

HEPiX bit-preservation WG update – Spring 2014

Dmitry Ozerov/DESY


Germán Cancio/CERN

HEPiX Spring 2014, Annecy



Agenda

- Bit-preservation WG one-slide summary
- Ongoing work
 - Recommendations for bit-preservation best practices
 - Bit preservation cost outlook
 - Cost model for 10,20,30 years archive

- **Mandate summary** (see w3.hepix.org/bit-preservation)
 - Collecting and sharing knowledge on bit preservation across HEP (and beyond)
 - Provide technical advise to 
 - Recommendations for sustainable archival storage in HEP
- **Survey on Large HEP archive sites carried out and presented at last HEPiX**
 - 19 sites; areas such as archive lifetime, reliability, access, verification, migration
 - HEP Archiving has become a reality by fact rather than by design
 - Overall positive but lack of SLA's, metrics, best practices, and long-term costing impact

Two work areas:

1. Preparing a set of **best-practice recommendations** for bit-level preservation within HEP
 - ~10 recommendations
 - Concentrate more on “what” rather than “how” to do
 - Will be circulated to WG participants and surveyed sites summer time
 - Feedback will be most appreciated
2. Defining a simple and customisable model for helping establishing the **long-term cost** of bit-level preservation
 - Useful for site planning/outlook
 - Input for DPHEP – significant fraction of overall Data Preservation cost!
 - The rest of this presentation

What is the approximate cost of a data archive over 10, 20 and 30 years?

- Generic archive (as on any HEP site)
- Start from scratch in terms of HW/media, with some initial data to be added
- Consider hardware, media, maintenance and electrical power costs
- 3 base scenarios
 - a) 10 PB initially, growing @ 50PB / year
 - b) 10 PB initially, growing @ 50PB +15% / year
 - c) 100 PB initially, no further data (“stable large archive preservation”)

Assumptions / limitations (1)

- Archive is tape based with a disk cache front-end
 - Single copy of data on tape
- Archived data is not compressible / deduplicable, tapes working at 100% capacity
- Access patterns:
 - write w/o high deletion
 - read of ~30% of archive/year, high latency for non-cached data
- Model based on 3 year cycles (10 cycles = 30 years)
 - Corresponding to HW generations and warranty lifetime
 - After each cycle, all disk cache servers and tape drives are replaced by new generation equipment
- Tape media is kept for 2 cycles
 - Enterprise-class equipment (not LTO)
 - All media repacked to higher density on second cycle
- Disk cache capacity for 10% of the archive
 - No replication (JBOD or RAID0)
 - Disk cache used for data influx, reading, repacking
- Duty cycle of 30% for both disk and tape servers
 - Relevant for power consumption

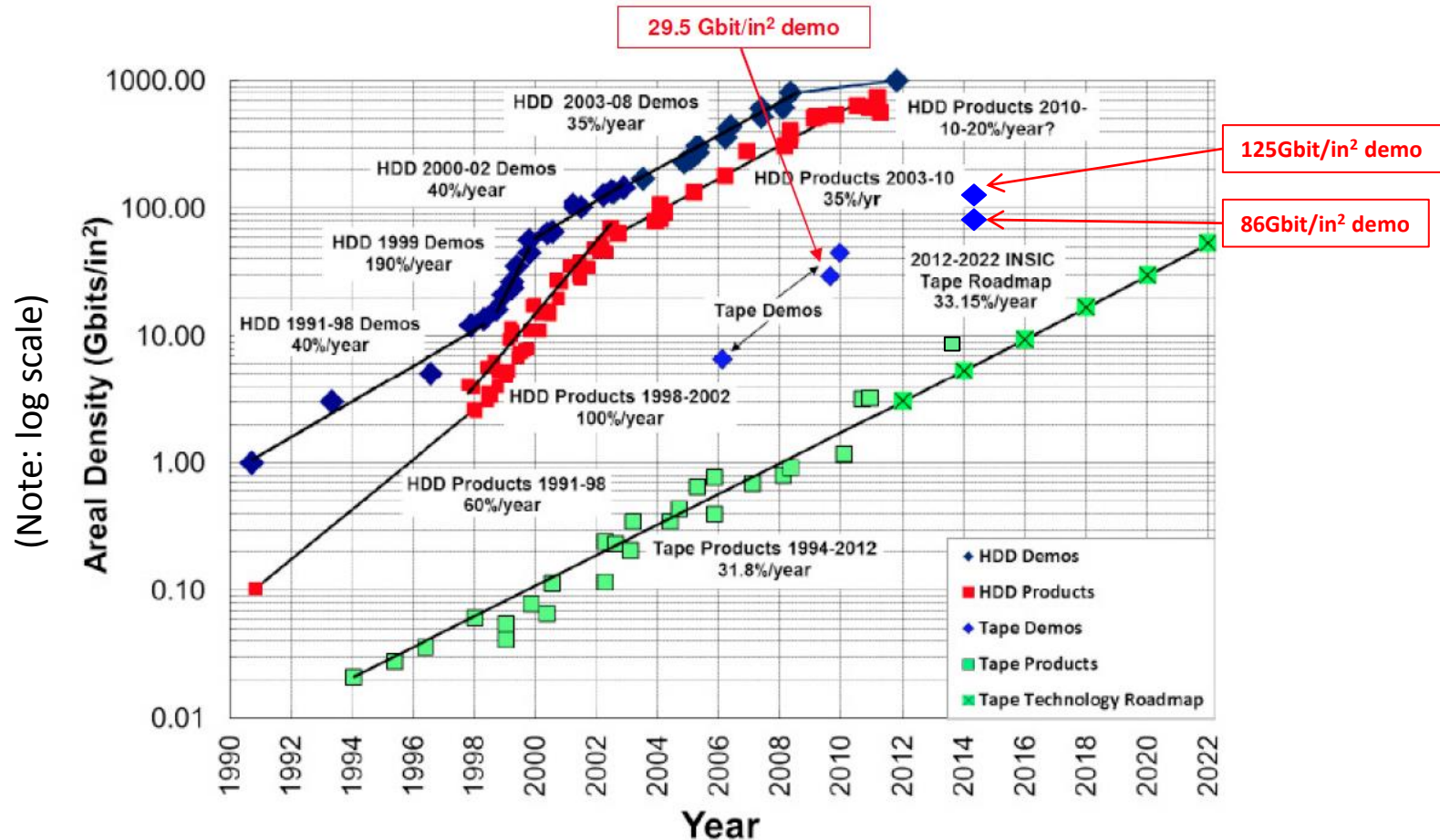
- Technology evolution forecast risky for 30 years
 - Model assumes no architecture paradigm shift (tape/disk)
 - Forecasts may hold true for 5 years, but longer-term extrapolation is risky
 - Will cloud storage affect storage capacity/pricing evolution?
- But, assuming similar storage capacity growth rates as over the last 30 years, archive cost becomes almost insignificant after 20 years
- Example: TODAY, CERN's 100PB archive requires 11.7K new-generation tapes (@ 8.5TB each)
- With 11.7K tapes, what were we able to store in the past?
 - 10 years ago (2004): tape @ 200GB -> 2.4 PB -> 277 of today's tapes
 - 20 years ago (1994): tape @ 20 GB -> 235 TB -> 28 " " "
 - 30 years ago (1984): tape @ 200MB -> 2.35 TB -> less than one of today's tapes!!!

Assumptions / limitations (3)

- Pricing mostly based on USD prices for a public US contracting alliance
 - Including educational discount
- Manpower costs not included
 - Estimations: 1FTE (engineer) + 0.5FTE (technician) for disk; 2 FTE (engineer) + 0.5 FTE (technician) for tape
- Software development / licensing costs not included
- General DC operations / floor space cost not included
- No assumptions on HW/media resale
 - Outdated / redundant HW/media is just decommissioned
- No inflation / interest rates; payments done upfront

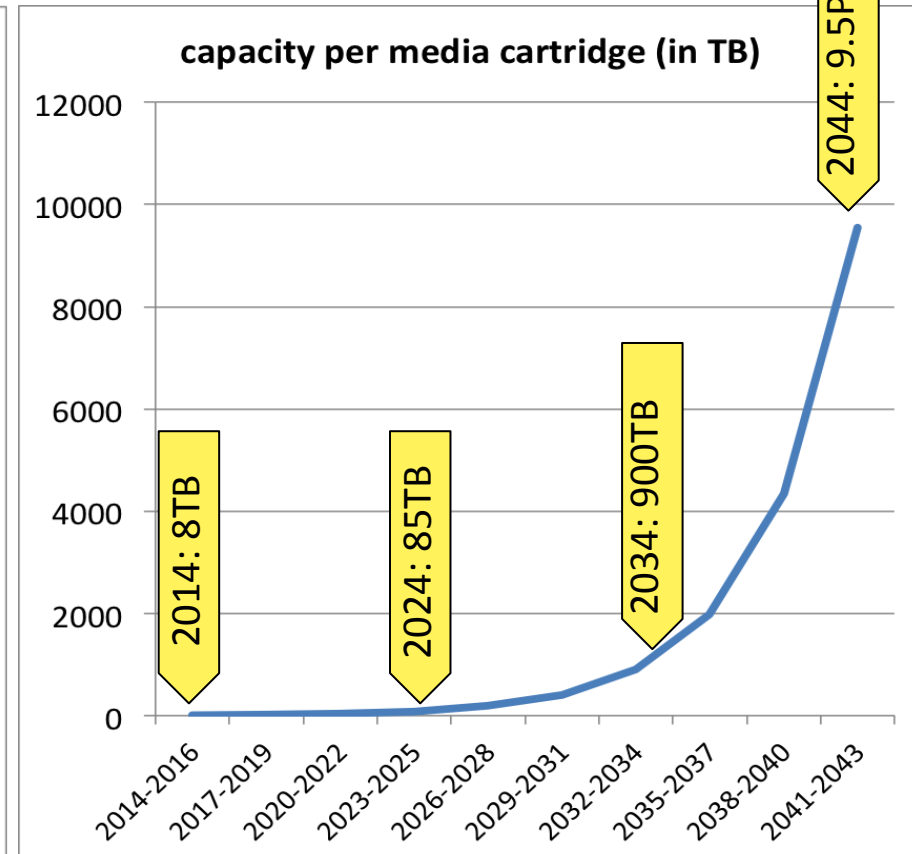
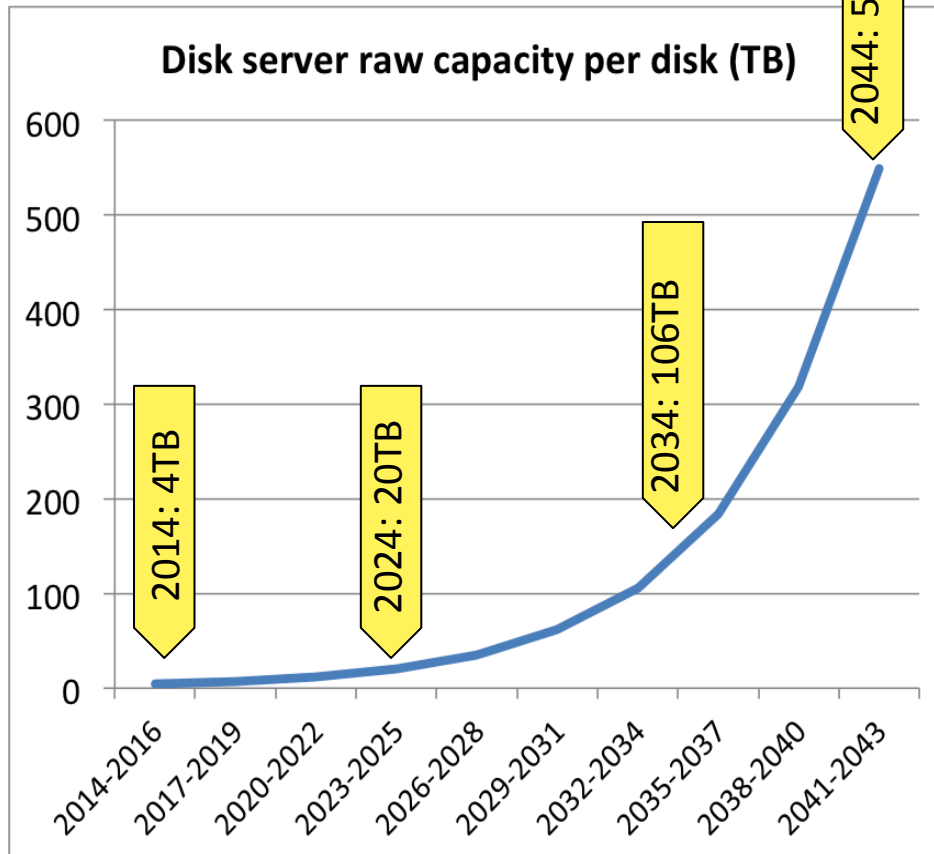
Technology evolution

- Assuming
 - +20% yearly disk capacity per constant \$
 - +30% yearly tape capacity per constant \$



Technology evolution

- Assuming
 - +20% yearly disk capacity per constant \$
 - +30% yearly tape capacity per constant \$ (+20%/yr I/O increase)



XLS spreadsheet

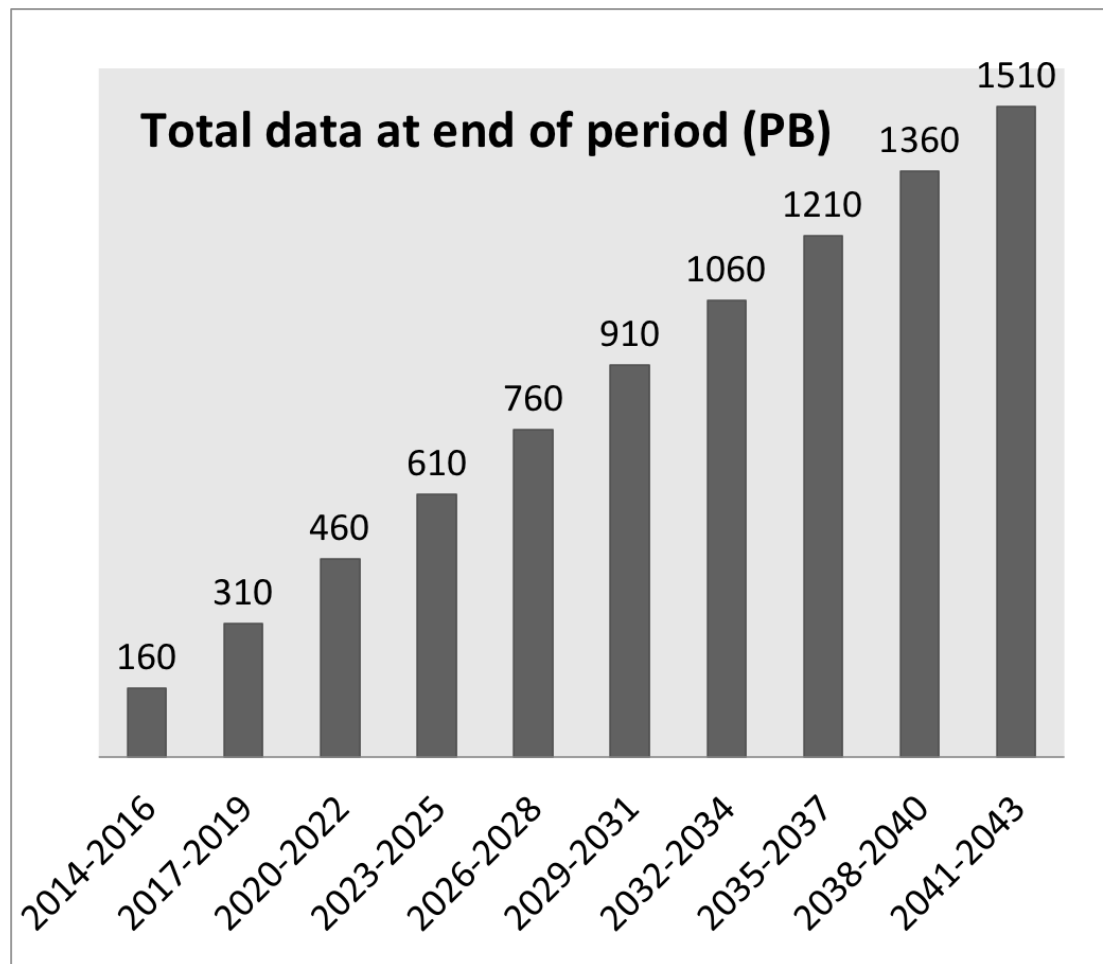
- Available on WG twiki page ([link](#))
- 1 tab for global parameters
- 1 tab for each scenario
 - Including graphs (scrolling down)
- Green fields == input data
- Please try it out and feed back 😊

Global parameters

cartridge capacity growth % per year	30%	33% according to INSIC
cartridge capacity growth factor (3 years)	2.20	
disk capacity growth % per year	20%	20% approx according to CERN IT CTO
disk capacity growth factor (3 years)	1.73	
slots per tape library	12000	12K - average btw Oracle, IBM, Spectralogic
cartridge / tape drive ratio (archiving access + repack + verification overhead)	500	500 at CERN
Overhead factor for decommissioning libraries	1.2	We don't decommission libraries immediately after removing cartridges, but keep a certain overhead
Disk cache total capacity (% of data at end period)	10%	10% sufficient for archiving + repacking functionality
Power consumption(W) tape library	550.00	Oracle SL8500 excluding drives, cf http://www.oracle.com/us/products/servers-storage/sun-power-calculators/calc/sl8500--power-calculator-161830.html
Power consumption(W) tape drive at 30% load	52.20	Oracle T10000D
Power consumption(W) disk server at 30% load	380.00	estimate
Power cost per kWh	\$0.14	(cf Wikipedia - Germany prices)
Power cost per W / 3 years	\$3.68	

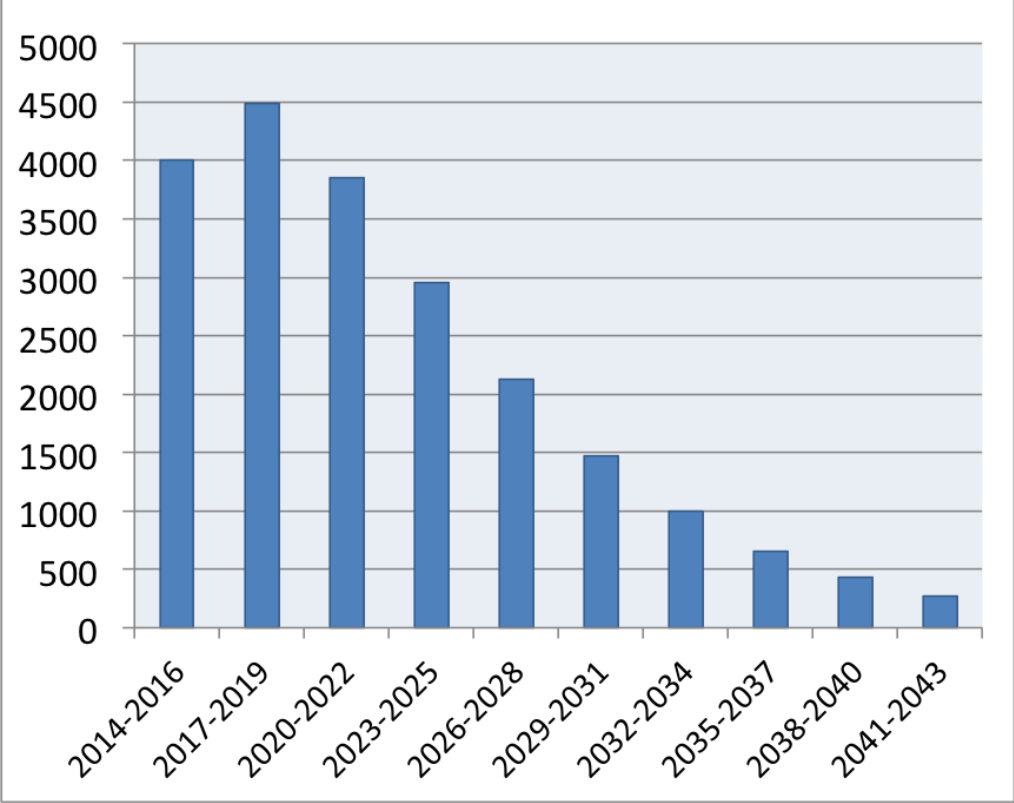
Case A) steady growth

Start with 10PB, then +50PB/year (150PB / 3y period)

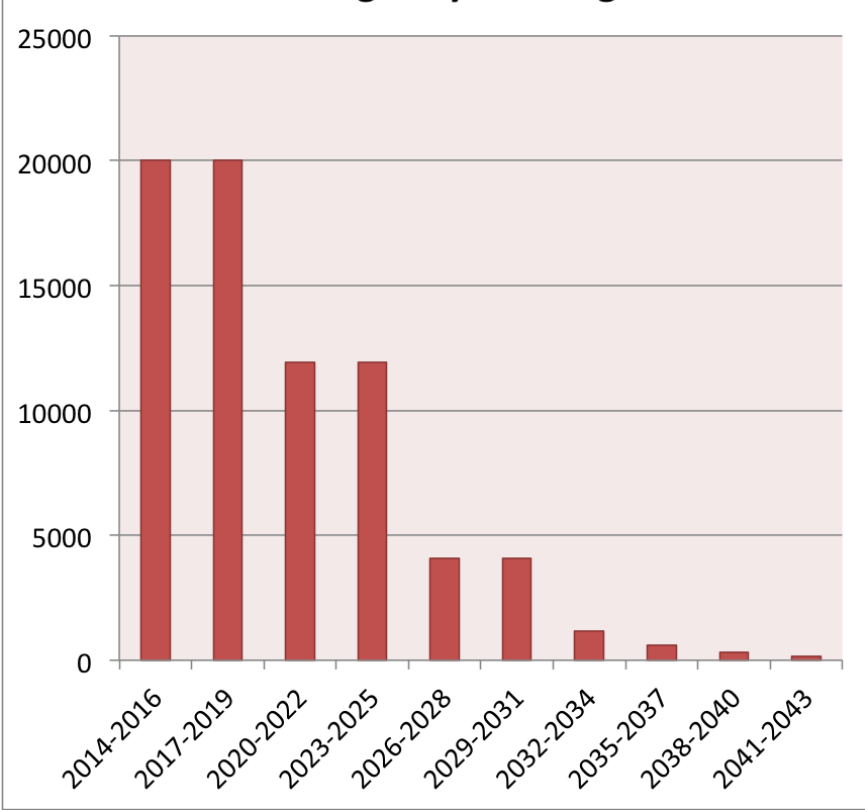


Case A) steady growth

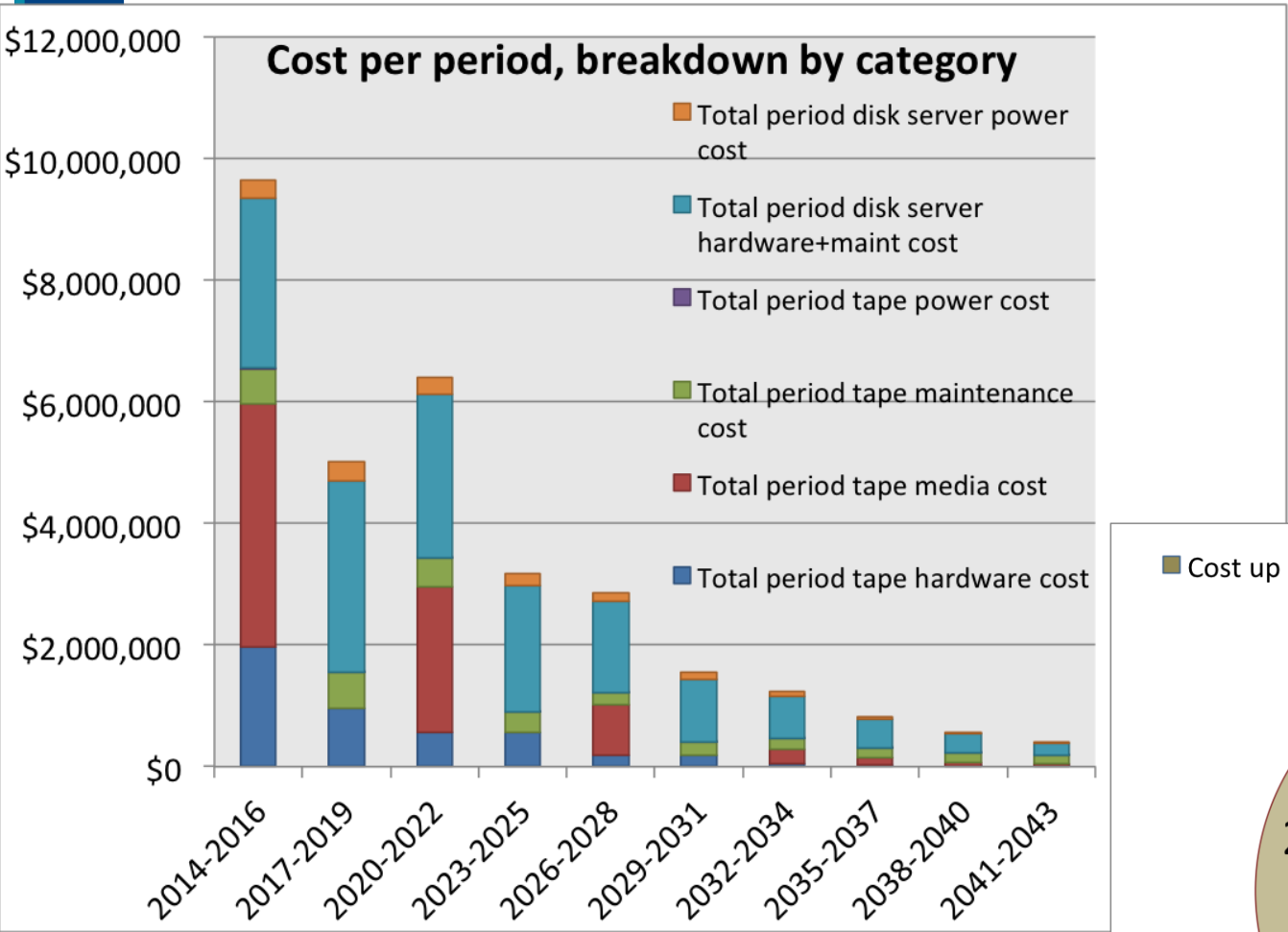
Total number of disks



Total cartridges by end of generation

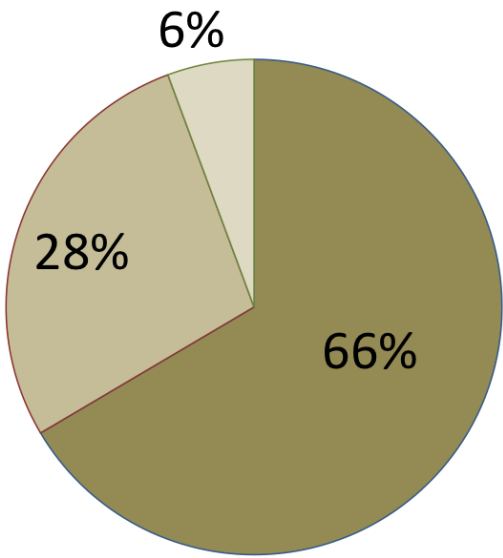


Case A) steady growth

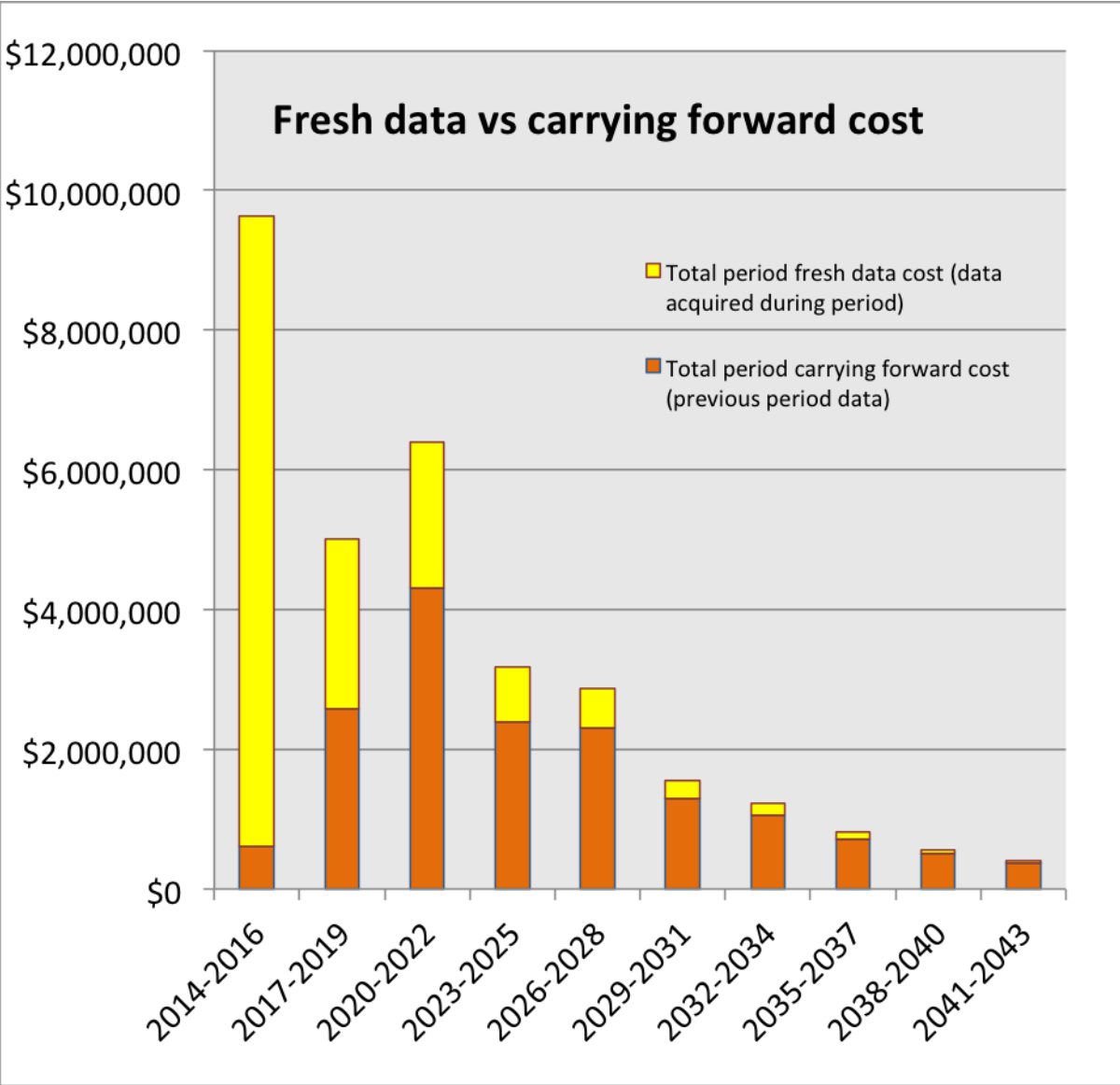


Total cost: ~31.6M\$
(~1M\$ / year)

Cost up to yr 9 Cost up to yr 21 Cost up to yr 30

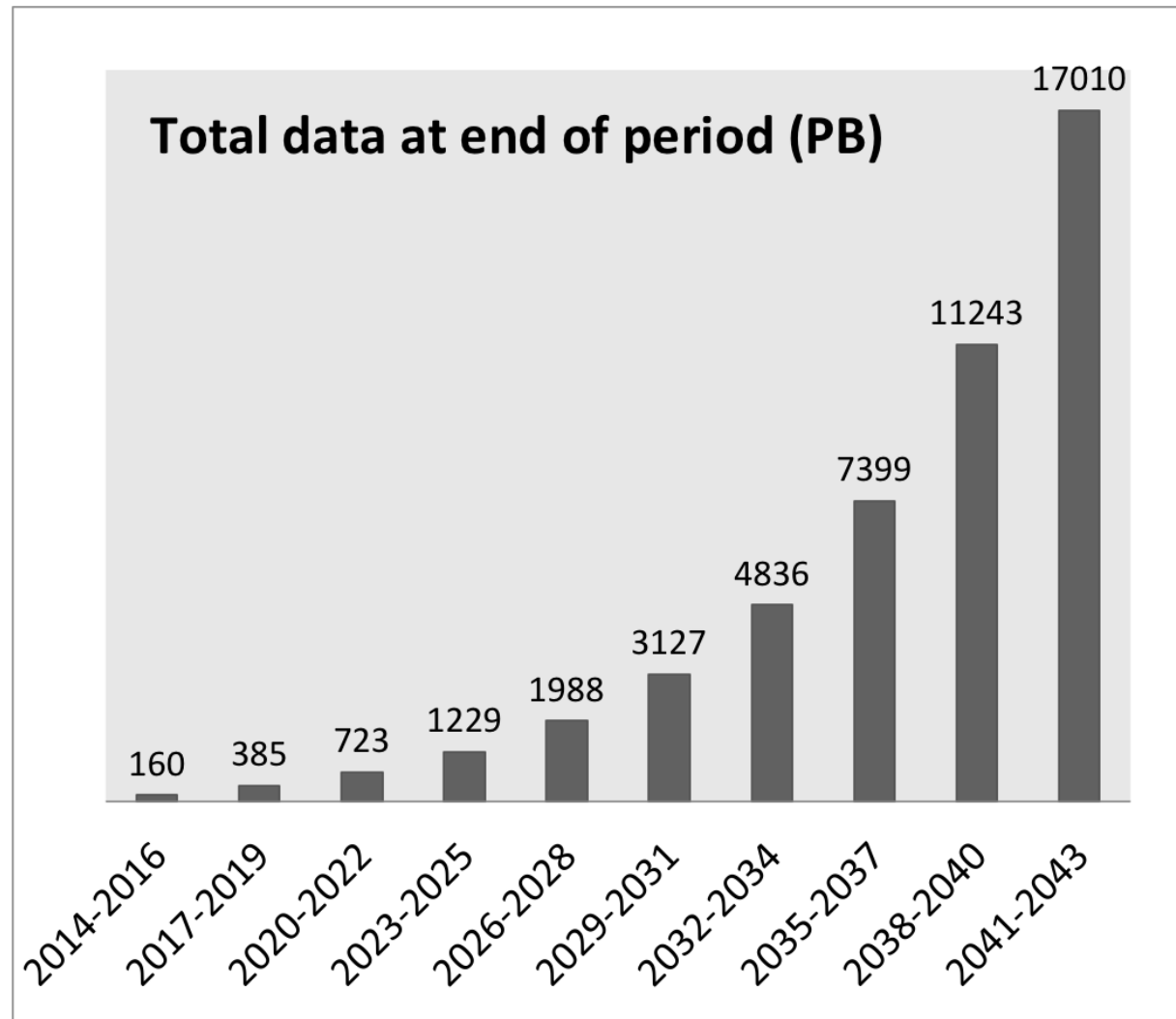


Case A) steady growth



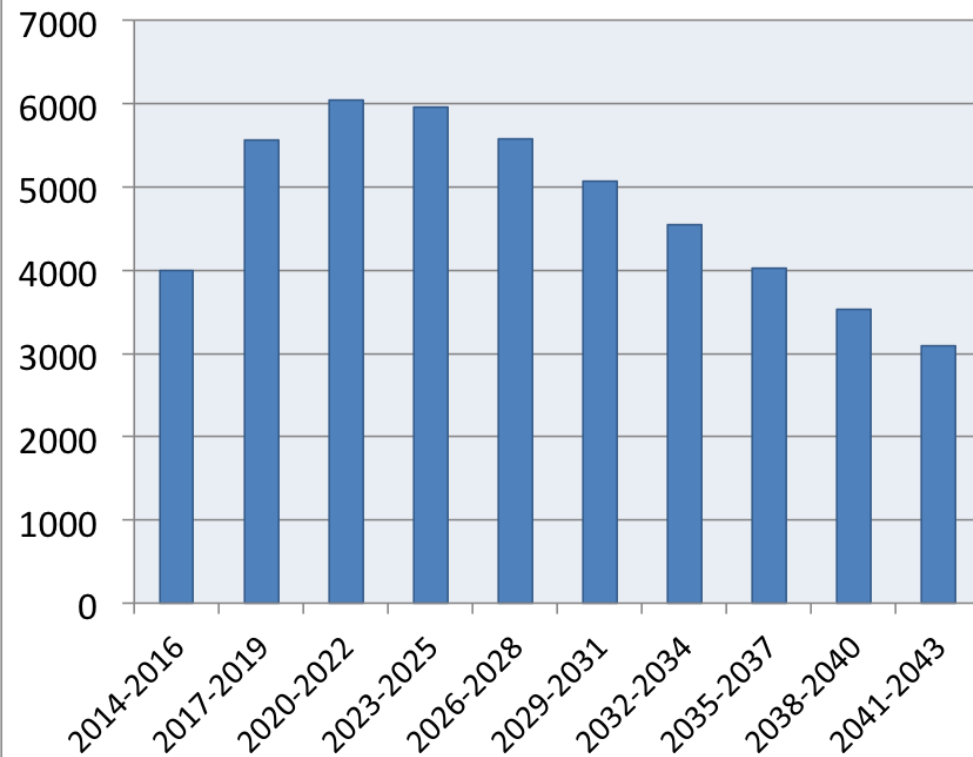
Case B) increasing archive growth

Start with 10PB, then +50PB/year, then +50% every 3y (or +15% / year)

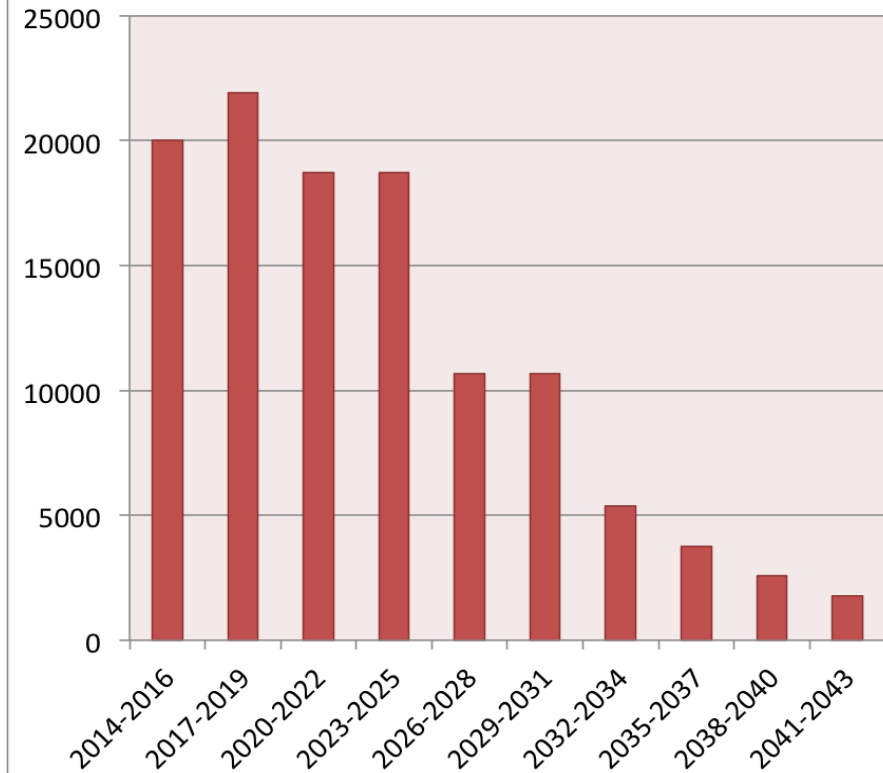


Case B) increasing archive growth

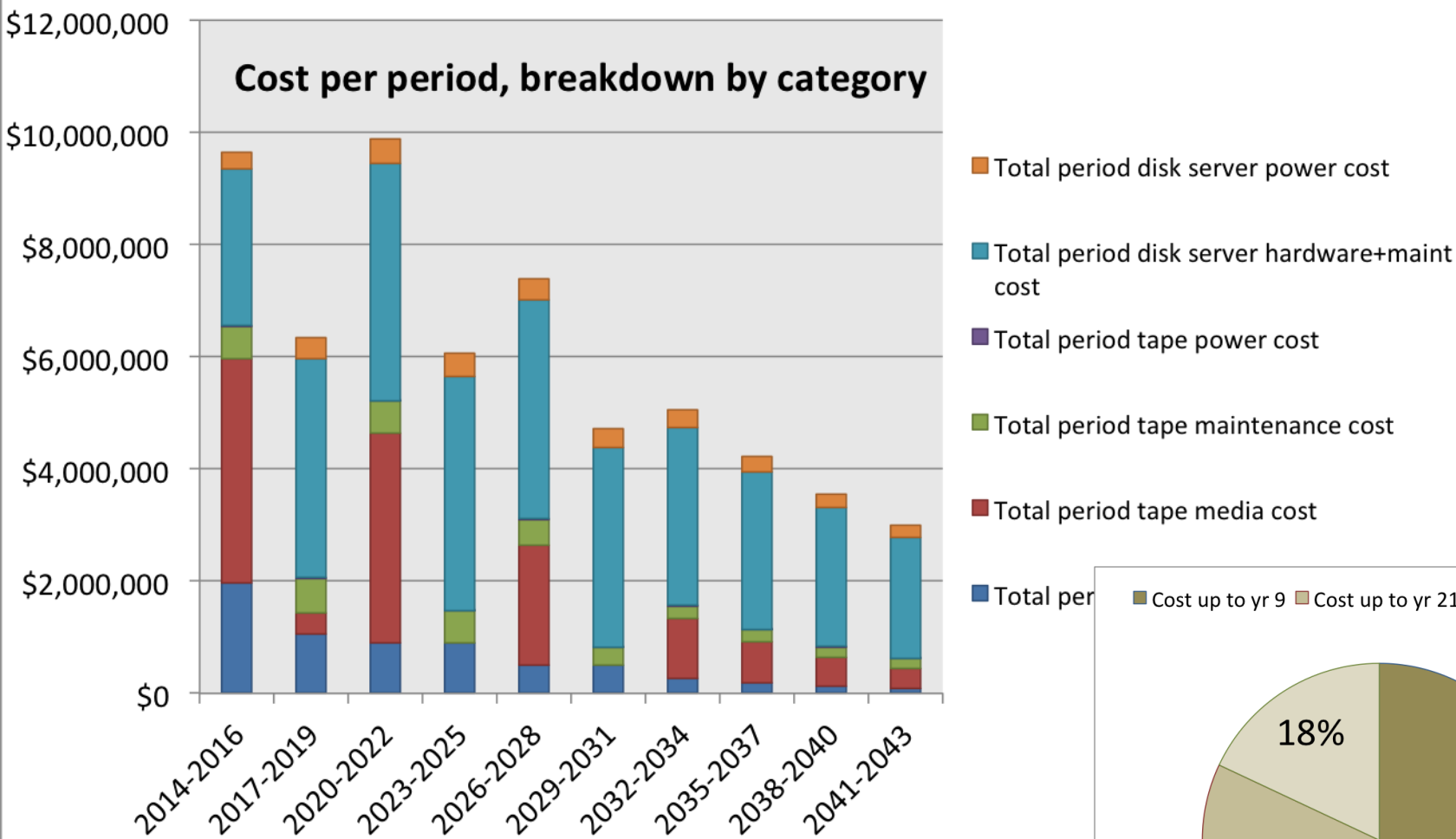
Total number of disks



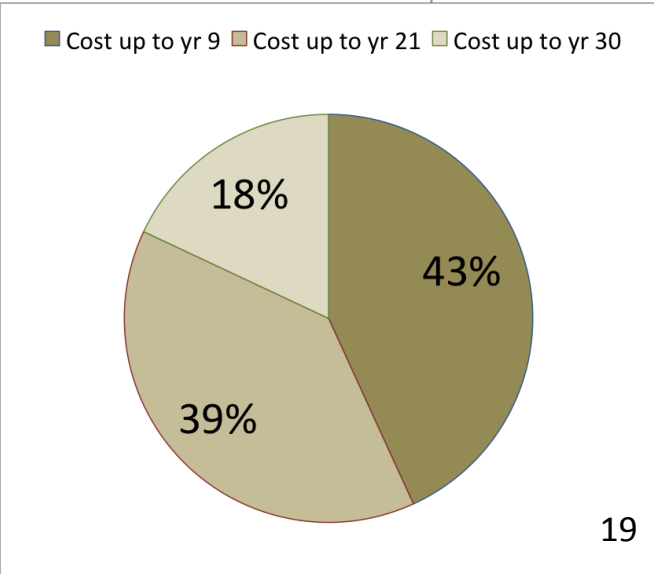
Total cartridges by end of generation



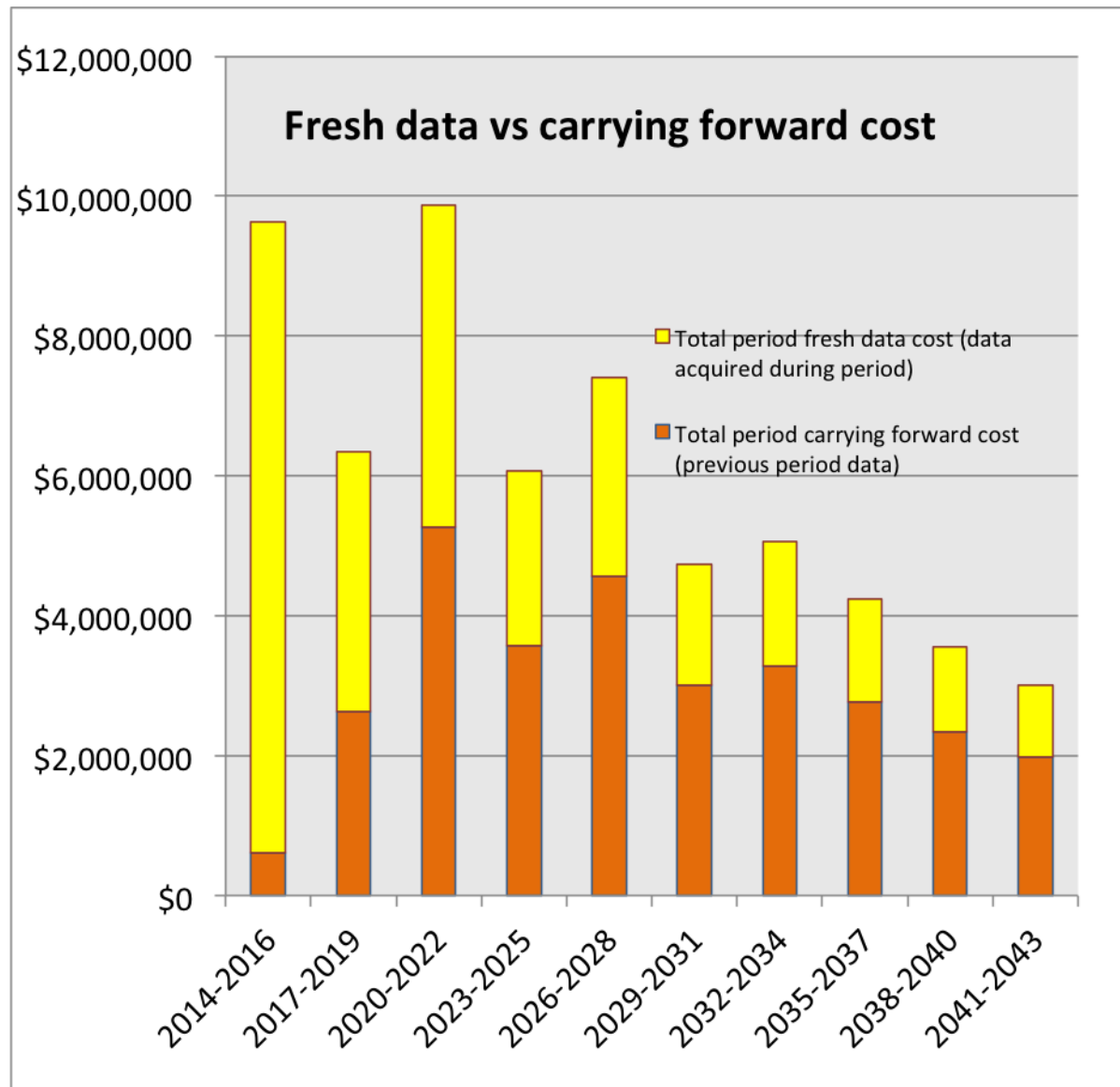
Case B) increasing archive growth



Total cost: ~59.9M\$
(~2M\$ / year)

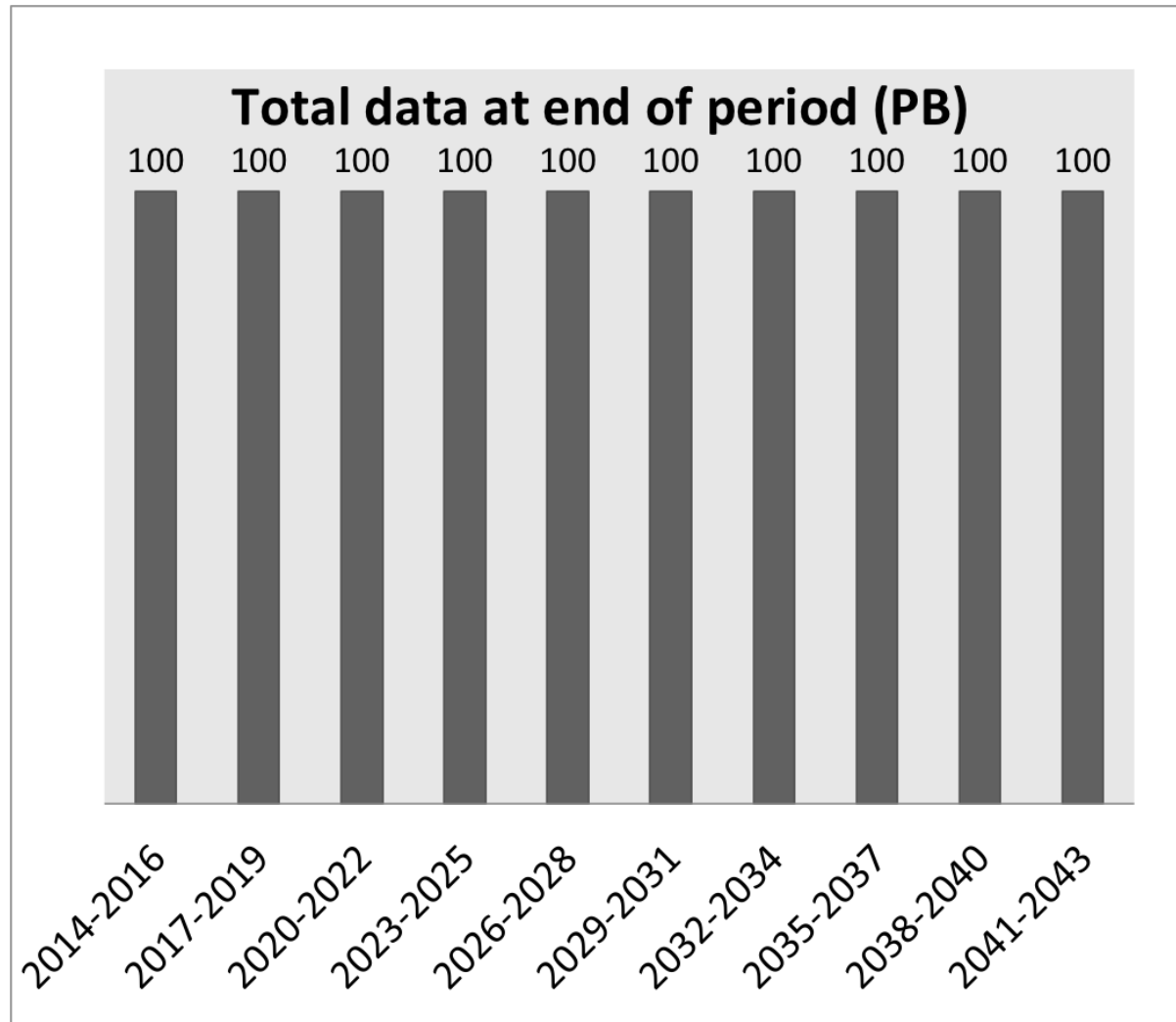


Case B) increasing archive growth



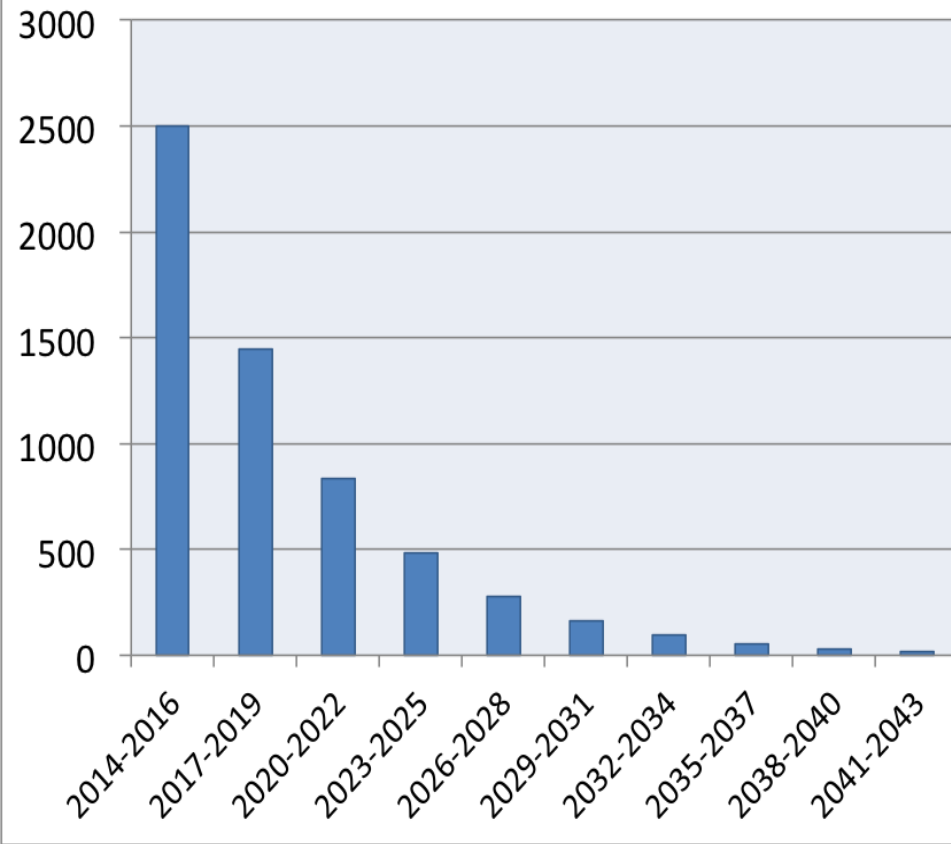
Case C) stable large archive

Start with 100PB, do not add any data

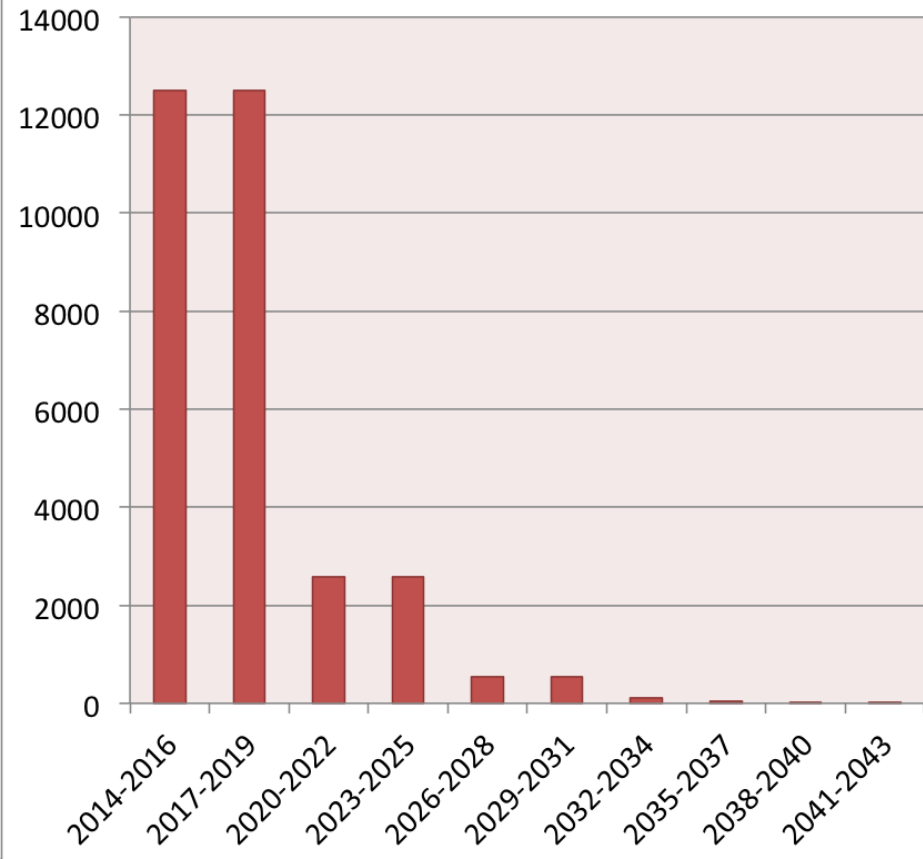


Case C) stable large archive

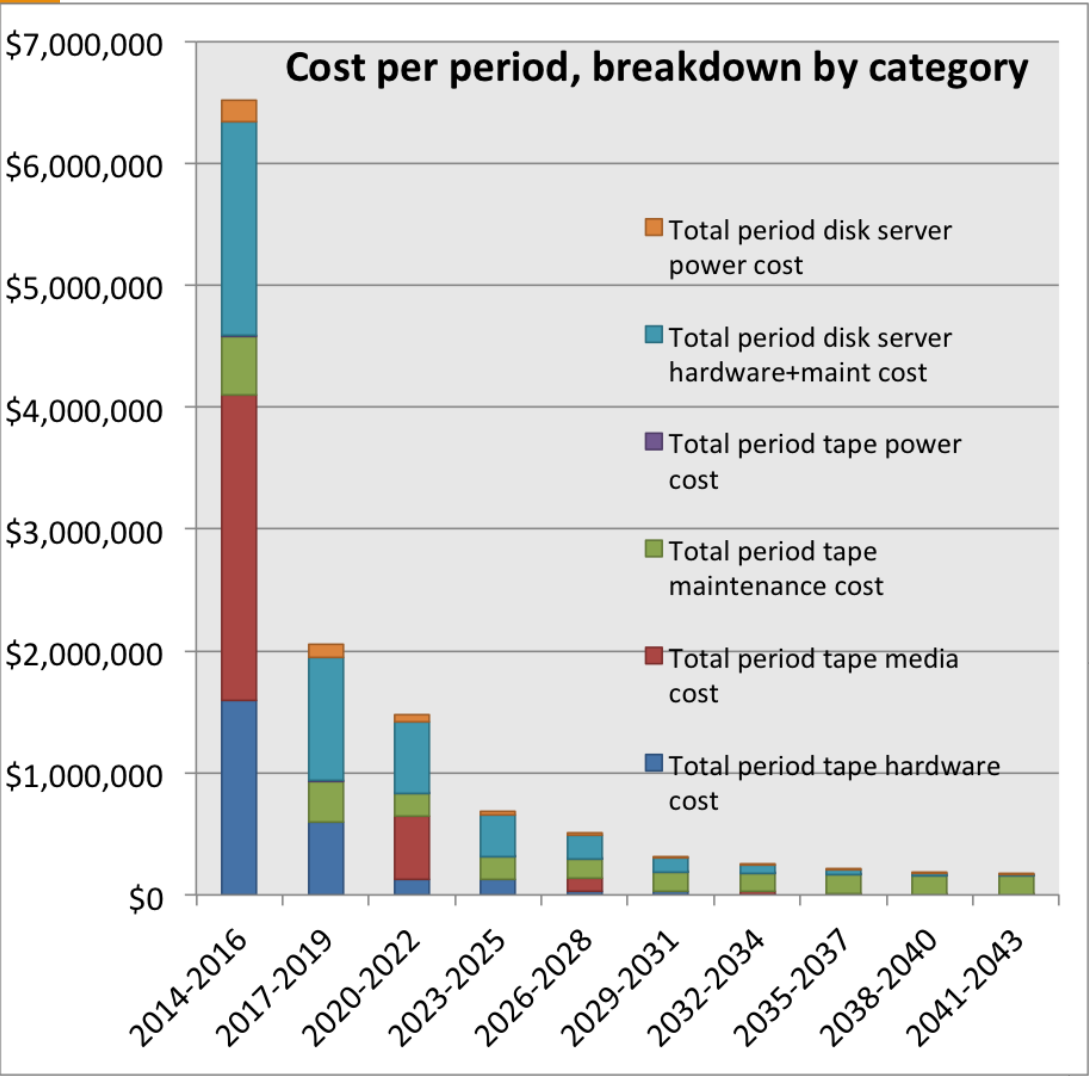
Total number of disks



Total cartridges by end of generation

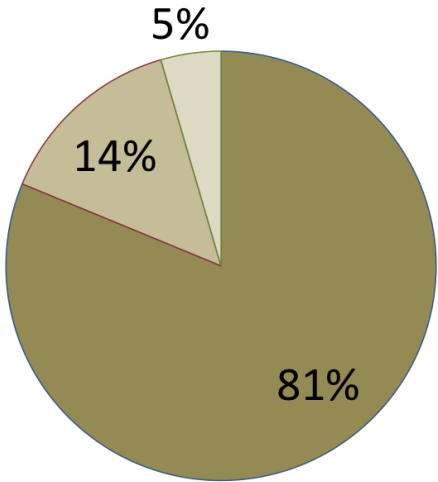


Case C) stable large archive



Total cost: ~12.3M\$
(400K\$ / year)

Cost up to yr 9 Cost up to yr 21 Cost up to yr 30



- INSIC Tape Roadmap, 2012 www.insic.org
- “A TCO analysis for Tape and Disk”, The Clipper Group, 2013 www.clipper.com
- “Enterprise Tape for Archival Storage?”, The Clipper Group, 2013 www.clipper.com
- “100 Year Archive Requirements Survey”, 2007 www.snia.org
- “Bit Preservation: A Solved Problem?”, D. Rosenthal, Stanford University, 2010 <http://www.ijdc.net>
- Oracle - Western States Contracting Alliance price list <http://www.oracle.com/us/corporate/pricing/wsca-homepage-081353.html>



Contributors welcome!

w3.hepix.org/bit-preservation

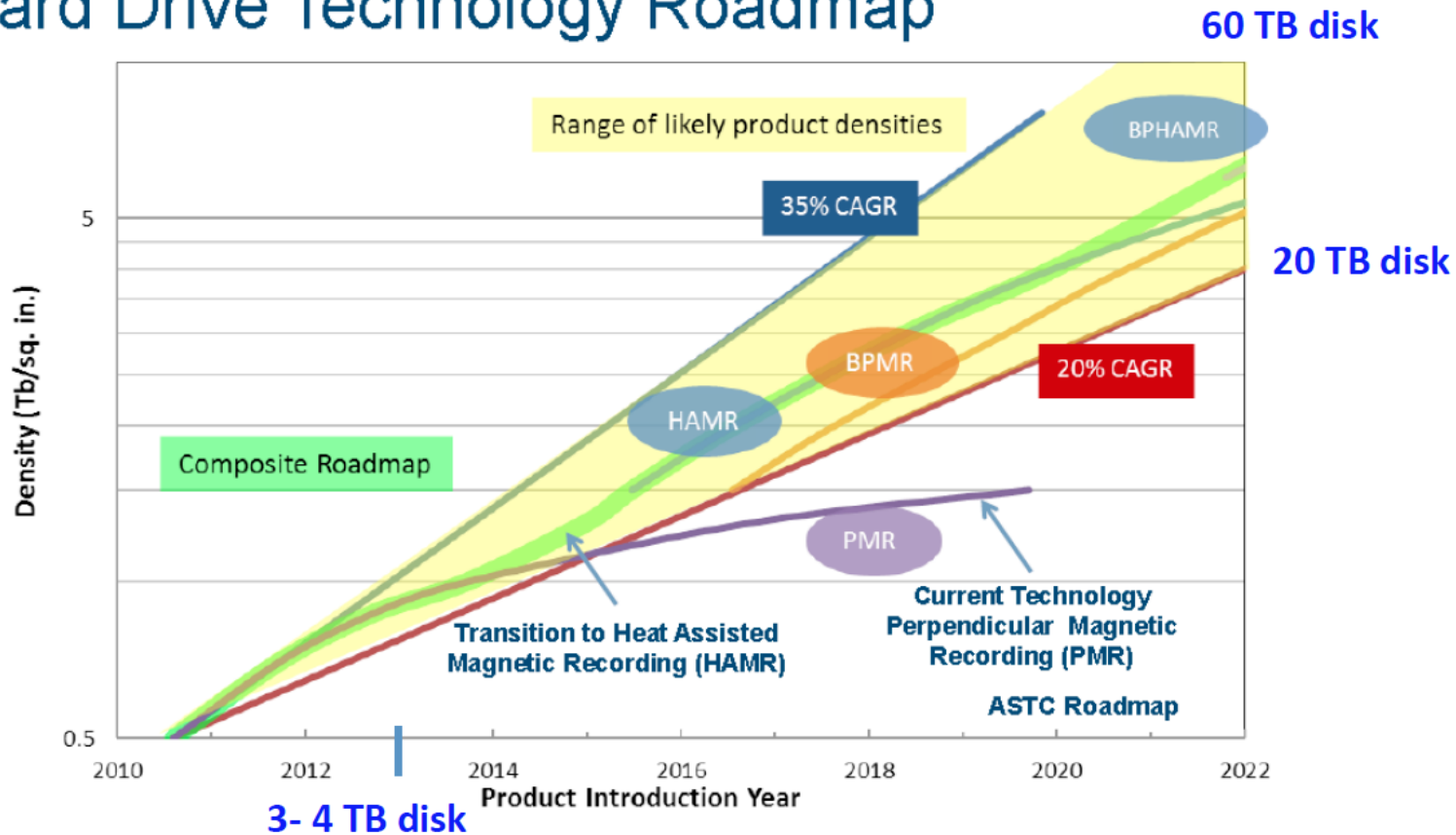
bit-preservation@hepix.org

Reserve slides

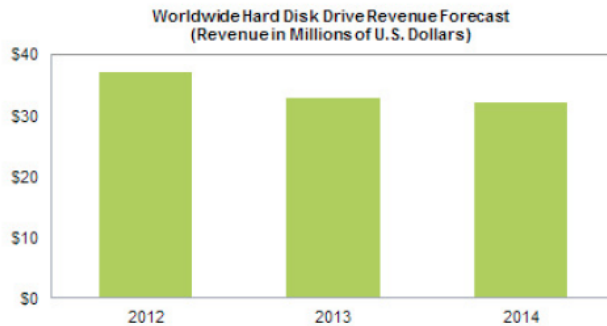
- The goal of the HEPiX Bit Preservation Working Group is to share ideas, practices and experience on bit stream preservation activities across sites providing long-term and large-scale archive services. Different aspects should be covered like: technology used for long-term archiving, definition of reliability, mitigation of data loss risks, monitoring/verification of the archive contents, procedures for recovering unavailable and/or lost data, procedures for archive migration to new-generation technology.
- The Working Group responds to a request by the [DPHEP](#) collaboration for advice on technical matters of bit preservation.
- The Working Group will produce a survey on existing practices across HEPiX and WLCG sites responsible for large-scale long-term archiving. The collaboration should ideally be extended to other large-scale archive sites from other research fields outside HEP.
- Based on best practices and development in storage preservation activities, the Working Group will provide recommendations for sustainable archival storage of HEP data across multiple sites and different technologies.

Technology (Disks)

Hard Drive Technology Roadmap



Markets (Disks)



Source: IHS iSuppli Research, February 2013

Decline of HDD market:

- Strong decline of desktop PCs (Q1 2013 -14%)
- Notebooks, Tablets and smartphones sales increase demands for flash memory (SSDs)
- consolidation of cloud storage (Steam, Netflix, iTunes)
- less copies, small caches

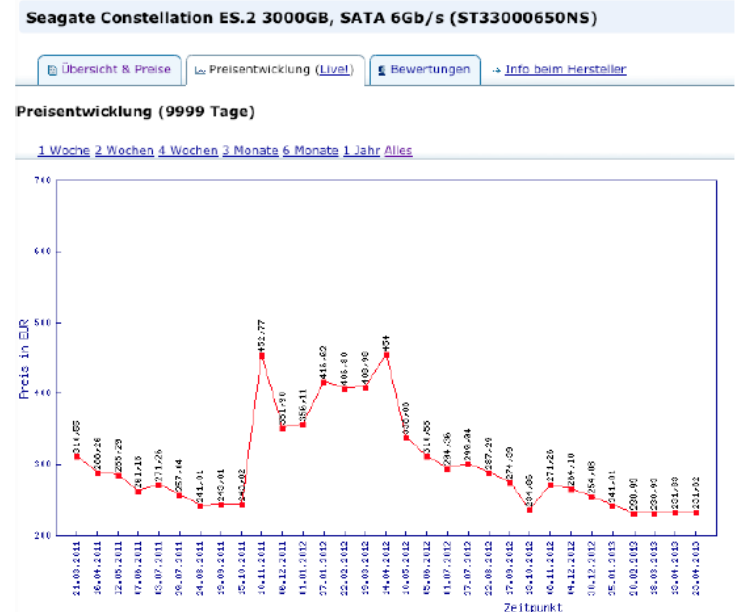
~80 million SSD shipments in 2013
Cost per GByte still 4-20 times higher than HDD

HDD WW Market Forecast (in million of units)

Year	Total HDD	Y/Y Growth
2012*	579.85	-6.8%
2013	565.76	-2.4%
2014	560.33	-1.0%
2015	556.56	-0.7%
2016	559.65	0.6%
2017	562.53	0.5%
	CAGR	-0.6%

* actual trendfocus

Fluctuating disk prices,
Difference between consumer disks and enterprise SATA disks is up to a factor 2
0.04 – 0.09 Euro/GB (raw disks)



Technology (Tapes)

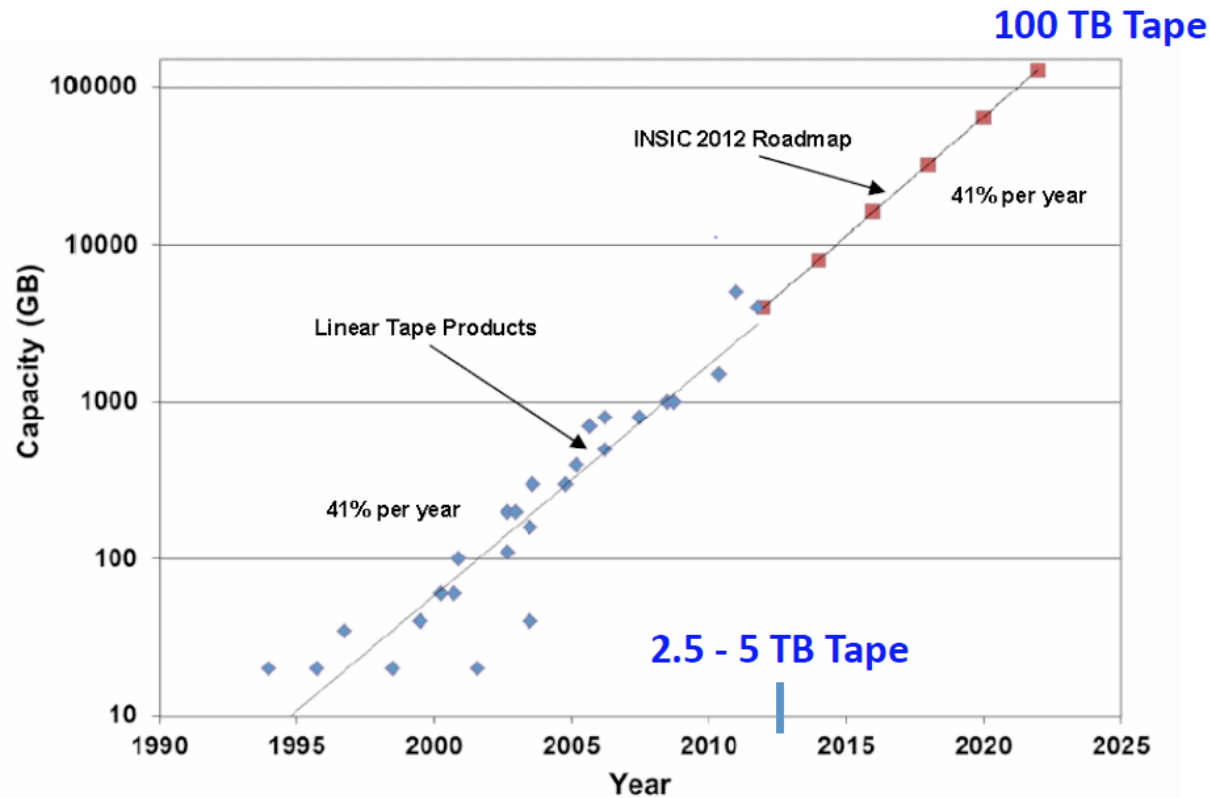


Figure 4: Tape Cartridge Capacity Trend.

© 2012 Information Storage Industry Consortium – All Rights Reserved
Reproduction Without Permission is Prohibited

International Magnetic Tape Storage Roadmap
May 2012

Markets (Tapes)

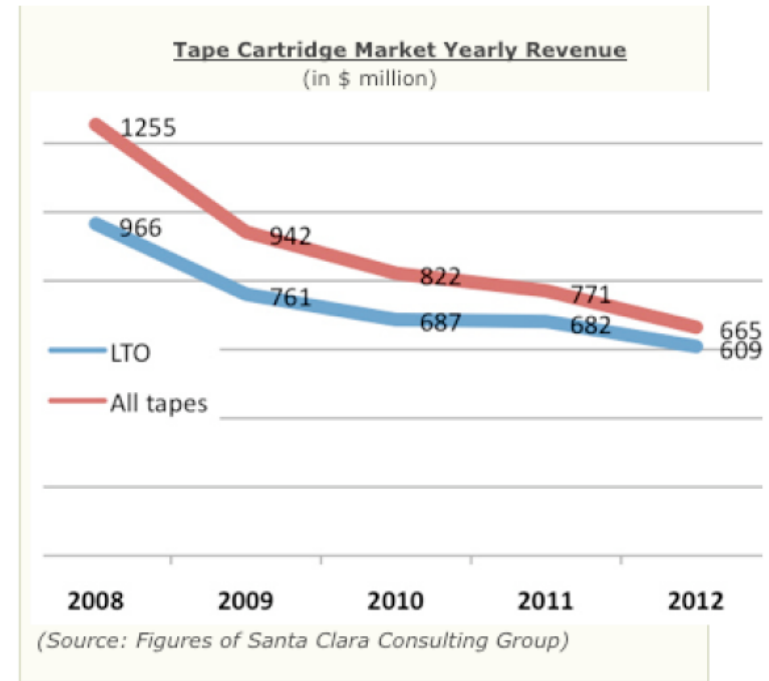
LTO has 93% market share
Enterprise tapes ~2%

Factor 2.5 cost decrease over 3 years
for cartridges

Today: 0.03 – 0.04 euro/GB

~23 Exabyte of tapes were sold in 2012
(backup large companies, scientific data, cloud storage)

To be compared with > 1000 Exabyte of worldwide data produced per year



Tape cartridge market:
-10% growth rate year by year