- 2020 $0.10 / GB

Will Exponential Improvement Continue?

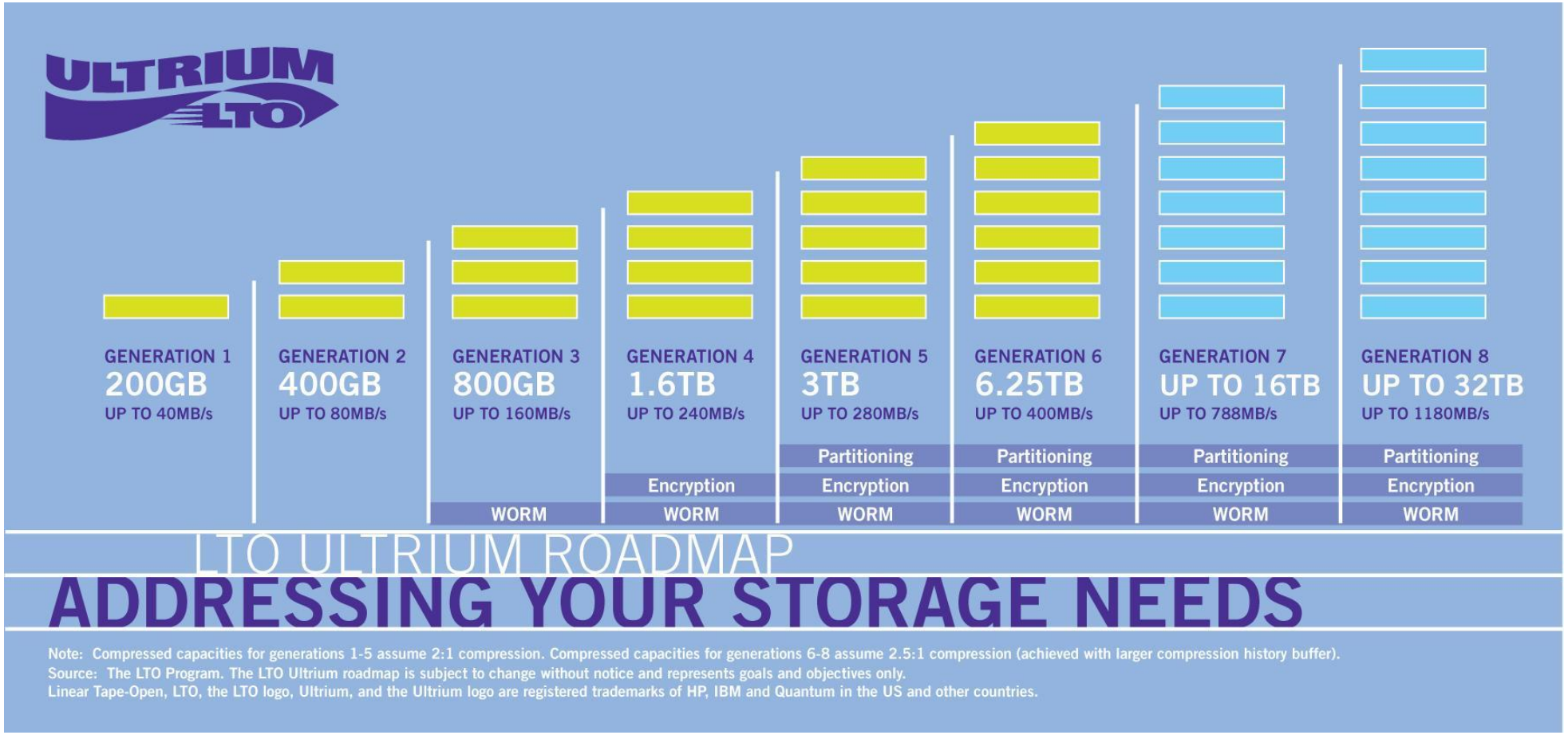
- Optical
  - 100GB -> 300GB, but long wait for 1TB
  - Based on DVD technology, losing consumer base
  - Reinvestment?
- Tape
  - 10 TB shipping, and 220TB demonstrated
  - High data rates (320MB/s, going to over 1GB/s)
  - Very long access times (90 seconds)
- 30 Optical Disks == 1 Tape capacity
- In 2020’s 100 Optical Disks == 1 Tape capacity

Magnetic tape (r)evolution

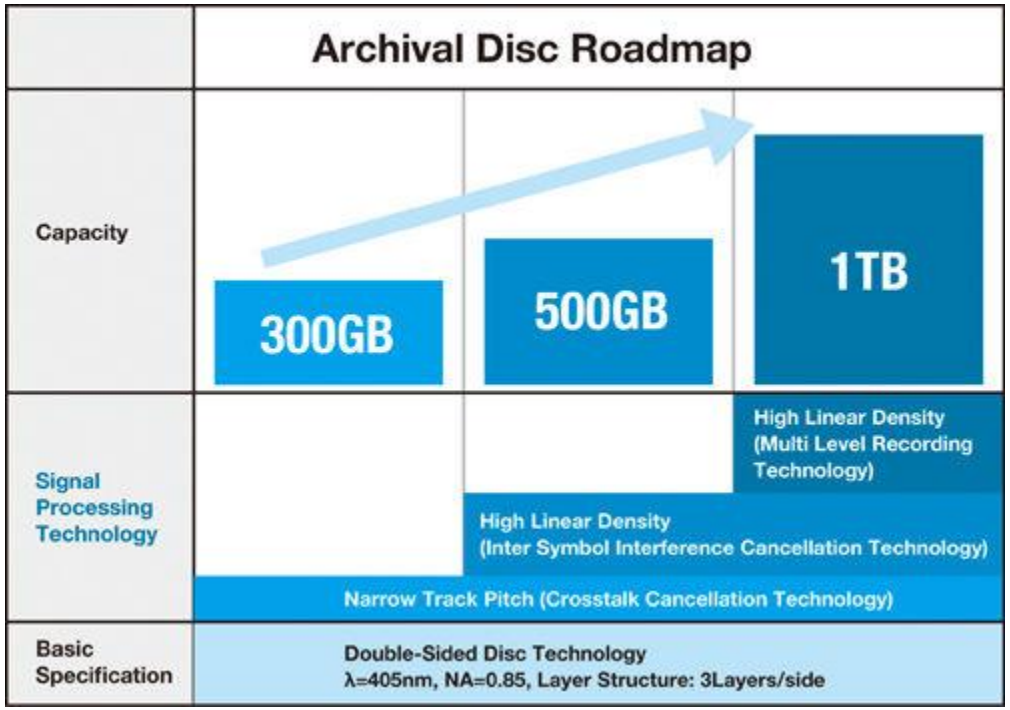
Product / Year:	IBM 726 /1952	LTO6 / 2012	TS1150 /2014	Demo 2015
Capacity:	2.3 MBytes	2.5 TBytes	10 TBytes	220 TBytes
Areal Density:	1400 bit/in^2	2.06 Gbit/in^2	6.7 Gbit/in^2	123 Gbit/in^2
Linear Density:	100 bit/in	385 kbit/in	510 kbit/in	680 kbit/in
Track Density:	14 tracks/in	5.35 ktracks/in	13.2 ktracks/in	181 ktracks/in

19.8 cm

LTO Tape roadmap



Archive Optical Disk Roadmap
1TB around 2020



What makes up the cost of storage?

- Data center space and power
  - $6-12M / MW ≅ $9/W
- Connection cost and power
  - Media Slot Tax: As high as 50% for hot storage, as low as 20% for cold storage
- Network infrastructure cost
  - Below 10% for hot storage, nominal for cooling storage
- Time to fill
  - 50% full means 2x the cost
- Media lifetime
  - Twice the life is half the cost <unless new media is cheaper than maintenance>

Separating the media from the system

- Both Tape and Optical pull the media out of the reader.
- 1 expensive part, and service any amount of media.
- Optical libraries fix the maximum ratio of drives / media
- Tape libraries can be more flexible in drive / media ratio
- New tape drives can store more data on older media
- Tape libraries have tighter environmental constraints
- Tape libraries require fiber channel infrastructure

Run through the media, and storage class

- Today, and 5 years from now. Future is estimated using current CAGR.
- Using “retail” pricing
- How much datacenter space will it take.
- How much power will it take.
- How much will it cost all up to store a ZB?
  - SSD
  - Hot Disk
  - Optical
  - Tape

SSD storage - smoking hot

- Assumptions:
- Current (2015)
  - High performance server ($4000), 500W
  - 32TB per server
  - 3 replica
  - $0.50 / GB
- Future (2020)
  - 128TB per server
  - $0.10 / GB

ZB on flash

	2015	2020
Cost / GB	3.45	0.71
Cost / ZB	$3.4 Trillion	$714 Billion
Cost / ZB / year	$1.15 Trillion	$238 Billion

Problem: less than 200 EB of flash is produced every year

ZB on HDD Storage

- Assumptions
- Current(2015)
  - Erasure codes within a DC, replication level of 1.3
  - 90 disks / server, 8TB disks
- Future (2020)
  - HAMR is made to work
  - 24 TB disks
  - Slightly more expensive per disk (freakin lasers)
  - Higher power per disk

ZB on HDD Storage

	2015	2020
Cost / GB	0.125	0.05
Cost / ZB	$125 Billion	$50 Billion
Cost / ZB / year	$41 Billion	$17 Billion

Optical WORM storage

- Assumptions
  - $100,000 per 3PB rack (300GB media)
  - Power 1kW/rack
  - 2 Servers per rack
  - 7 year lifetime
  - Zero re-write
- Future
  - $150,000 per 10PB rack (1TB media)

ZB on Optical

	2015	2020
Cost / GB	0.19	0.028
Cost / ZB	$192 Billion	$29 Billion
Cost / ZB / year	$27 Billion	$4 Billion

Difficulty: you must know you workload, rewrite rate of 1 time every 3.5 years doubles cost.

ZB on Tape

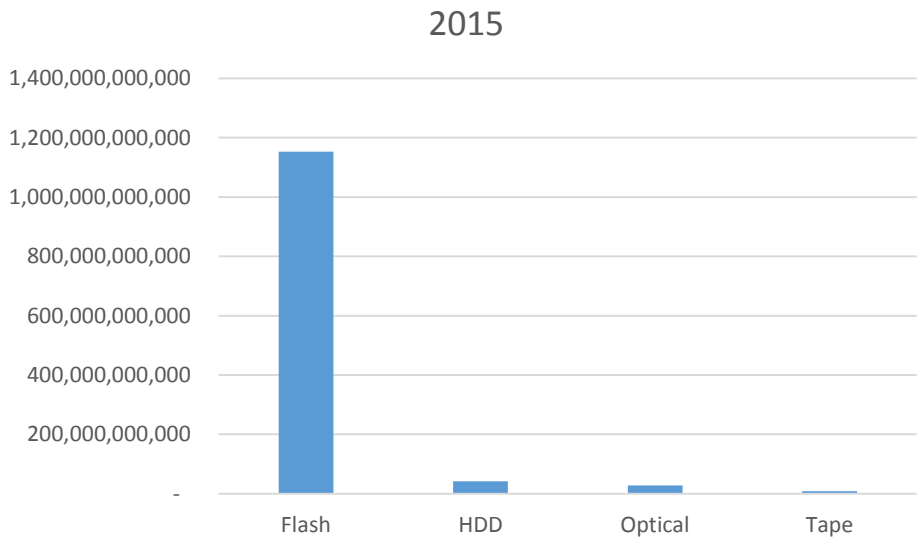
Assumptions:

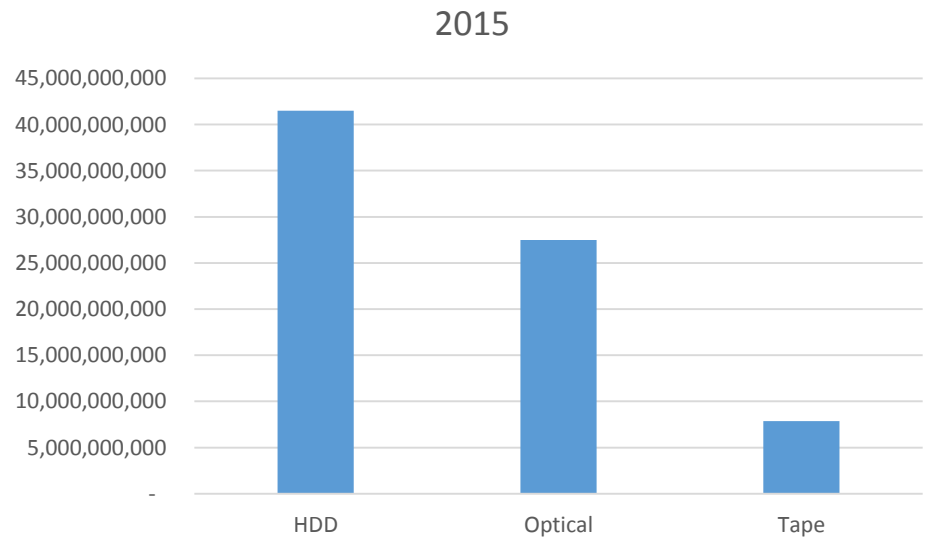
- Current (2015)
  - 10 TB Enterprise Tape
  - Gateway Servers to Connect DC to Fiber Channel
- Future (2020)
  - 40 TB Tape (following 30 CAGR)
  - Tape Drives on Ethernet

ZB on Tape

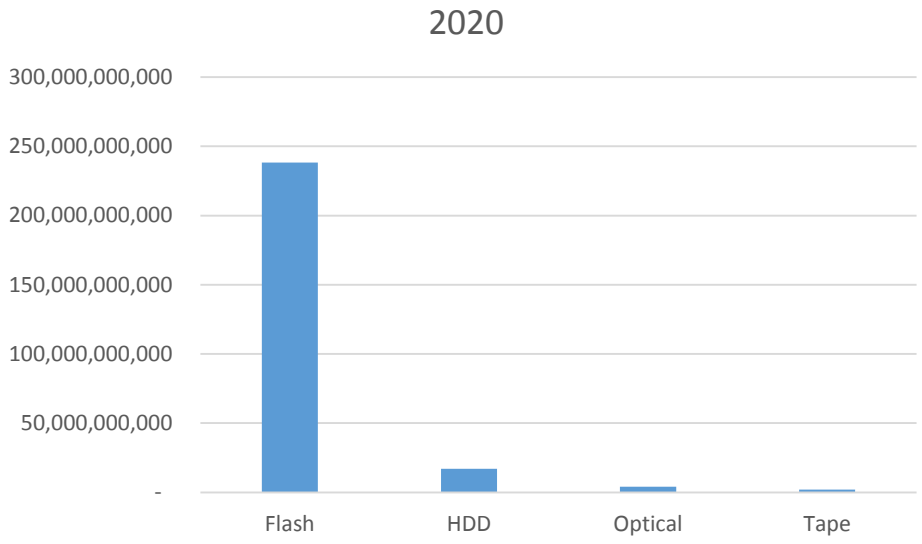
	2015	2020
Cost / GB	0.06	0.013
Cost / ZB	$55 Billion	$13 Billion
Cost / ZB / year	$8 Billion	$1.9 Billion

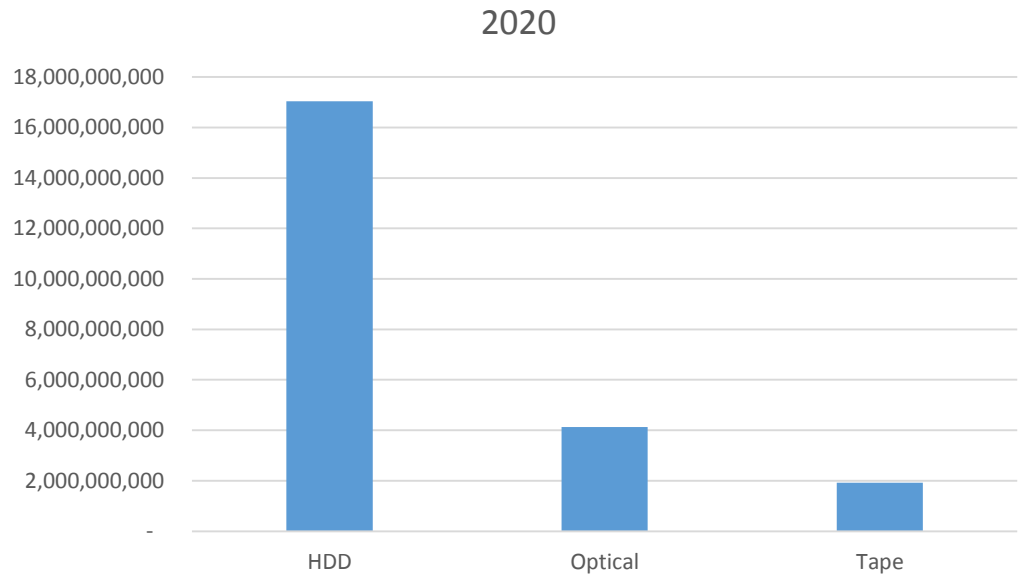
Cost to store a ZB for a year in 2015





Cost to store a ZB for a year in 2015





Summary

- Storing a ZB will be financially feasible in 2020
- Optimizing Storage Tiers to match the workload is critical to cost control.
- Media manufacturers must ramp up significantly.
- Some cold Archive Tier must be part of any strategy to cost effectively store a ZB.
- Investments in ways to move data between tiers smoothly will be a critical area of cloud development over the next decade.