Data at Scale

Overview

Data at Scale Characteristics

- Data center storage capacity worldwide is at 1450 Exabyte. (1.45 *10^{21} B)
- Enterprise storage systems market is $13.2 billion second quarter of 2018 with growth i nthe order of 20% year to year

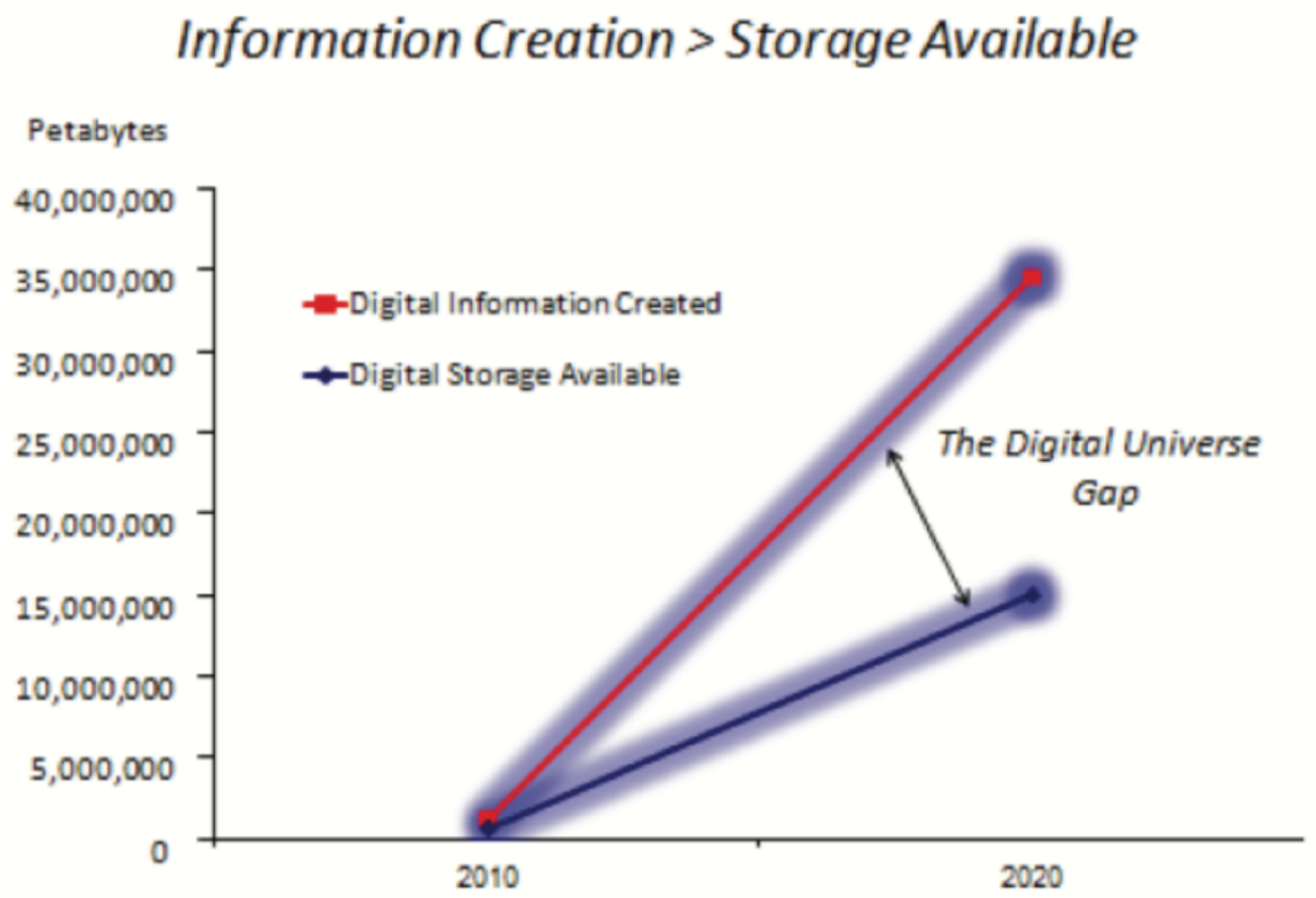
Fundamental Questions

- How to store all this data
- How to use all this data

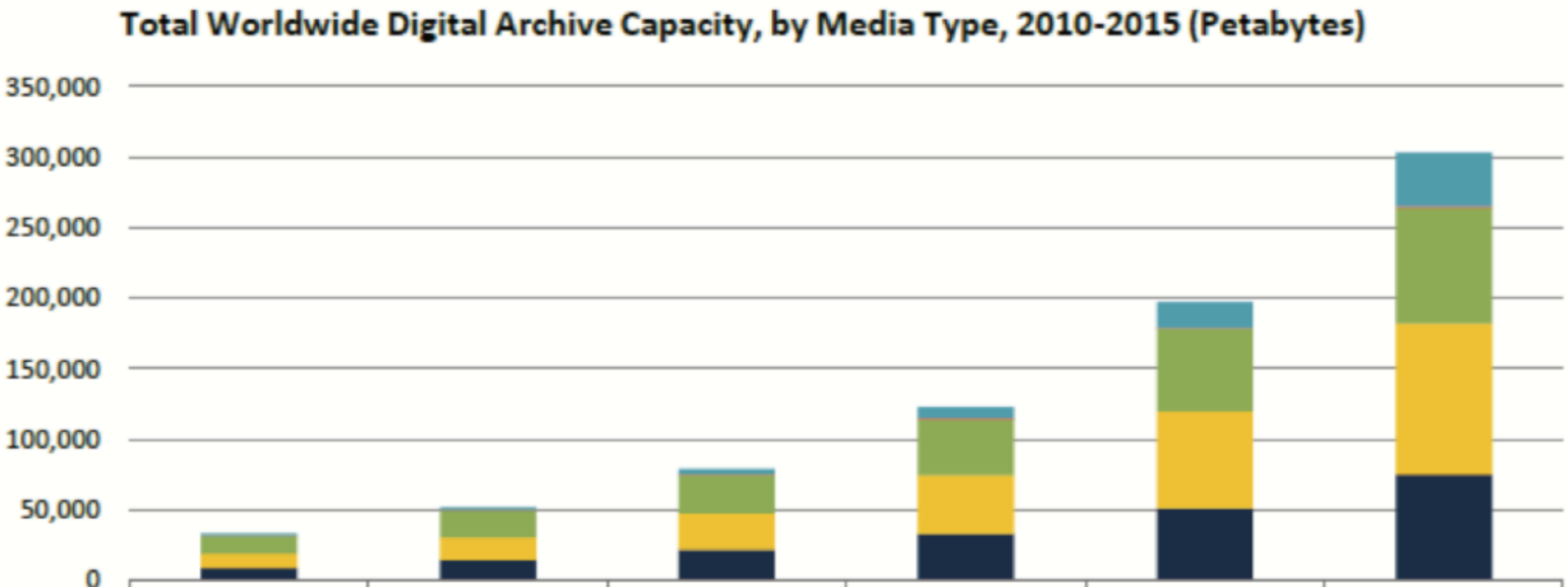
Fundamental Question

- How do we store all this data?
  - More fundamental question
    - Capability to store data determines how data can be processed

Emerging Gap between Digital Information and Available Storage

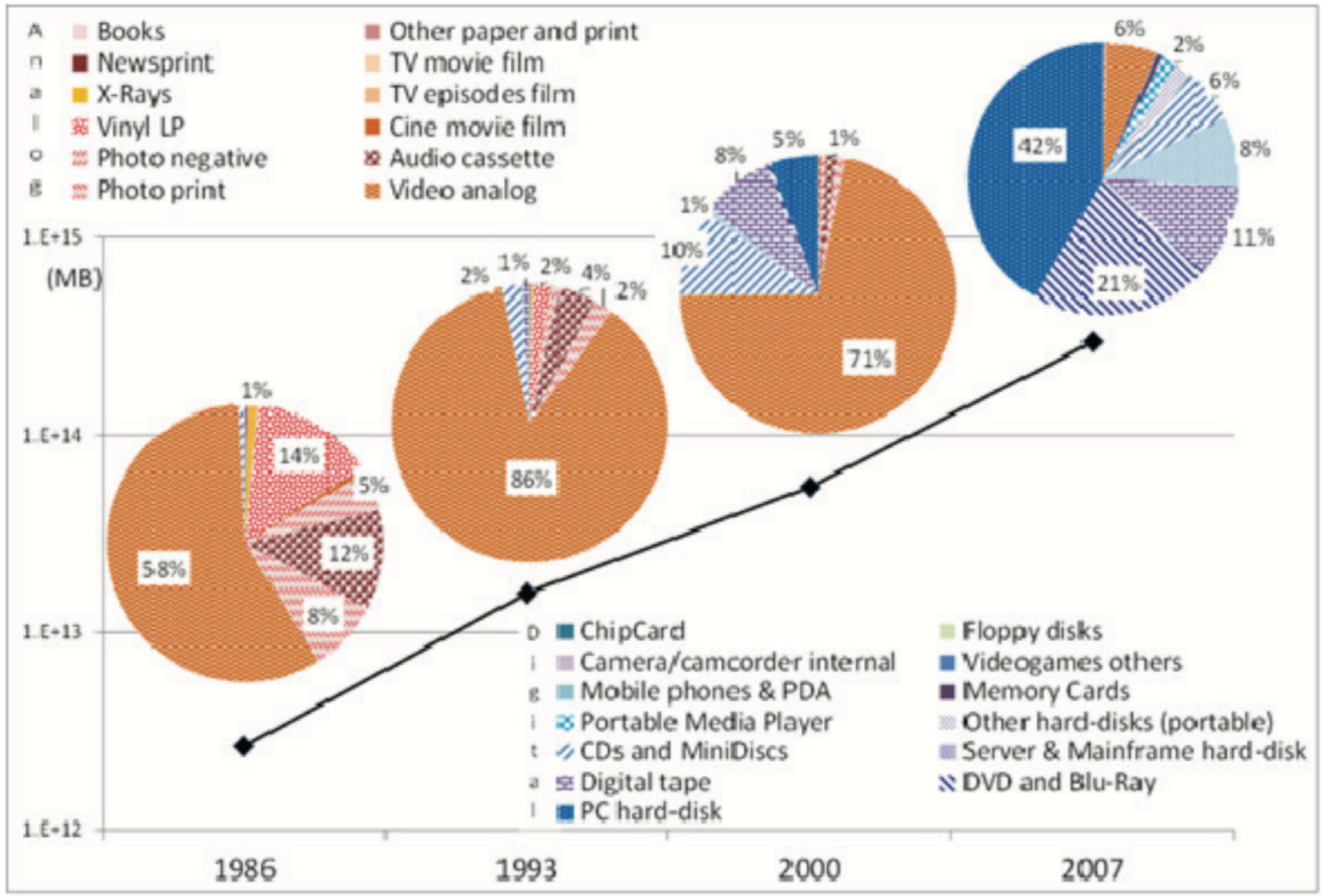


Digital Archive Capacity



	2010	2011	2012	2013	2014	2015
Cloud	768	1,708	3,676	8,033	17,908	37,846
Optical	303	410	537	714	968	1,247
Tape	12,784	18,939	27,116	39,362	58,220	81,562
External disk	9,712	16,053	25,637	41,501	68,430	106,830
Internal disk	9,650	14,881	22,185	33,547	51,707	75,510

World's Capacity to Store Information



Storage System Characteristics

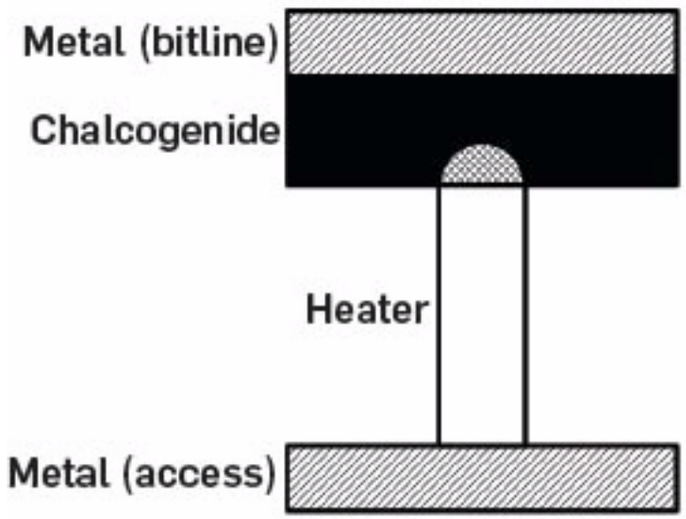
- Hierarchy (in 2022):
  - Persistent RAM (new technology coming to market)
  - Flash based storage (to be replaced by persistent RAM)
  - Disk based storage
  - Tape based storage

Storage Systems Characteristics

	Access Time	Capacity	Managed by
Level 1 Cache	1 cycle	4 KB	Software, Compiler
Level 2 Cache	2-4 cycles	64 KB	Hardware
Level 3 Cache	40 cycles	256 KB	Hardware
Main Memory	200-500 cycles	10 GB	Software / OS
Flash Drive	10 - 10 µsec	100 GB	Software / OS
Hard Disk	10 msec	10 TB	Software / OS

Storage Systems Characteristics

- Persistent RAM



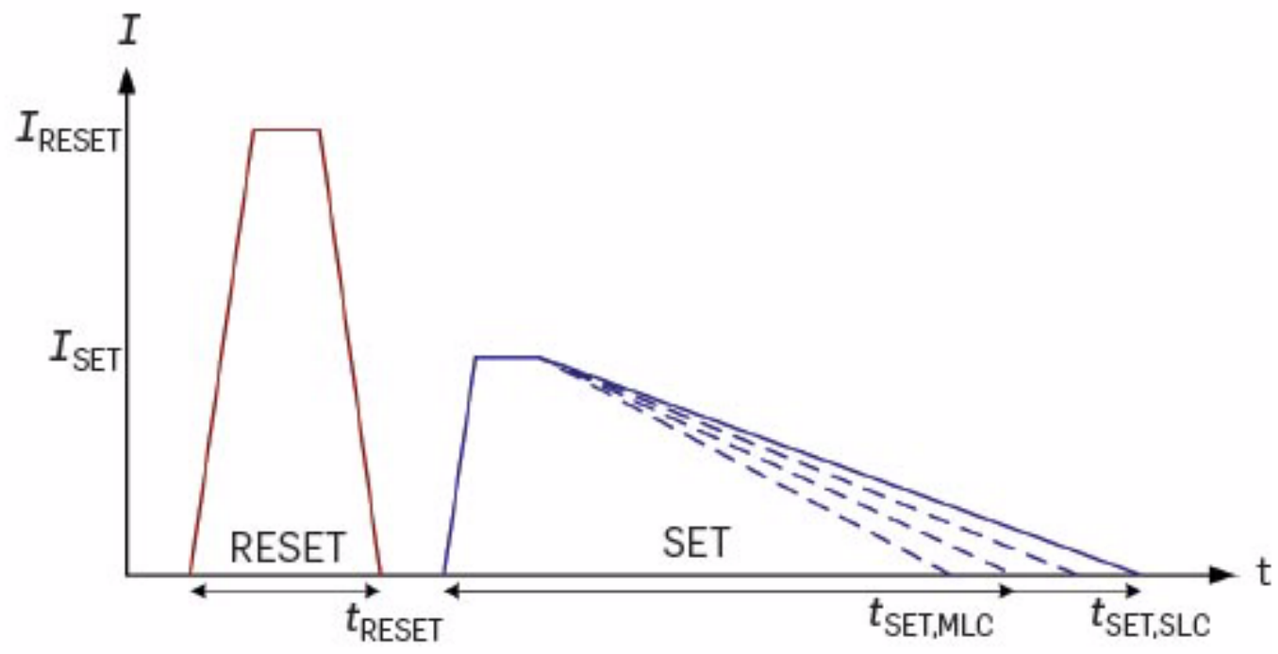
- Phase shift cell

Storage Systems Characteristics

- Persistent RAM: PCM
  - Byte addressable
  - Access times better / equal to DRAM
  - Resets destructive, heat needs to be controlled
  - More life-time cycles than FLASH
  - Cheaper than DRAM once entered production

Storage Systems Characteristics

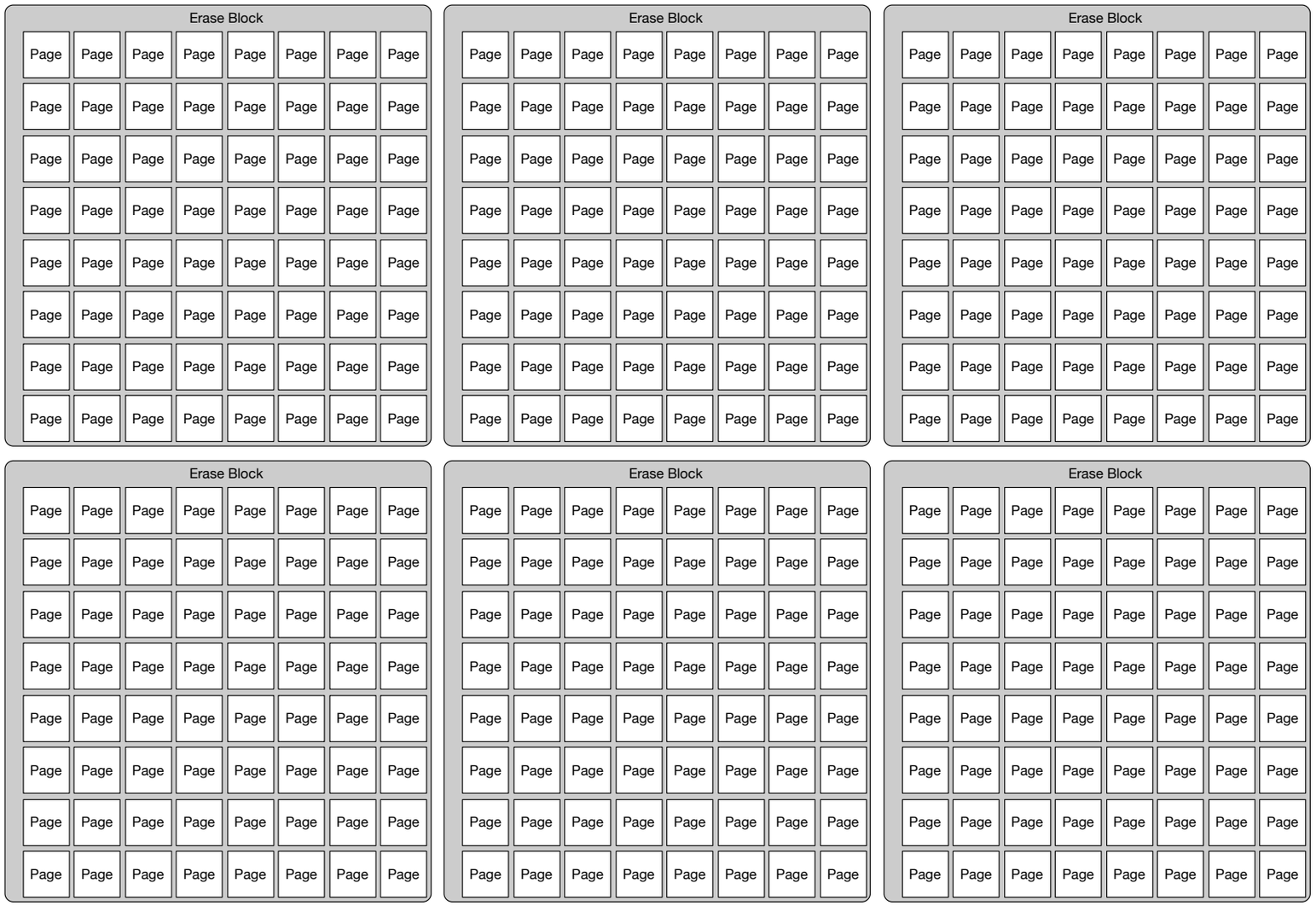
- Persistent RAM
  - Uneven write times
  - Research: How to minimize bit flips



Storage Systems Characteristics

- Flash
  - Stores several bits per cell
  - Write pages, each page has additional parity data
  - Erasure is destructive and takes longer
  - Data is erased in erase blocks

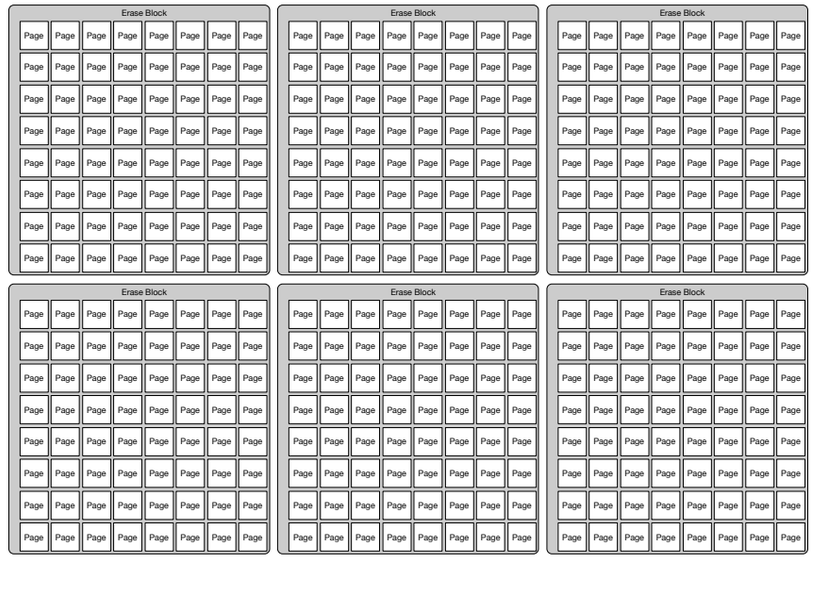
Flash Memory



- Garbage collection:
  - Copy active pages in order to create empty erase blocks
  - Erase the erase block
  - Write new pages in the new block

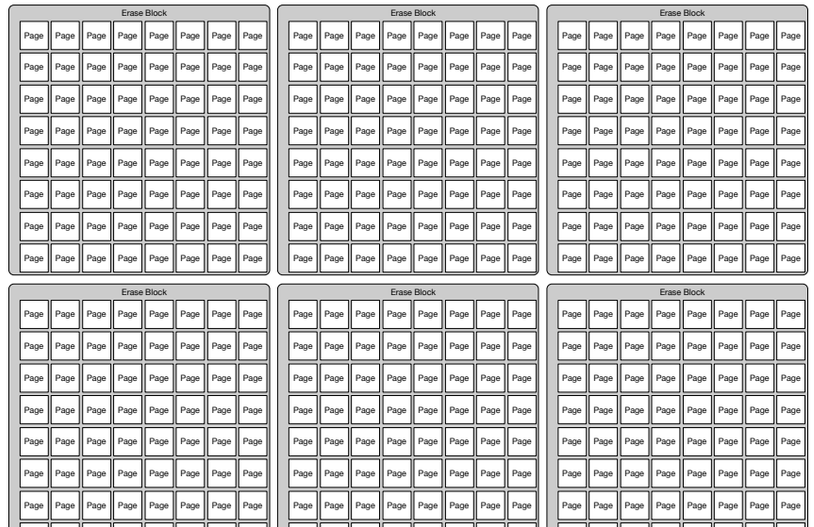
Flash Memory

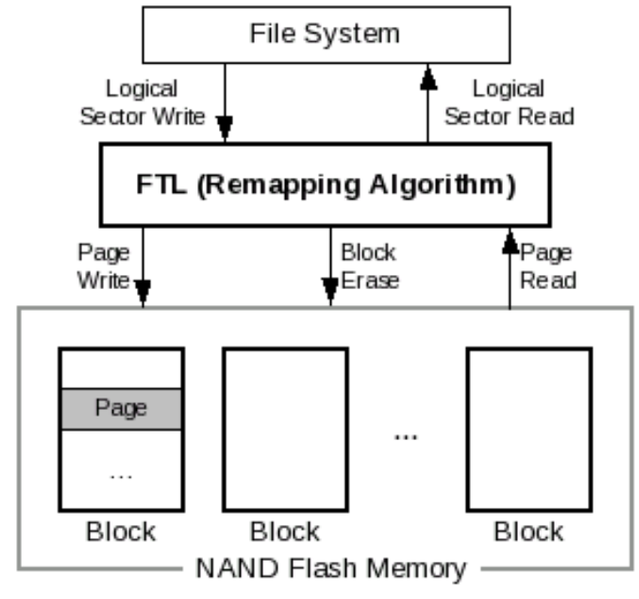
- Writing lots of data is fast at the beginning and then slows down as the drive starts garbage collection
- Write amplification
  - Writing one page can result in more than one page being written



Flash Memory

- Wear leveling:
  - Insures that erase blocks get erased about the same number of times
  - Lets file system write to a block (= a 4 KB page in flash-speak) at a given direction
  - But instead writes to a different page
  - With increasing flash size, can no longer just have a page translation page.



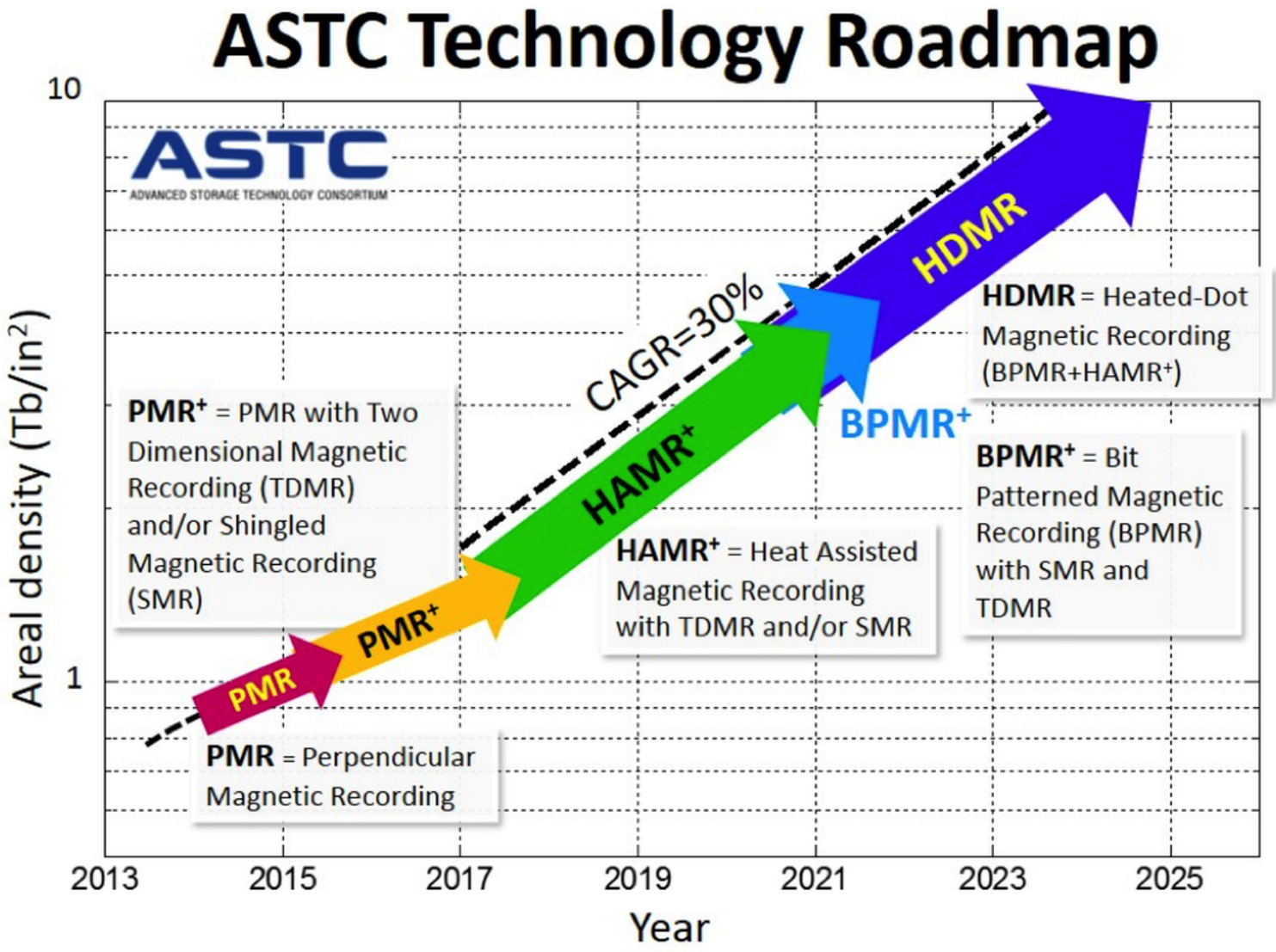


Flash Memory

- Reliability
  - As pages get overwritten, there is more likelihood of a cell no longer capable of holding data
  - Increases read times as it takes longer to sense voltage
  - Eventually, page contents can no longer be restored with error correcting parity data (1/32 of each page)
  - Page gets virtually swapped with a spare one, but eventually, capacity decreases

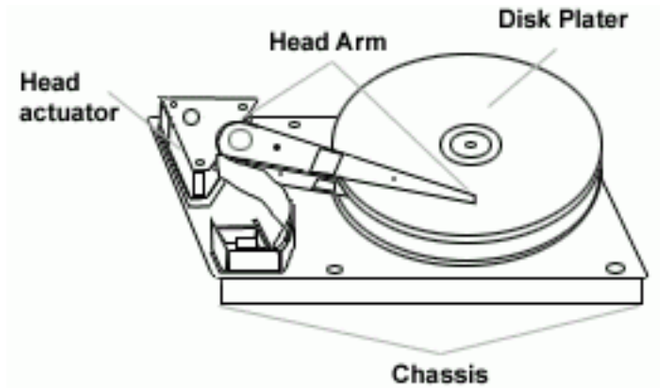
Disk Drives

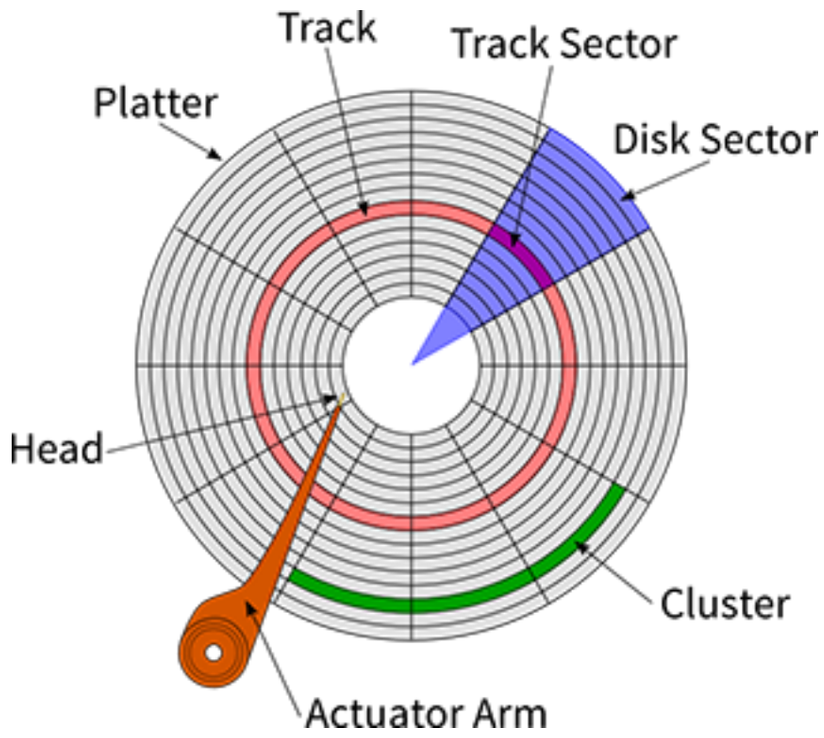
- Access Times around 10 msec
  - Has not changed much over the years
- Capacity is charging ahead



Disk Drives

- Store data by magnetizing a thin magnetic layer on a platter





Flash vs. Disk Access Times

- Flash data access: 550 MB per second read, 520 MB per second write (peak)
- Hard Drive access: 128 MB per second read, 120 MB per second write
- Flash costs per GB $0.20
- HDD costs per GB $ 0.03

Near Disk Storage

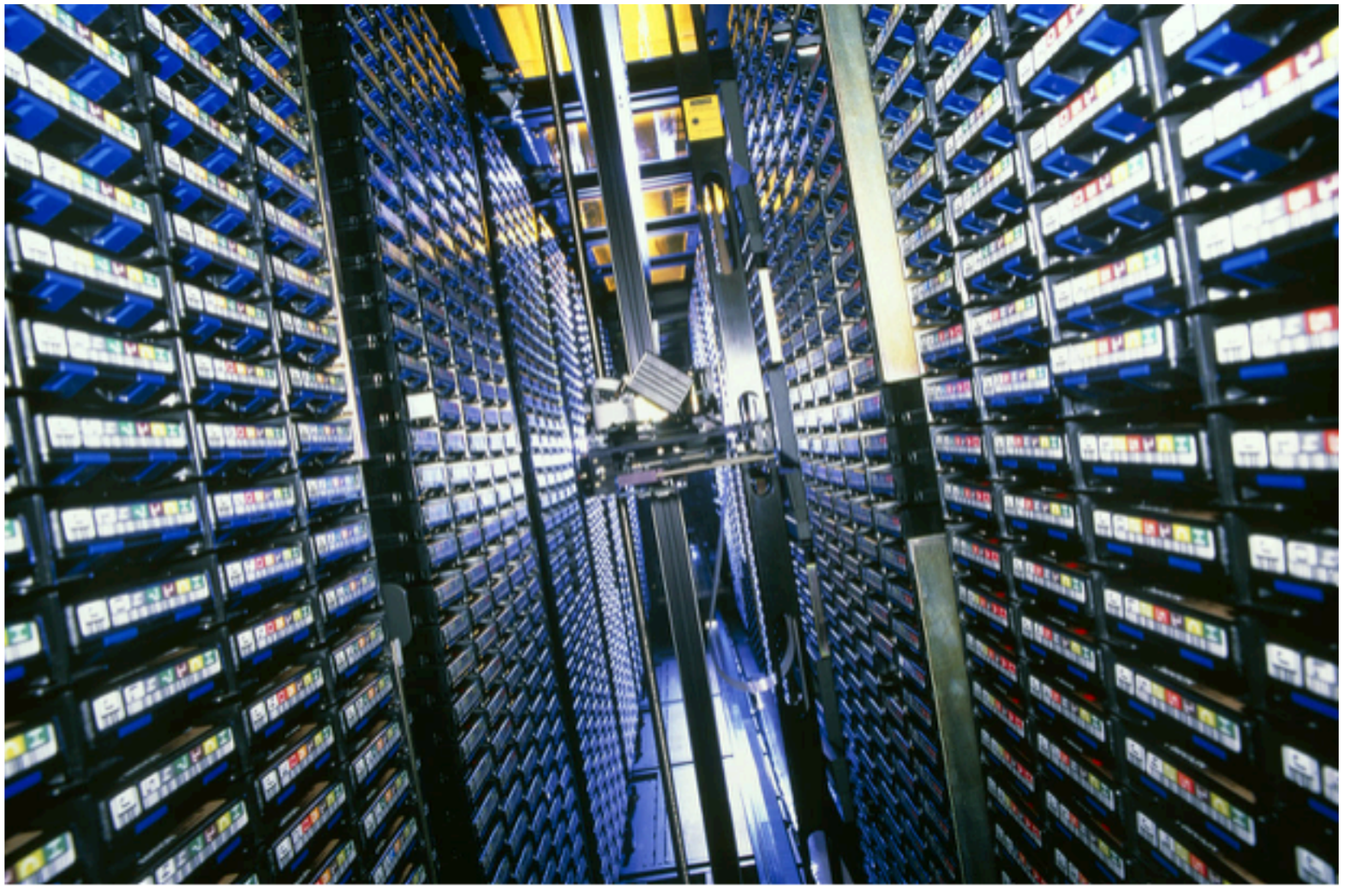
- Near disk storage :
  - Disks are not immediately accessible but need to be loaded (like a tape)

Tape

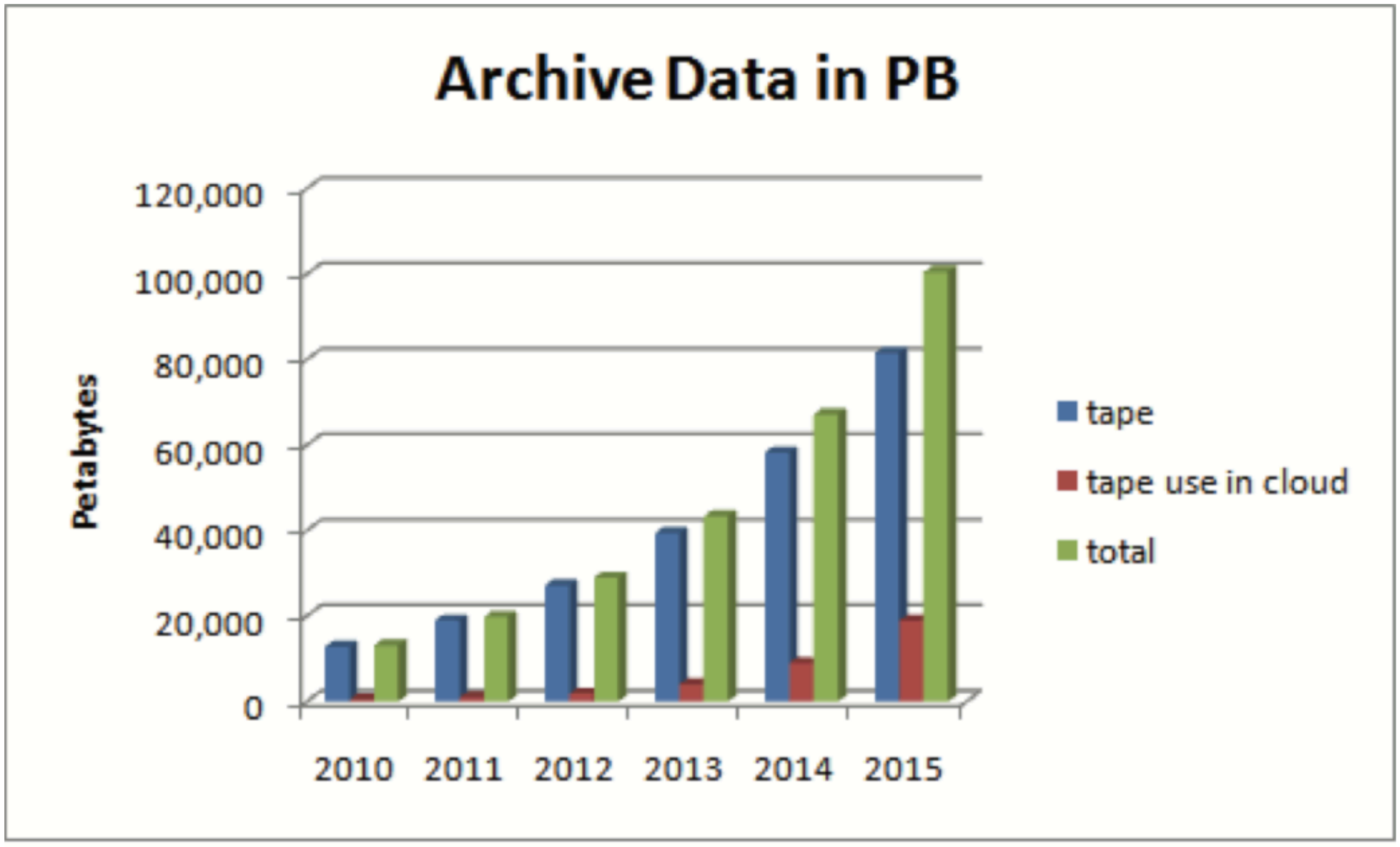
- Oldest storage medium, but still developed

Linear Tape Open (LTO) cartridge

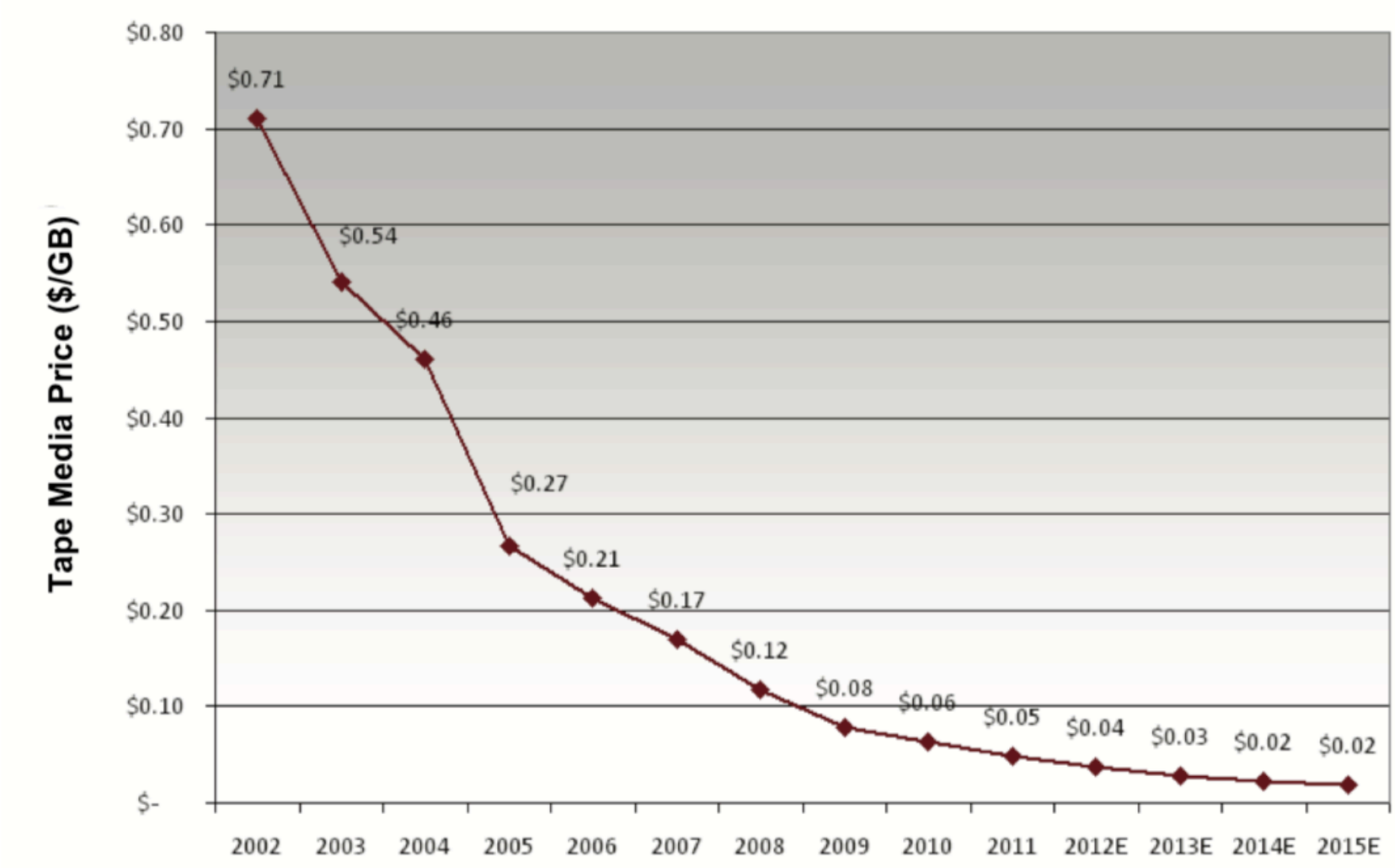
Tape Storage System



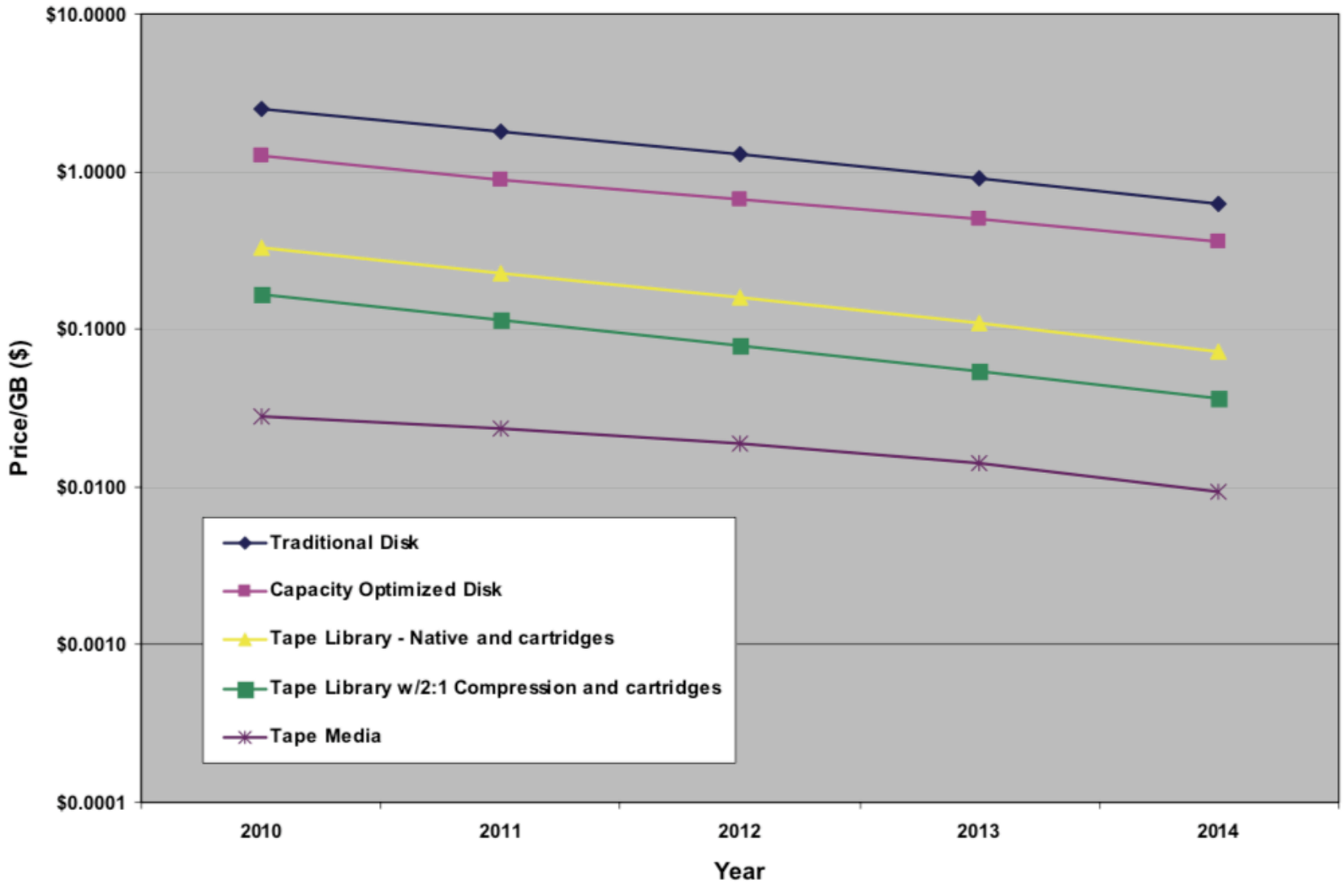
Tape Storage System



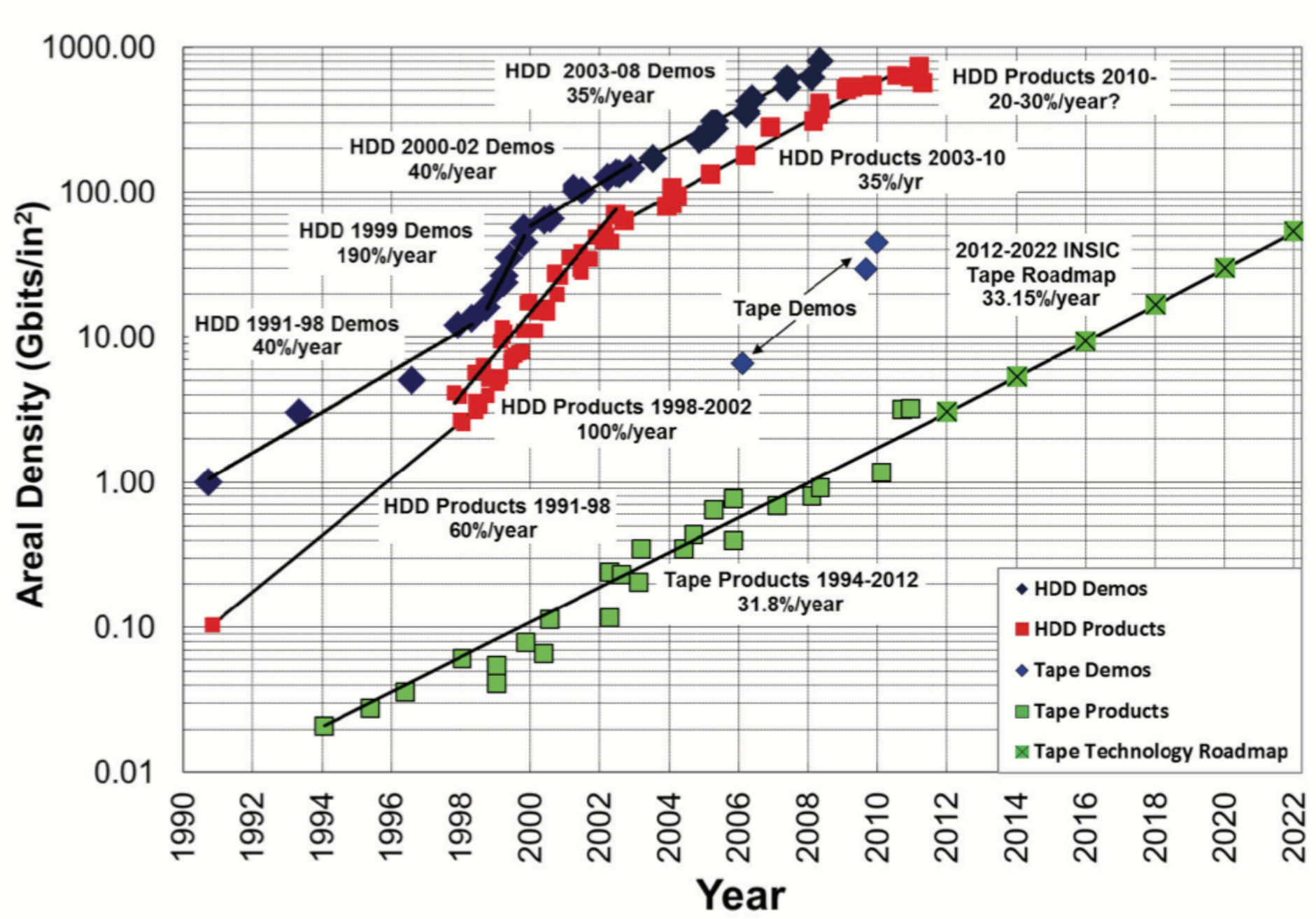
Tape Storage System



Tape Storage Systems



Tape Storage Systems



Summary

- We will continue to have a memory - storage hierarchy
- I predict for 2025
  - DRAM replaced largely by PCM
  - Flash replaced largely by PCM
  - Disk drives take on a more archival role
  - Tape systems continue to be “deep backup”

Summary

- Caching:
  - Keep important data near towards top of the stack
  - Move less important data towards the bottom of the stack
- Backup
  - Make copies of important data
  - Store in less accessible, cheaper storage

Reliability

- No storage device is infallible
- Failure rates can be hard to specify
  - Example: HDD
    - Common failure mode: Device no longer functions
    - Latent sector failures: Sector becomes unreadable
- Failure rates can be as low as MTTF of 2,000,000 hrs (233 years) (same as flash), but can be in the 5000 hrs.

Reliability

- How to deal with failures:
  - Detect failures
    - Heart beat monitoring
    - Error detecting coding
  - Restored failed data
    - Replication
    - Error correcting codes
    - Backups

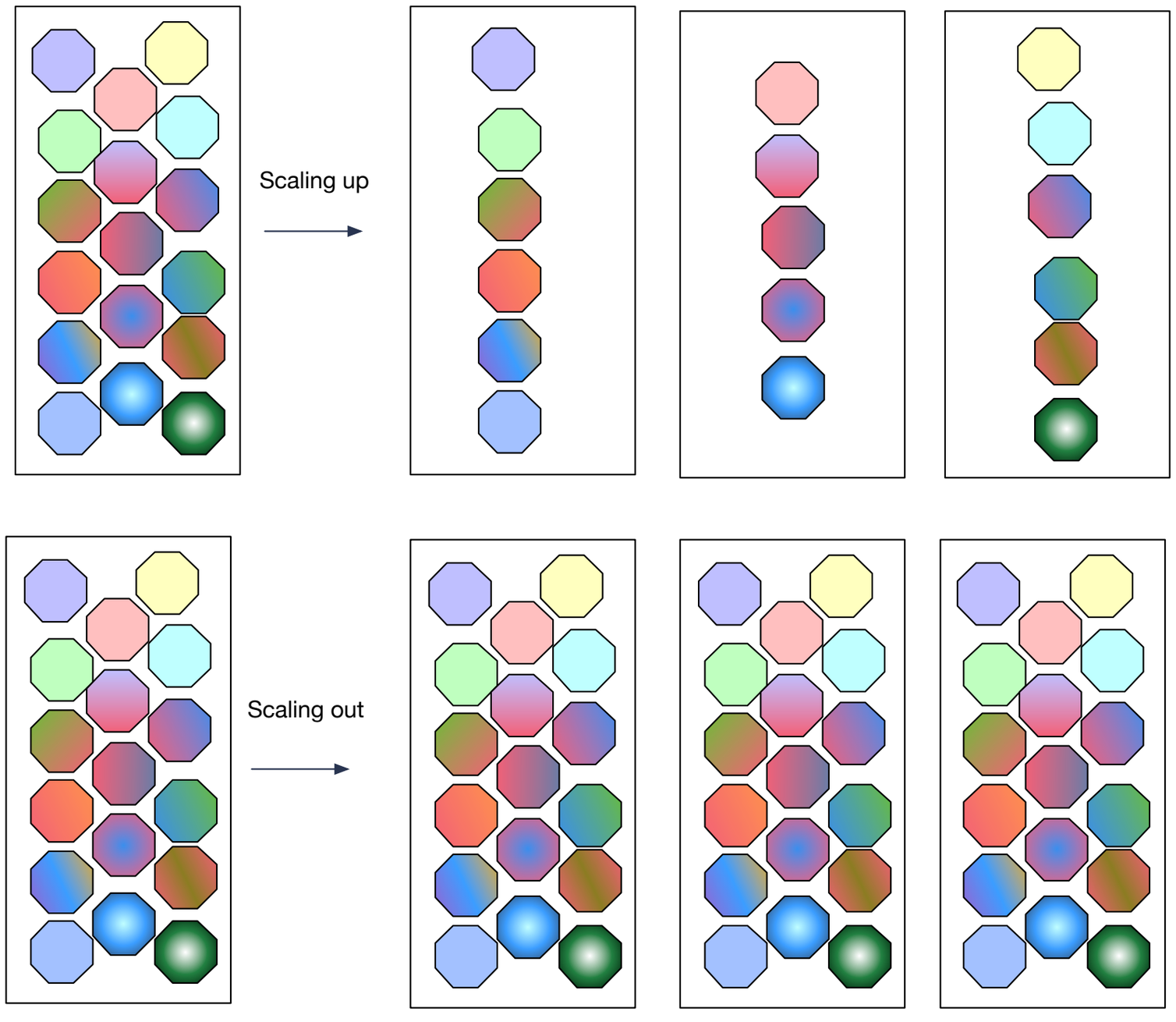
Backups

- Complete backup can be impossible because of the time it takes to access all data
  - Delta backups
    - Backup only data that has been changed
    - Industry standard mixes complete backups with delta backups
- Snapshots:
  - Freeze a file system conceptually at one time.
    - Changes are either before or after
    - Snapshot is always a consistent view
      - Has difficulties with *monster transactions:*
        - E.g. a transaction that changes all records

Replication

- Replication for :
  - Availability
    - Survives failure affecting all but one copy
  - Scalability
    - Scaling up: Use more systems
    - Scaling out: Replicate data on more than one system

Replication

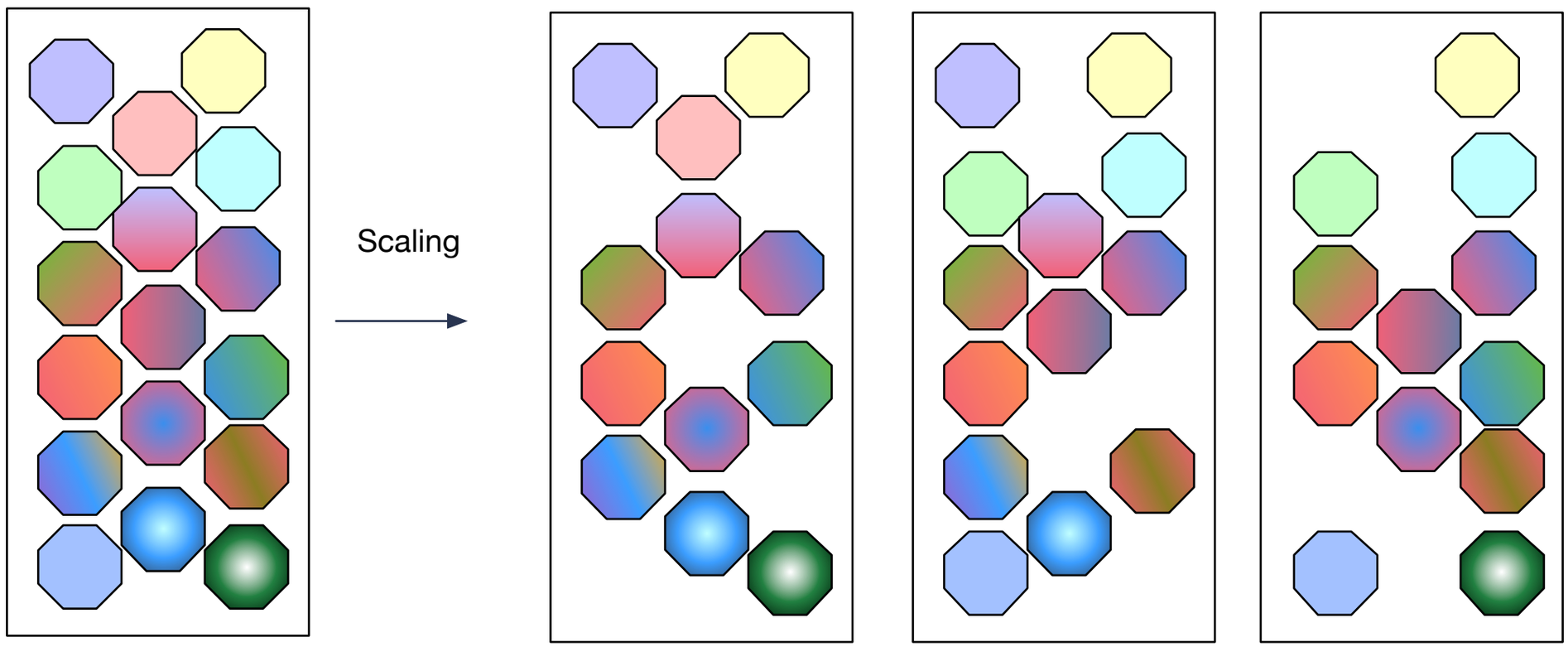


Provides performance, does not protect against failure

Provides robustness, provides performance, needs consistency control

Replication

- Hybrid approaches:
  - Combine scaling up and scaling out



Replication

- Scaling provides elasticity
  - At the cost of administrative decisions
  - Under - or over-provisioning of services

Replication

- Replica placement
  - Best close to the client
    - Acamai: Intelligent edge platform
  - Server initiated replica distribution
    - Systems decide whether to migrate or replicate data
    - Typical for Content Delivery Networks
- Client initiated replica distribution
  - Not constrained by system rules

Replication

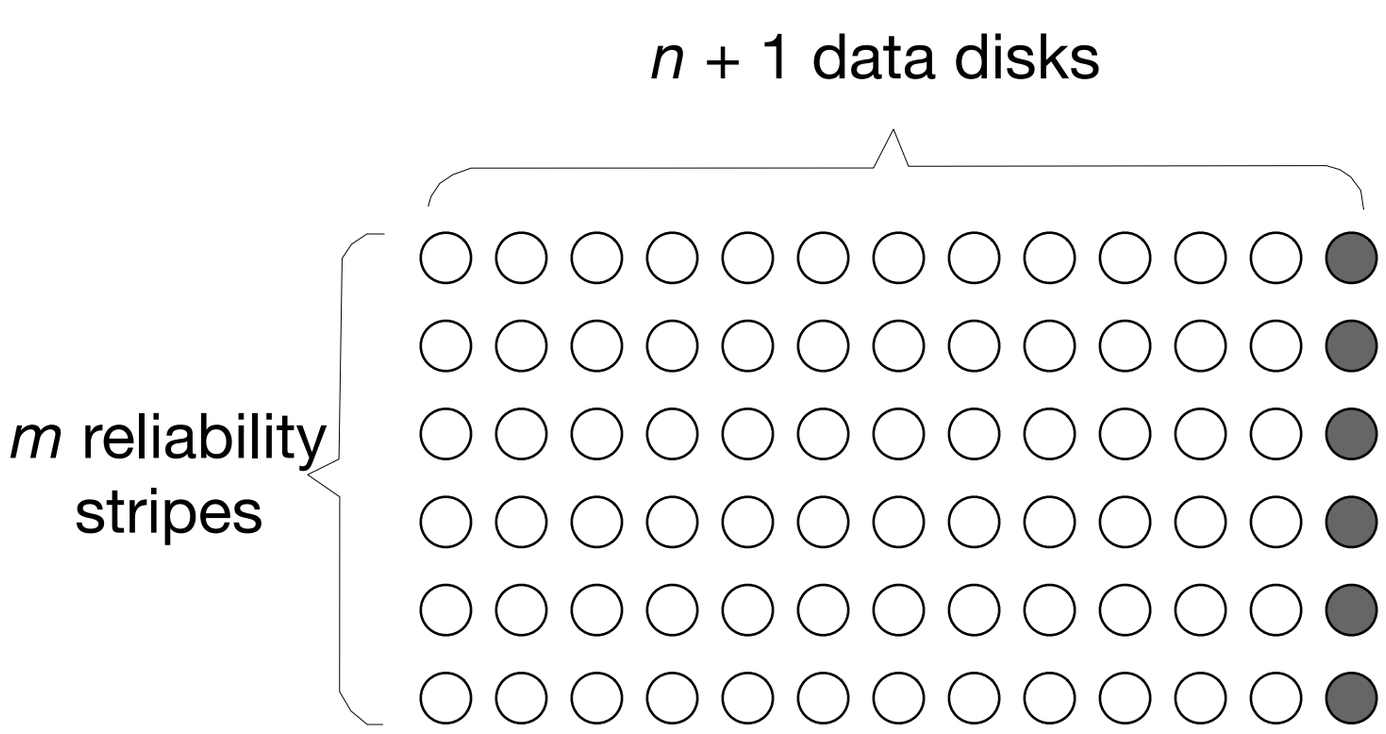
- Consistency challenges
  - Replica need to be equal
  - Clients can change replica
  - Strong consistency:
    - A read from any replica will always produce the same result, regardless of a concurrent update
  - Low consistency:
    - Reads will eventually converge to the same value in the absence of udpates

Erasure Correcting Coding

- Combine *n* data items into a reliability stripe
- Add *k* parity items calculated with an ECC
- Can recover all data items if *n* out of *n+k* data items are still available

Erasure Correcting Coding

- RAID Level 5
  - Consists of stripes of *n*+1 disks
  - Last disk contains the bit-wise parity of the other disks



Erasure Correcting Coding

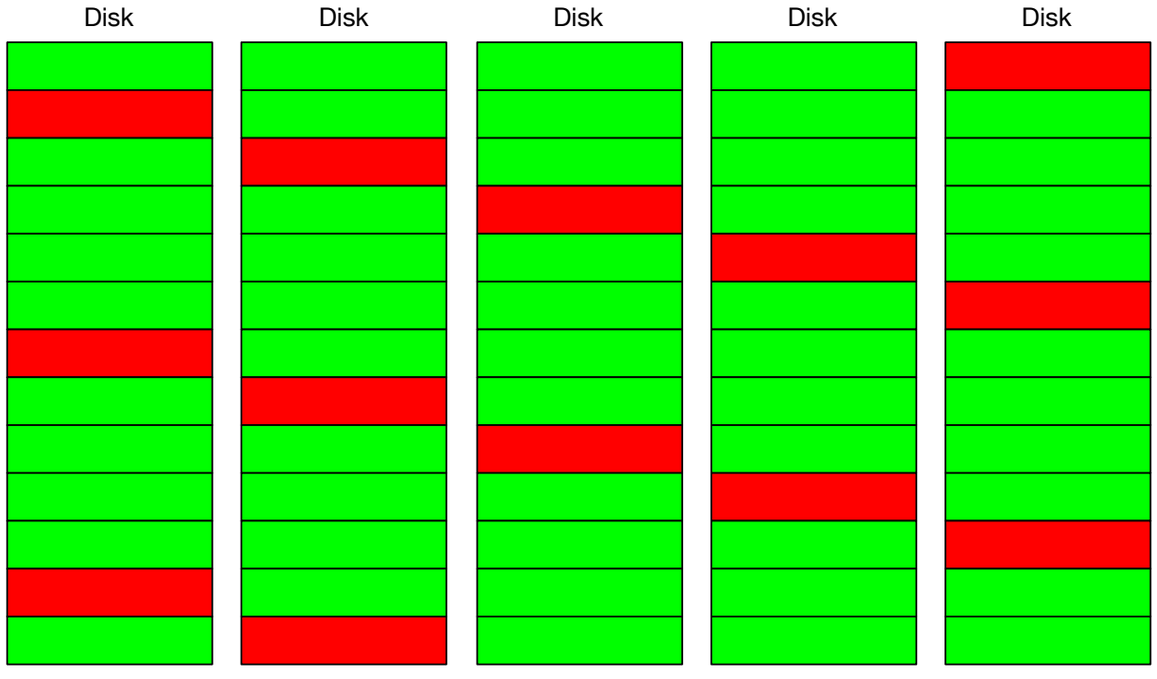
- If we write to a single disk, we will also need to update the parity disk
  - “Small write operation”
    - Read old data
    - Calculate the exclusive-or of the old data with the new data
    - Read the old parity
    - Calculate the exclusive-or of the old parity and the exclusive-or of the new data and the old data and write it to parity
    - Write the new data
- Total: 2 reads and 2 writes
- A disk will need to rotate twice

Erasure Correcting Codes

- Efficient:
  - Rewrite complete stripes
  - For partial stripe writes, create new parity block, then replace old parity block with new one

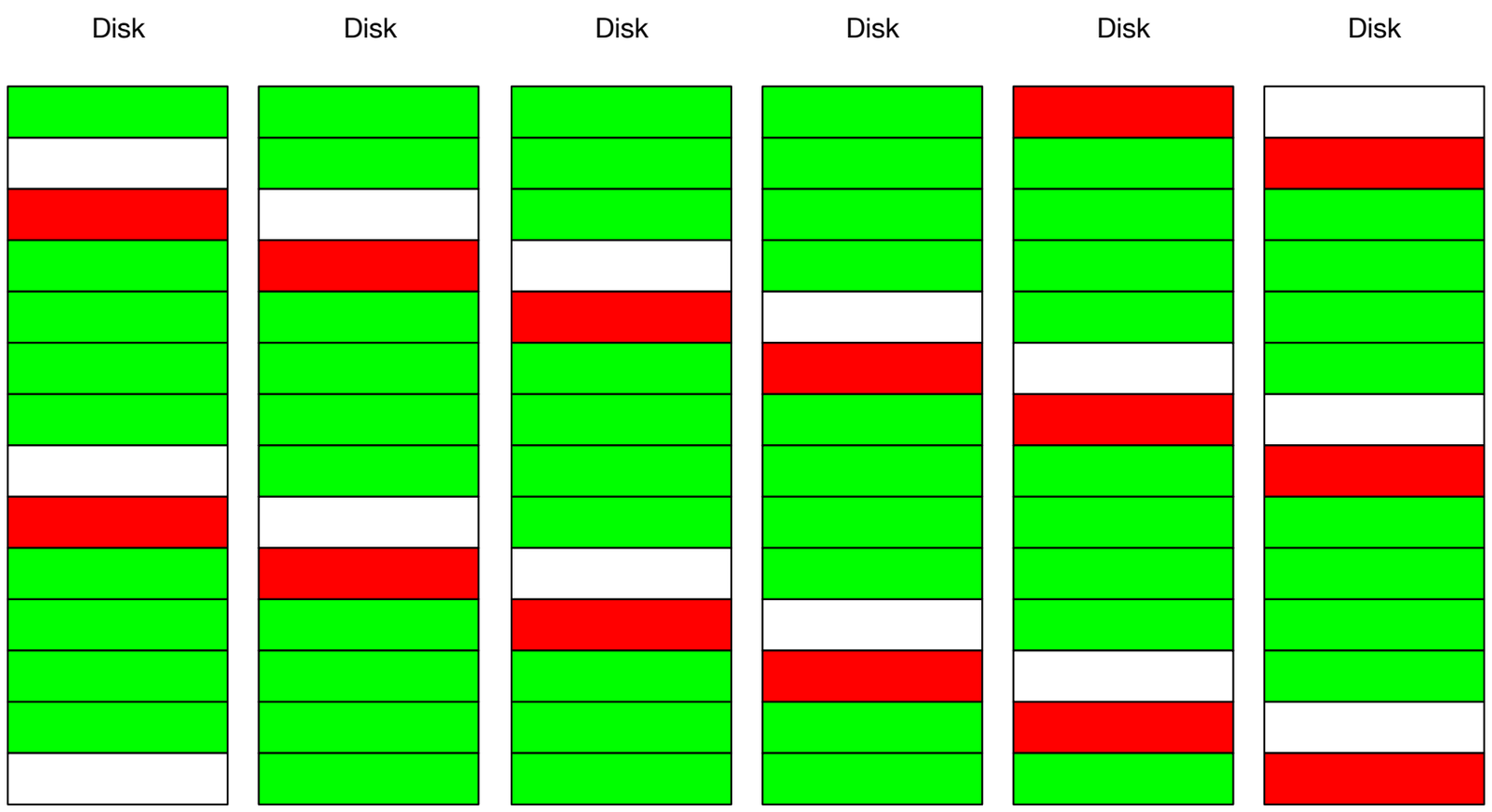
Erasure Correcting Coding

- Writes are concentrated in the parity drives
- Reads are exclusively to the data drives (unless there is a failure)
- Distribute roles by dividing disk into disklets
- Rotate roles



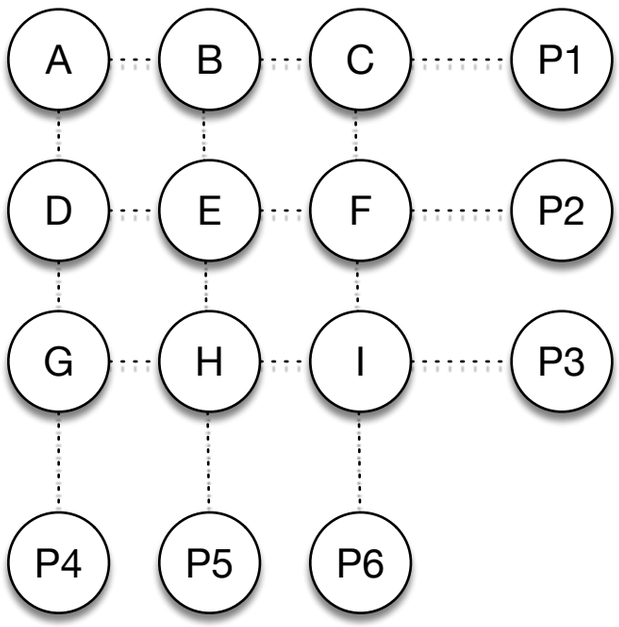
Erasure Correcting Coding

- If we design a disk array, we might want to add a spare disk (or two)
- Can distribute this as well



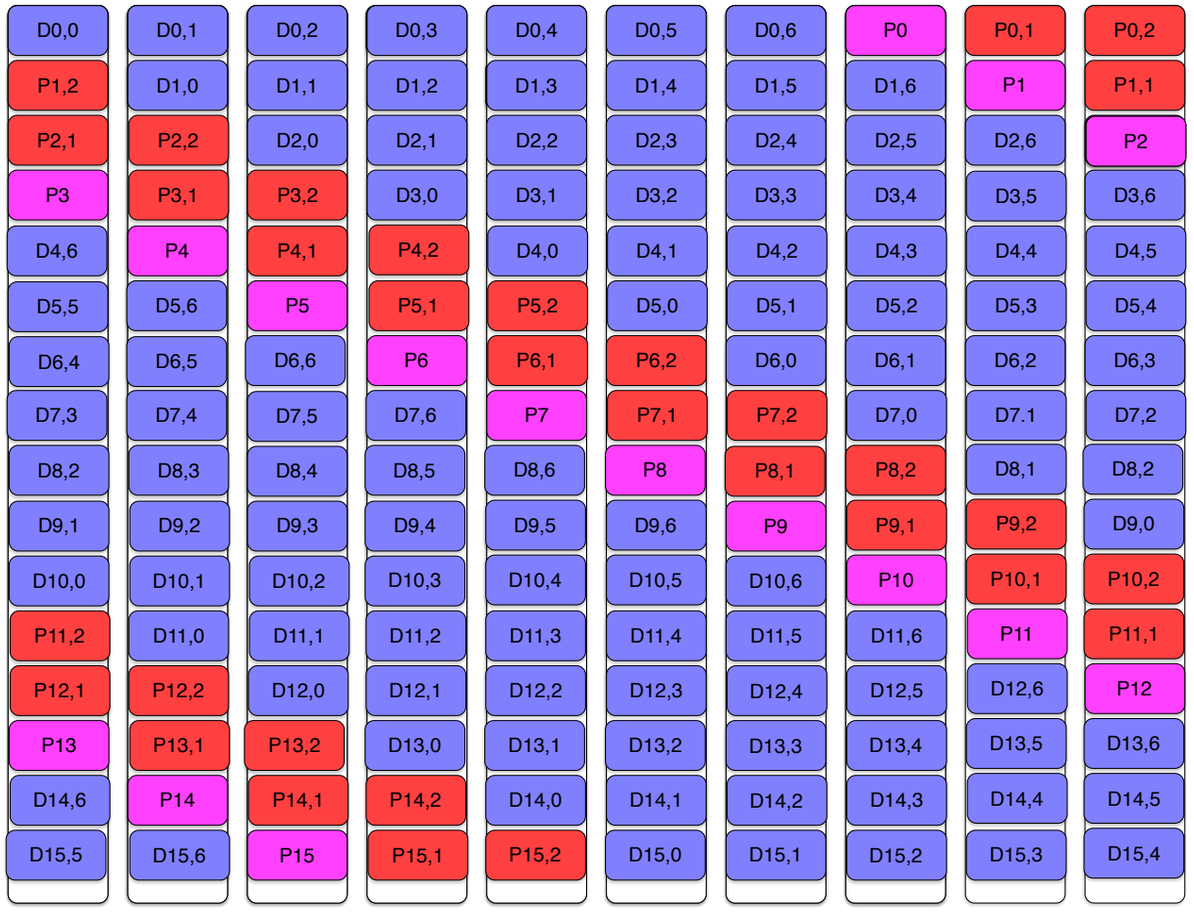
Erasure Correcting Coding

- Can add additional parity
  - Usually defined via Galois field operations
  - Or can use alternative layouts



Erasure Correcting Graphs

- RAID with three parity, one exclusive-or, the other more involved



Erasure Coding vs. Parity

- Erasure coding uses less storage for similar robustness
- Replication does not involve parity calculation
- Replication allows reads from multiple sources
- Both need consistency, but replication can have eventual consistency

Data Modeling

- To interpret data, we need to have structured data
- Semantic data models:
  - Semantic information add basic meaning to the data and the relationships between data
- Semantic Web:
  - Use hyperlinks and other semantic markup so that computers can understand and process information automatically