Almacenamiento de Estado Sólido

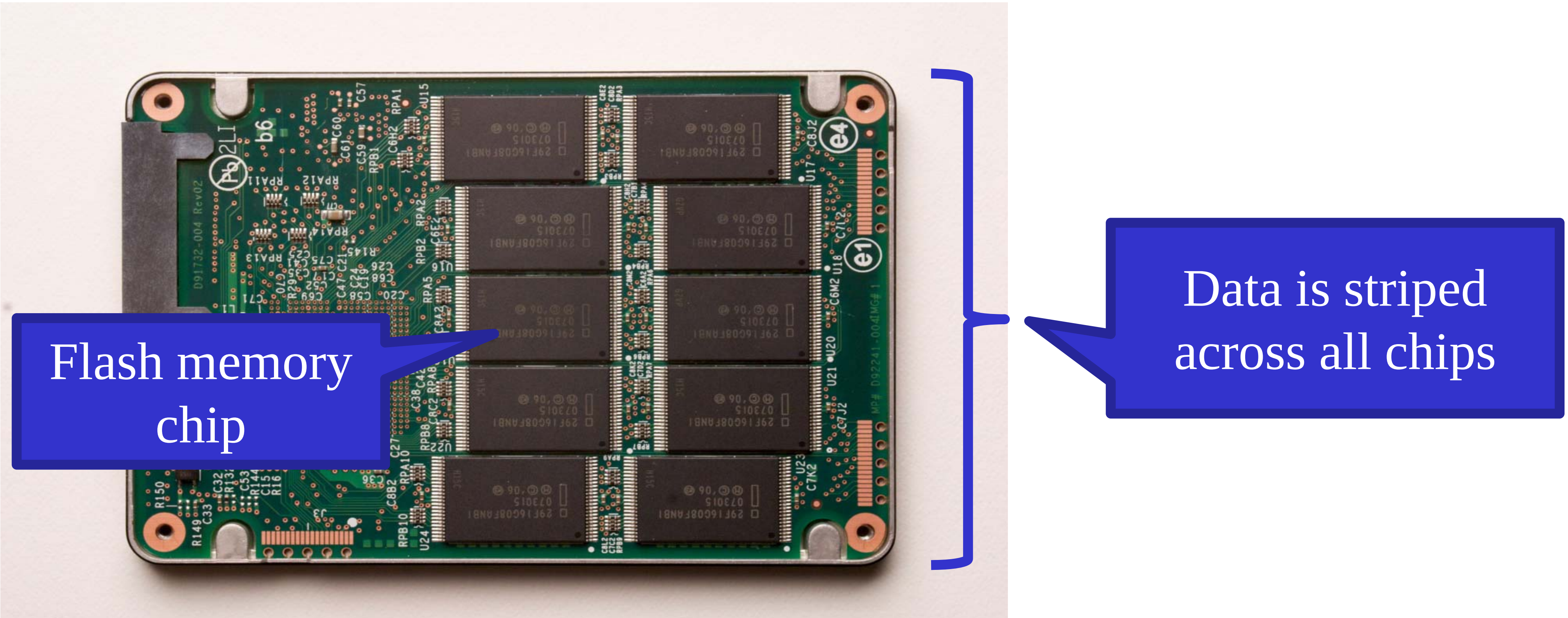
Visión General

Beyond Spinning Disks

- **Hard drives have been around since 1956**
  - The cheapest way to store large amounts of data
  - Sizes are still increasing rapidly
- **However, hard drives are typically the slowest component in most computers**
  - CPU and RAM operate at GHz
  - PCI-X and Ethernet are GB/s
- **Hard drives are not suitable for mobile devices**
  - Fragile mechanical components can break
  - The disk motor is extremely power hungry

Solid State Drives

- **NAND flash memory-based drives**
  - High voltage is able to change the configuration of a floating-gate transistor
  - State of the transistor interpreted as binary data



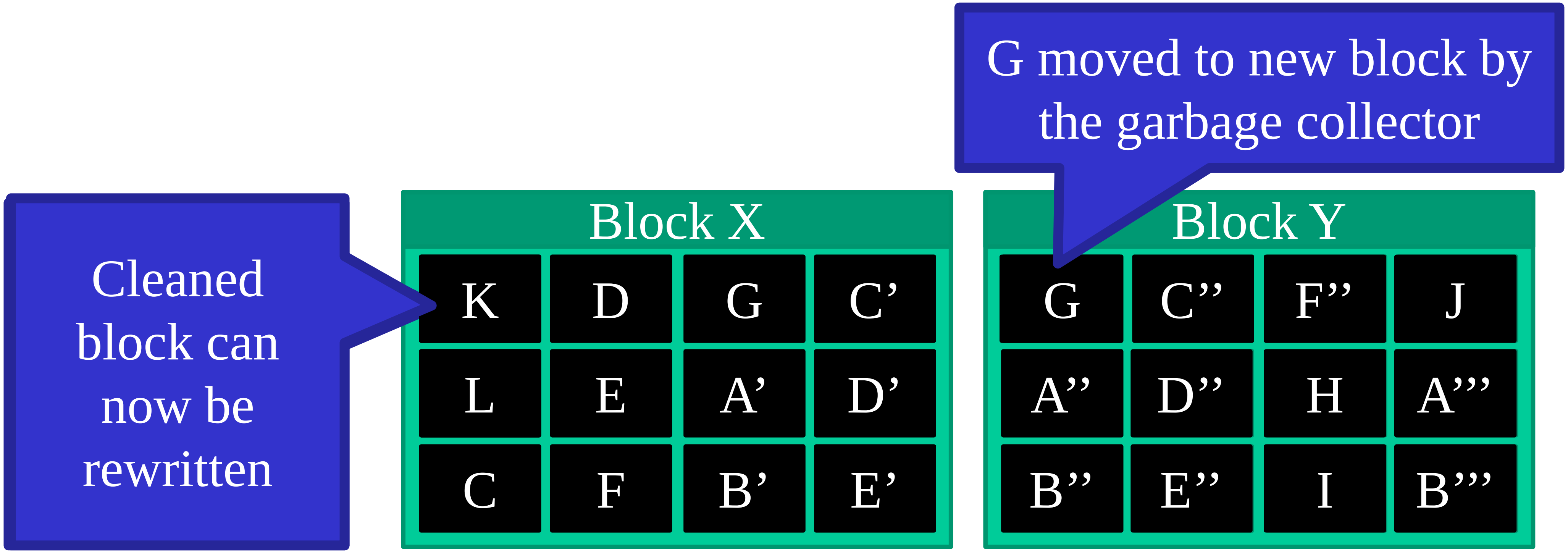
Advantages of SSDs ☺

- **More resilient against physical damage**
  - No sensitive read head or moving parts
  - Immune to changes in temperature
- **Greatly reduced power consumption**
  - No mechanical, moving parts
- **Much faster than hard drives**
  - >500 MB/s vs ~200 MB/s for hard drives
  - No penalty for random access
    - Each flash cell can be addressed directly
    - No need to rotate or seek
  - Extremely high throughput
    - Although each flash chip is slow, they are RAIDed

Challenges with Flash ☹

- **Flash memory is written in pages, but erased in blocks**
  - Pages: 4 – 16 KB, Blocks: 128 – 256 KB
  - Thus, flash memory can become fragmented
  - Leads to the **write amplification** problem
- **Flash memory can only be written a fixed number of times**
  - Typically 3000 – 5000 cycles for MLC
  - SSDs use **wear leveling** to evenly distribute writes across all flash cells

Write Amplification



- **Once all pages have been written, valid pages must be consolidated to free up space**

- **Write amplification: a write triggers <u>garbage collection/compaction</u>**
  - One or more blocks must be read, erased, and rewritten before the write can proceed

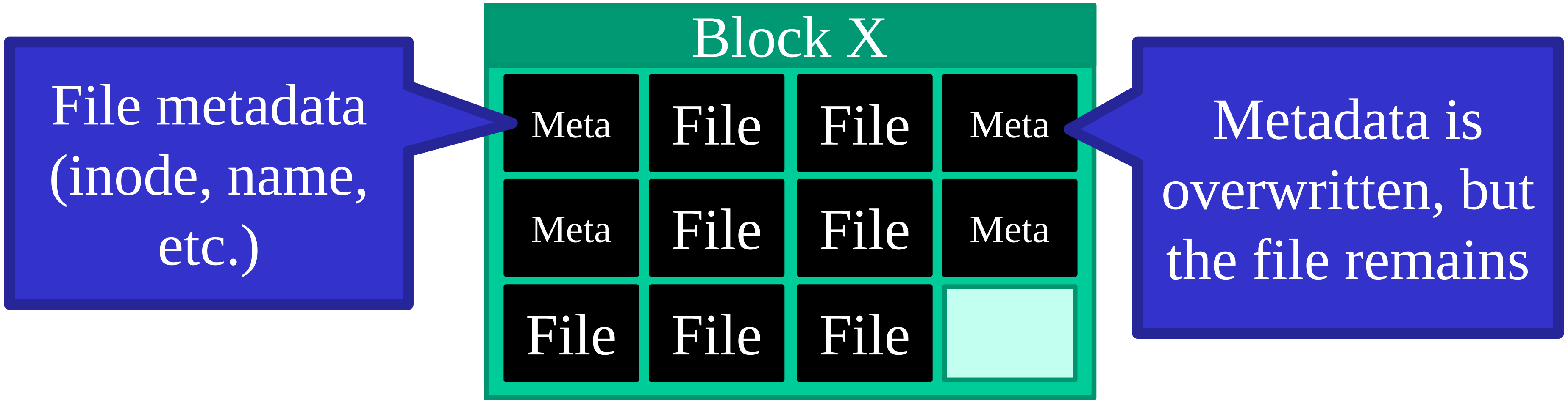
Garbage Collection

- **Garbage collection (GC) is vital for the performance of SSDs**
- **Older SSDs had fast writes up until all pages were written once**
  - Even if the drive has lots of "free space," each write is amplified, thus reducing performance
- **Many SSDs over-provision to help the GC**
  - 240 GB SSDs actually have 256 GB of memory
- **Modern SSDs implement background GC**
  - However, this doesn't always work correctly

The Ambiguity of Delete

- **Goal: the SSD wants to perform background GC**
  - But this assumes the SSD knows which pages are invalid
- **Problem: most file systems don’t actually delete data**
  - On Linux, the “delete” function is unlink()
  - Removes the file meta-data, but not the file itself

Delete Example

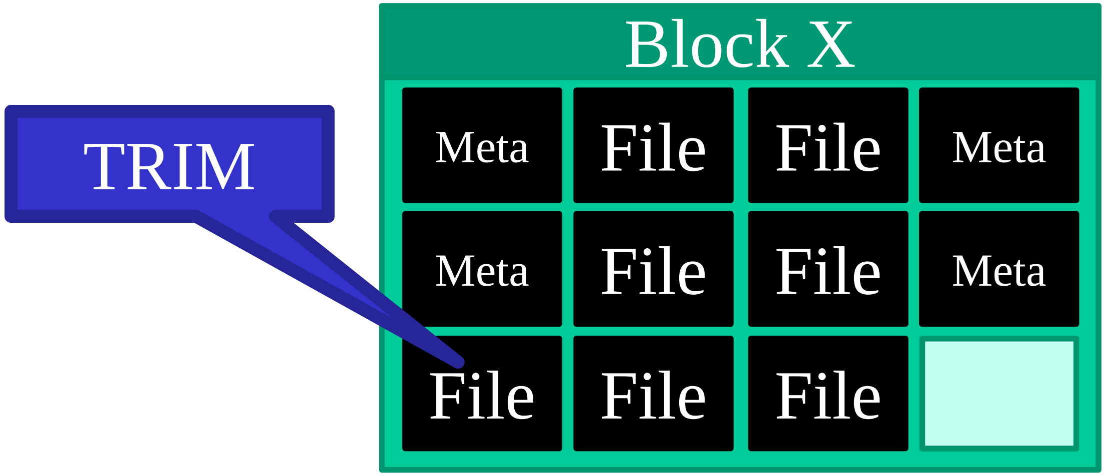


1. **File is written to SSD**
2. **File is deleted**
3. **The GC executes**
   - 9 pages look valid to the SSD
   - The OS knows only 2 pages are valid

- Lack of explicit delete means the GC wastes effort copying useless pages
- Hard drives are not GCed, so this was never a problem

TRIM

- **New SATA command TRIM (SCSI – UNMAP)**
  - Allows the OS to tell the SSD that specific LBAs are invalid, may be GCed



- OS support for TRIM
  - Win 7, OSX Snow Leopard, Linux 2.6.33, Android 4.3
- Must be supported by the SSD firmware

Almacenamiento de Estado Sólido

Limitación del número de ciclos de borrado

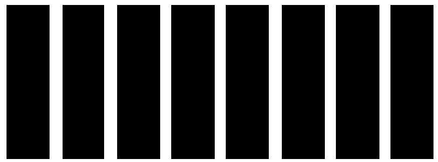
Block Usage

- Need to maintain supply of empty blocks to add to write allocation pool.

- Cleaning involves moving valid pages from one block to another block.

- Flash blocks have limited lifetime
  - Fixed number of erasures

- Greedy cleaning
  - Choose blocks with best cleaning efficiency



- Goal: use all blocks uniformly

(Source – SSD USENIX, 08)

Wear Leveling (Nivelación del desgaste)

- **Recall: each flash cell wears out after several thousand writes**
- **SSDs use wear leveling to spread writes across all cells**
  - Typical consumer SSDs should last ~5 years

Wear-leveling

- **Write/Erase cycle of NAND is limited to 100K for SLC and 10K for MLC.**

- **Reducing Wear Level:**

    - Write data to be evenly distributed over the entire storage.

    - Count # of Write/Erase cycles of each NAND block.

    - Based on the Write/Erase count, NAND controller re-map the logical address to the different physical address.

    - Wear-leveling is done by the NAND controller (FTL), not by the host system.

(Source - Ken Takeuchi INRET, 08)

Static Vs. Dynamic wear-leveling

Static data

Data that does not change such as system data (OS, application SW).

Dynamic data

Data that are rewritten often such as user data.

Dynamic wear-leveling

Wear-level only over empty and dynamic data.

Static wear-leveling

Wear-level over all data including static data.

(Source - Ken Takeuchi INRET, 08)

Wear Leve

If the GC runs now, page G must be copied

Dynamic Wear Leveling

Wait as long as possible before garbage collecting

Block X			
K	D	G	C’
L	E	A’	D’
C	F	B’	E’

Block Y			
F’	C’’	F’’	G’
A’’	D’’	H	A’’’
B’’	E’’	I	B’’’

Static Wear Leveling

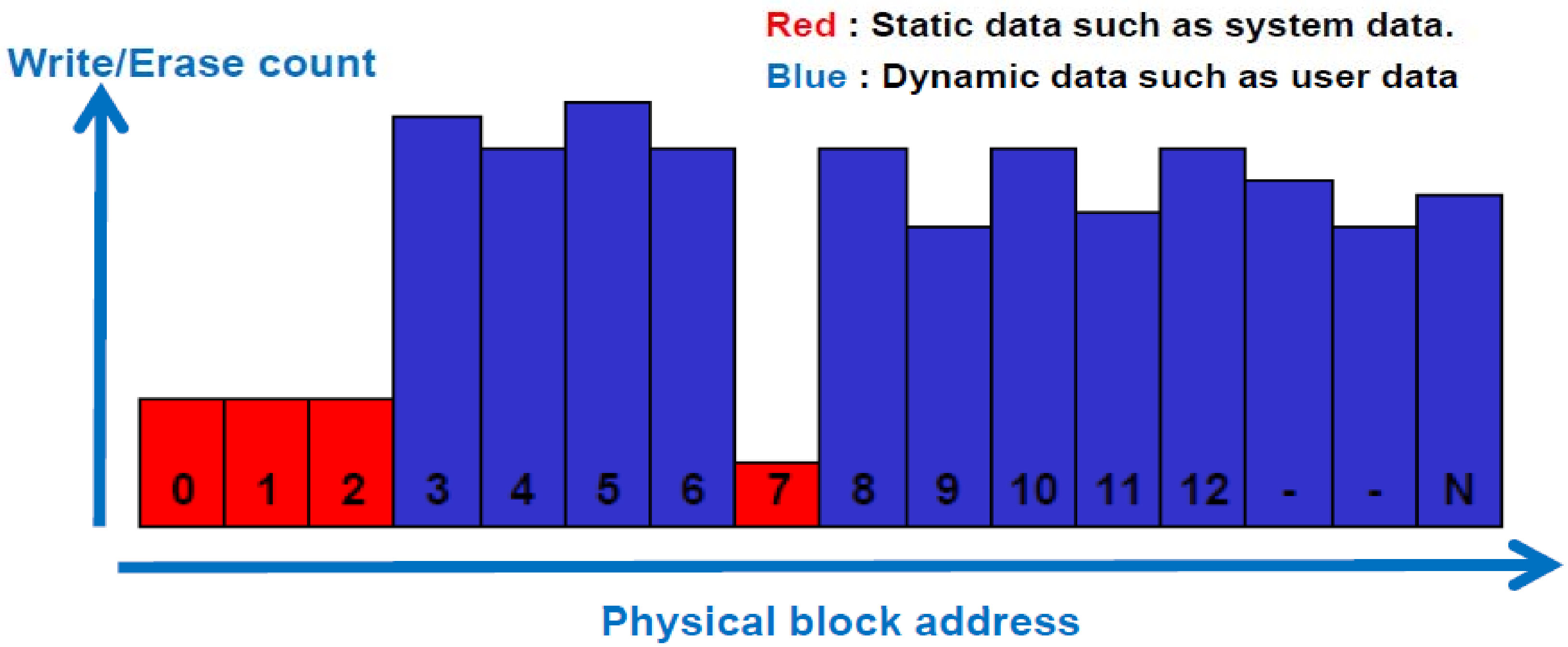
Blocks with long lived data receive less wear

Block X			
M*	D	G	J
N*	E	H	K
O*	F	I	L

Block Y			
A	D	G	J
B	E	H	K
C	F	I	L

SSD controller periodically swap long lived data to different blocks

Dynamic wear-leveling

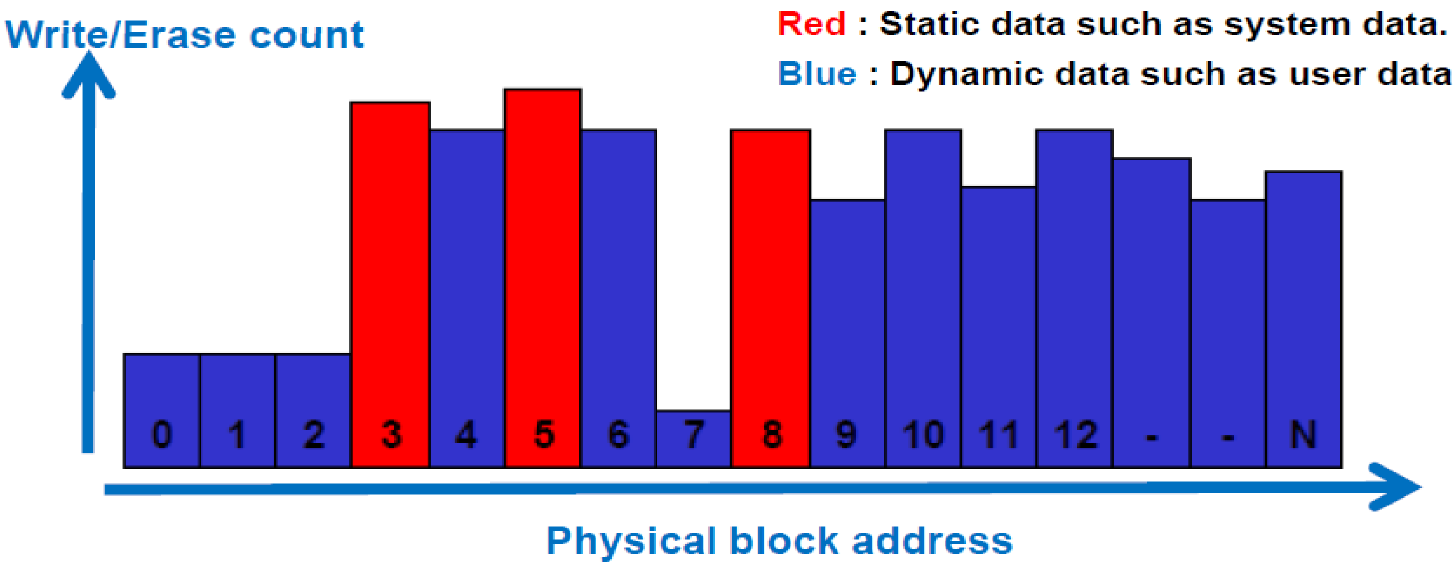


- Block with static data is NOT used for wear-leveling.
- Write and erase concentrate on the dynamic data block.

N.Balan, MEMCON2007.
SiliconSystems, SSWP02

(Source - Ken Takeuchi INRET, 08)

Static wear-leveling



- Wear-level more effectively than dynamic wear-leveling.
- Search for the least used physical block and write the data to the location. If that location Is empty, the write occurs normally.
- Contains static data, the static data moves to a heavily used block and then the new data is written.

N.Balan, MEMCON2007.
SiliconSystems, SSWP02

(Source - Ken Takeuchi INRET, 08)

Summary

SSD Advantage :

- Low power consumption
- High mechanical reliability, no spinning parts
- Fast Read performance
- No data loss as failure occurs on write (another cell can be used for write), rather than read on HDD

SSD Disadvantage :

- High cost
- Low capacity compared to HDD
- Slow random write (due to slow block erase)
- Limited write/erase cycles

SSD Controllers

- SSDs are extremely complicated internally

- **All operations handled by the SSD controller**
  - Maps LBAs to physical pages
  - Keeps track of free pages, controls the GC
  - May implement background GC
  - Performs wear leveling via data rotation

- **Controller performance is crucial for overall SSD performance**

SSD Controller

HIL – Support host interconnect (USB/PCI/SATA/PCIe).

Buffer Manager – Holds pending and satisfied request along primary data path.

Flash Demux/Mux – emits command and handles transport of data along serial connection to flash.

Processing engine – manages request flow and mapping from Logic block address to physcial flash location.

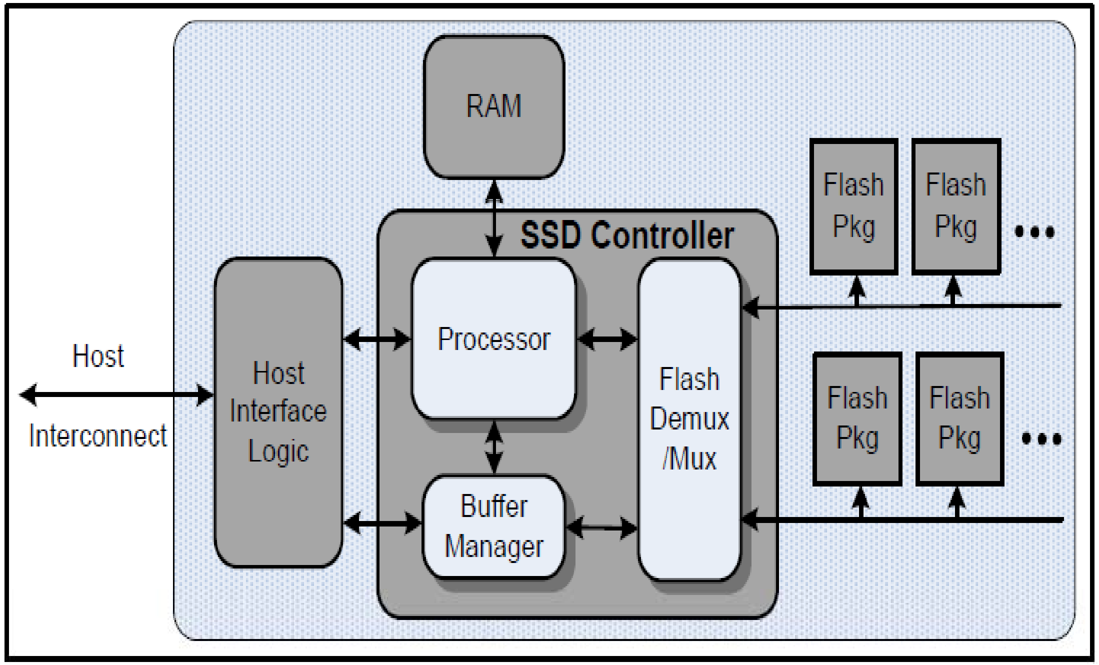


Figure 3: SSD Logic Components

(Source – SSD USENIX, 08)

Flash Technology

- **Fujio Masuoka invents flash memory in 1984 while working for Toshiba.**
  - Capable of being erased and re-programmed multiple times, flash memory quickly gained a loyal following in the computer memory industry.
  - Toshiba's failure to reward his work, and Masuoka quit to become a professor at Tohoku University.
  - Bucking Japan's culture of company loyalty, he sued his former employer demanding compensation, settling in 2006 for a one-time payment of ¥87m ($758,000).

http://www.computerhistory.org/timeline/1984/#169ebbe2ad45559efbc6eb357202d1e7 NAND flash memory-based drives

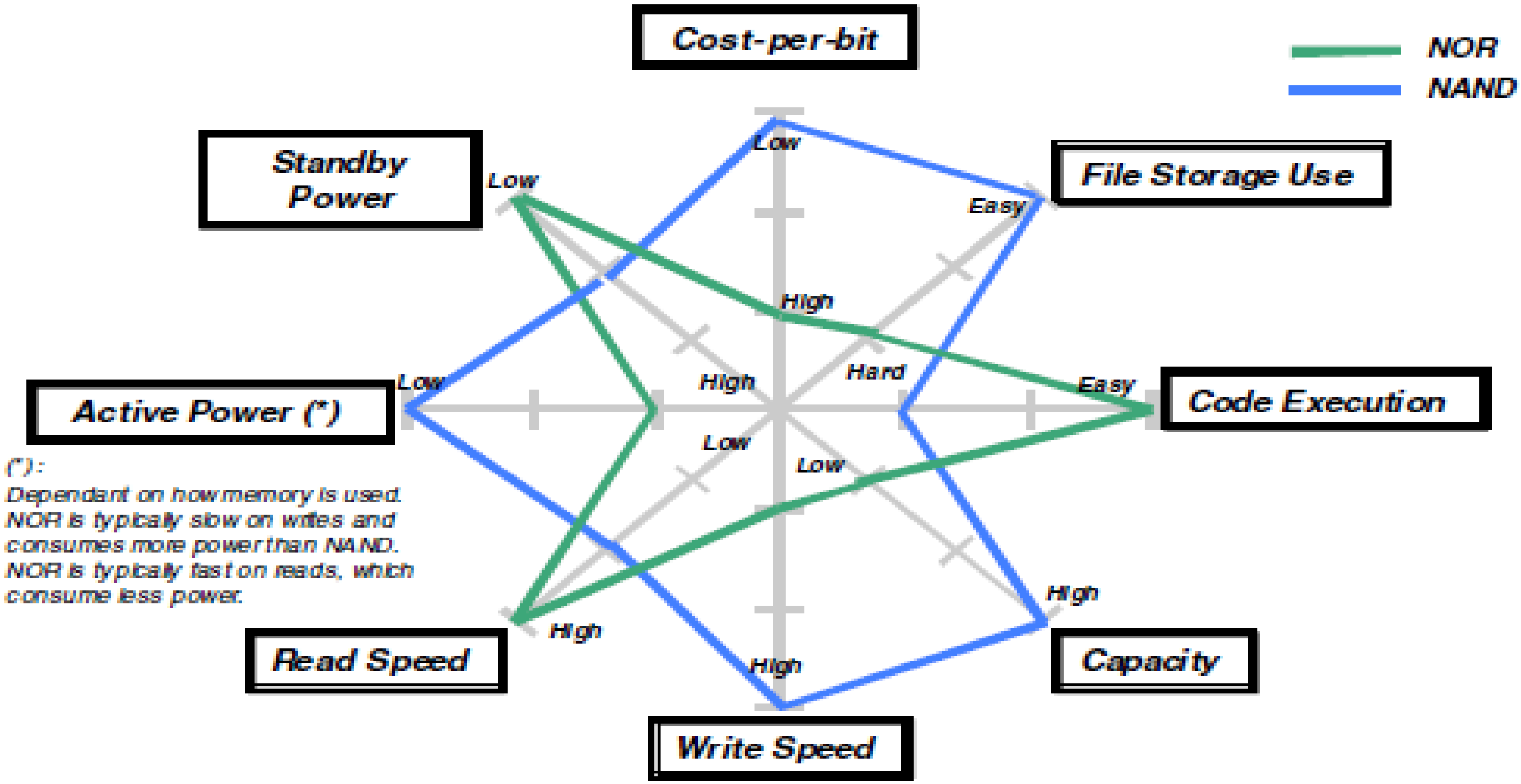
Flash Technology overview

- **Two major forms NAND flash and NOR flash**
  - NOR Flash has typically been used for code storage and direct execution in portable electronics devices, such as cellular phones and PDAs.
  - NAND Flash, which was designed with a very small cell size to enable a low cost-per-bit of stored data, has been used primarily as a high-density data storage medium for consumer devices such as digital still cameras and USB solid-state disk drives.
- **Toshiba was a principal innovator of both NOR type and NAND-type Flash technology in the 1980's.**

(Source: Toshiba)

NAND vs. NOR Flash Memory

Fig. 1 Comparison of NOR and NAND Flash



(Source Toshiba)

When should one choose NAND over NOR ?

For a system that needs to boot out of Flash, execute code from the Flash, or if read latency is an issue, NOR Flash may be the answer.

For storage applications, NAND Flash's higher density, and high programming and erase speeds make it the best choice.

Power is another important concern for many applications. For any write-intensive applications, NAND Flash will consume significantly less power.

What if a system, such as a camera phone, has a requirement both for code execution and high capacity data storage?

(Source Toshiba)

Flavors of NAND Flash Memory

Multi-Level Cell (MLC)

- **Multiple bits per flash cell**
  - For two-level: 00, 01, 10, 11
  - 2, 3, and 4-bit MLC is available
- **Higher capacity and cheaper than SLC flash**
- **Lower throughput due to the need for error correction**
- **3000 – 5000 write cycles**
- **Consumes more power**

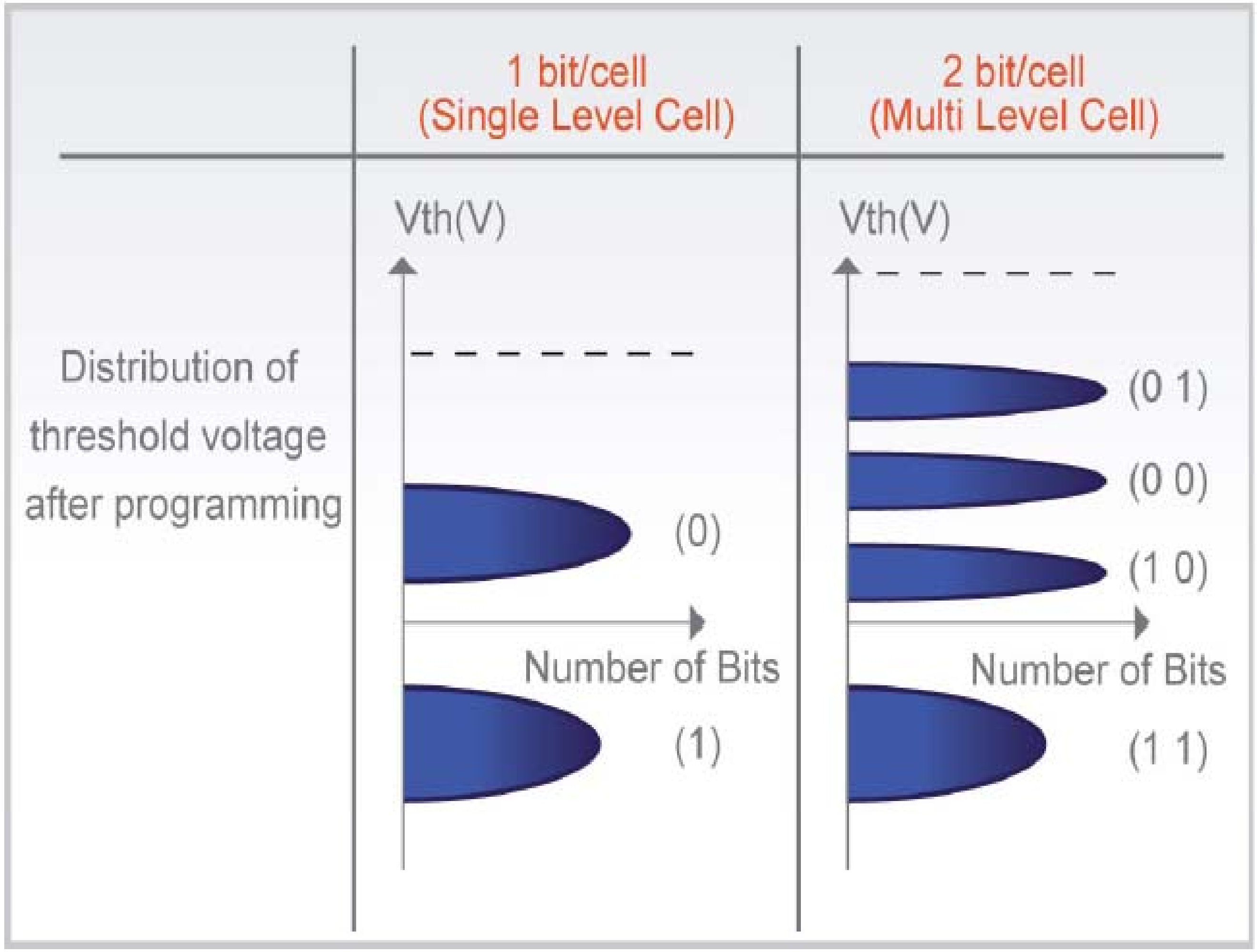
Consumer-grade drives

Single-Level Cell (SLC)

- **One bit per flash cell**
  - 0 or 1
- **Lower capacity and more expensive than MLC flash**
- **Higher throughput than MLC**
- **10000 – 100000 write cycles**

Expensive, enterprise drives

NAND SLC vs. MLC Technology



Source: Toshiba. 2008

(Source Toshiba)

HDD vs. SDD

Random access

	Read	Write	Erase
NAND (SLC)	25us	300us	1ms
NAND (MLC)	50us	800us	1ms
HDD	3ms	3ms	N.A.

Erase are hidden by operating the erase during the idle period.

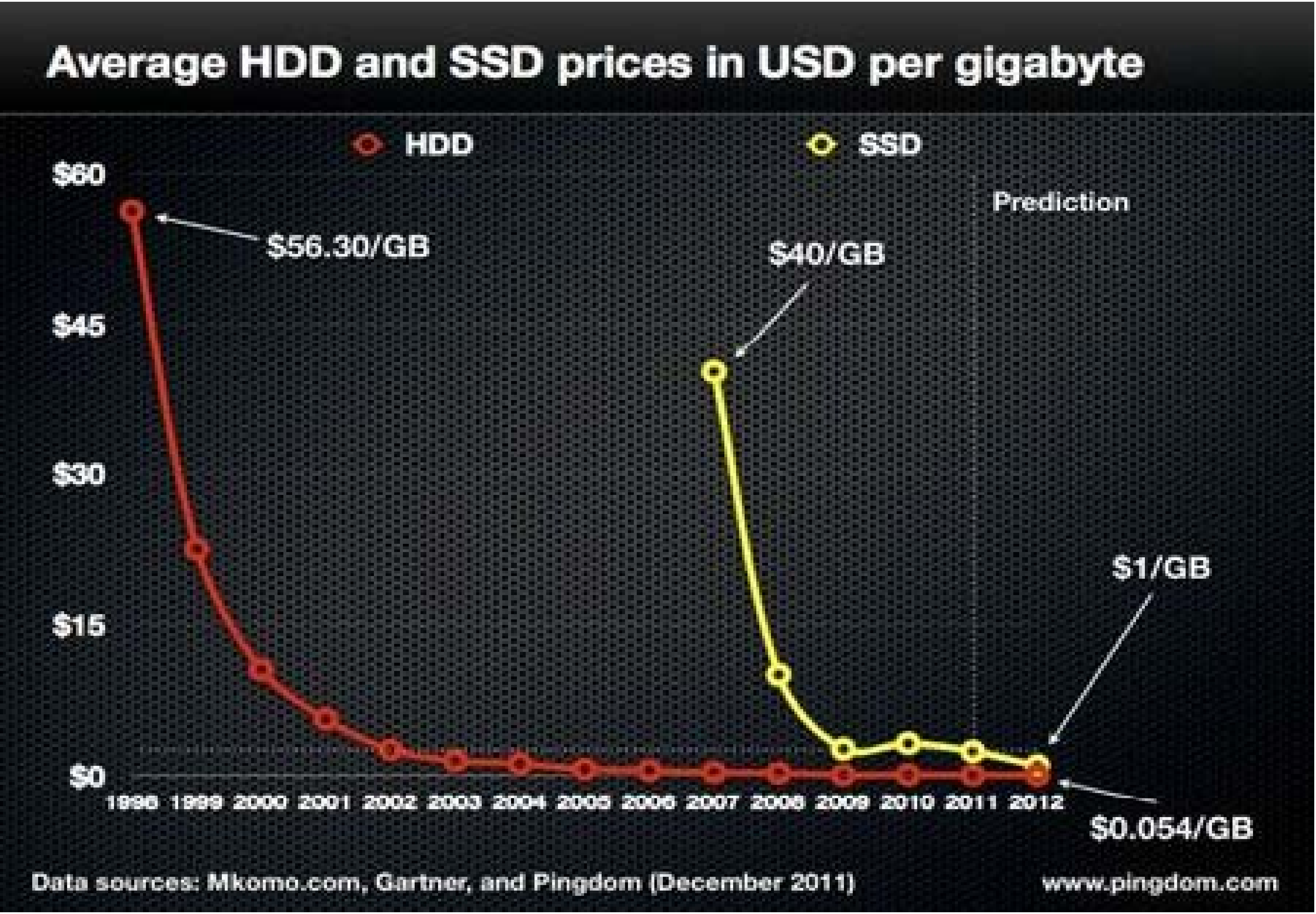
Sequential access

	NAND : Single chip operation		NAND : 4 chip interleaving	
	Read	Write	Read	Write
NAND (SLC)	25MB/sec	20MB/sec	100MB/sec	80MB/sec
NAND (MLC)	20MB/sec	10MB/sec	80MB/sec	40MB/sec
HDD	80MB/sec	80MB/sec	-	-

(Source - Ken Takeuchi INRET, 08)

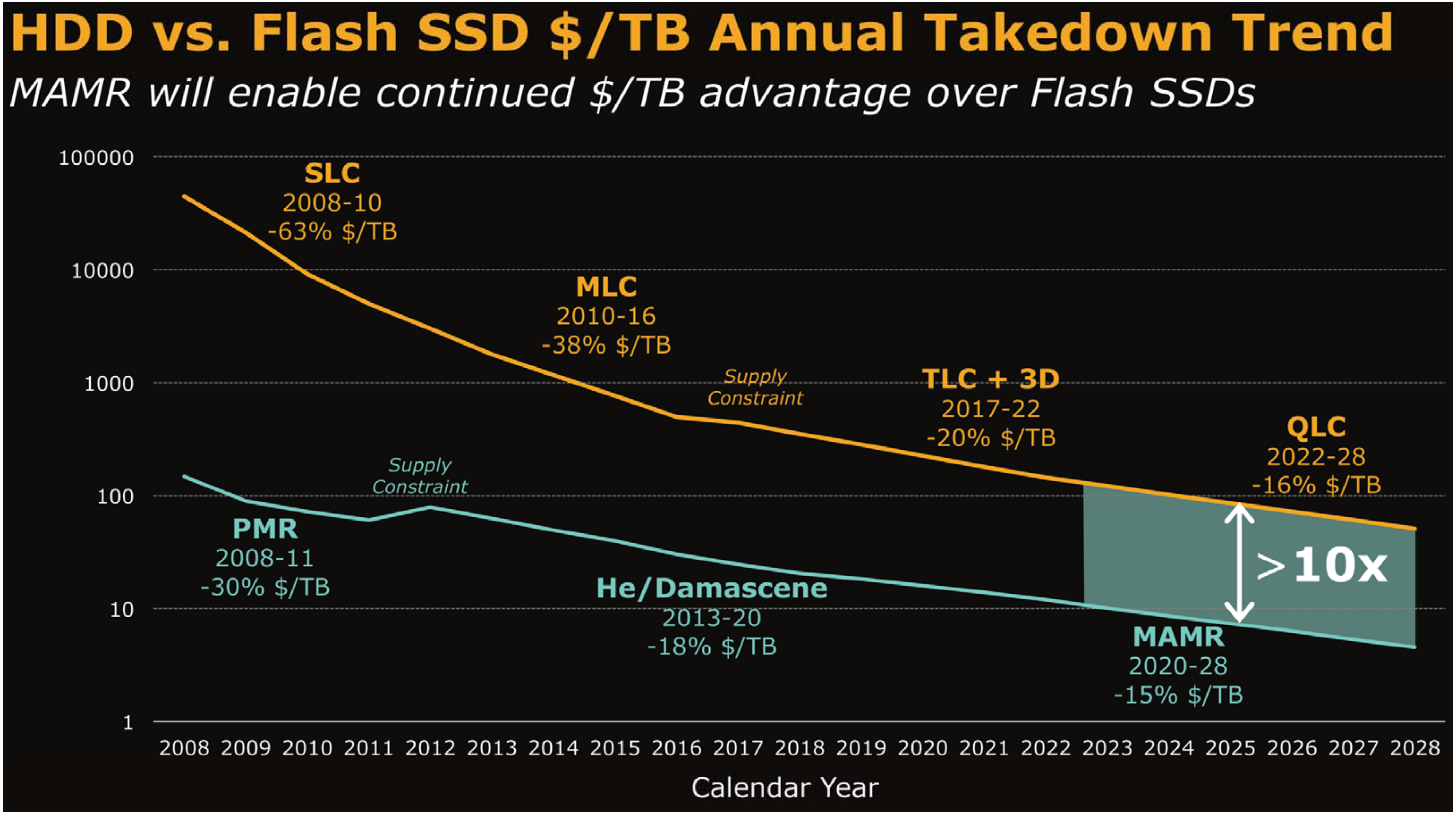
Flash NAND alternativas

- **SLC (Capa Simple)**
  - Cada Celda almacena 1 bit de información
- **MLC (Multi Capa)**
  - 2 bits por celda
  - 4 estados posibles 00, 01, 10, 11.
  - Son mas lentas por que tenemos que distinguir más estados
- **TLC (Triple Capa)**
  - tres bits por celdas.
  - 8 estados, 000, 001, 010, 011, 100, 101, 110, 111
- **3D NAND (Apiladas Verticalmente)**
  - Celdas Apiladas Verticalmente que llegan hasta las 32 Capas
- **QLC (Quadrupple Level Cell)**
  - 4 bits por célula de datos
  - Capacidades de hasta 128 TB

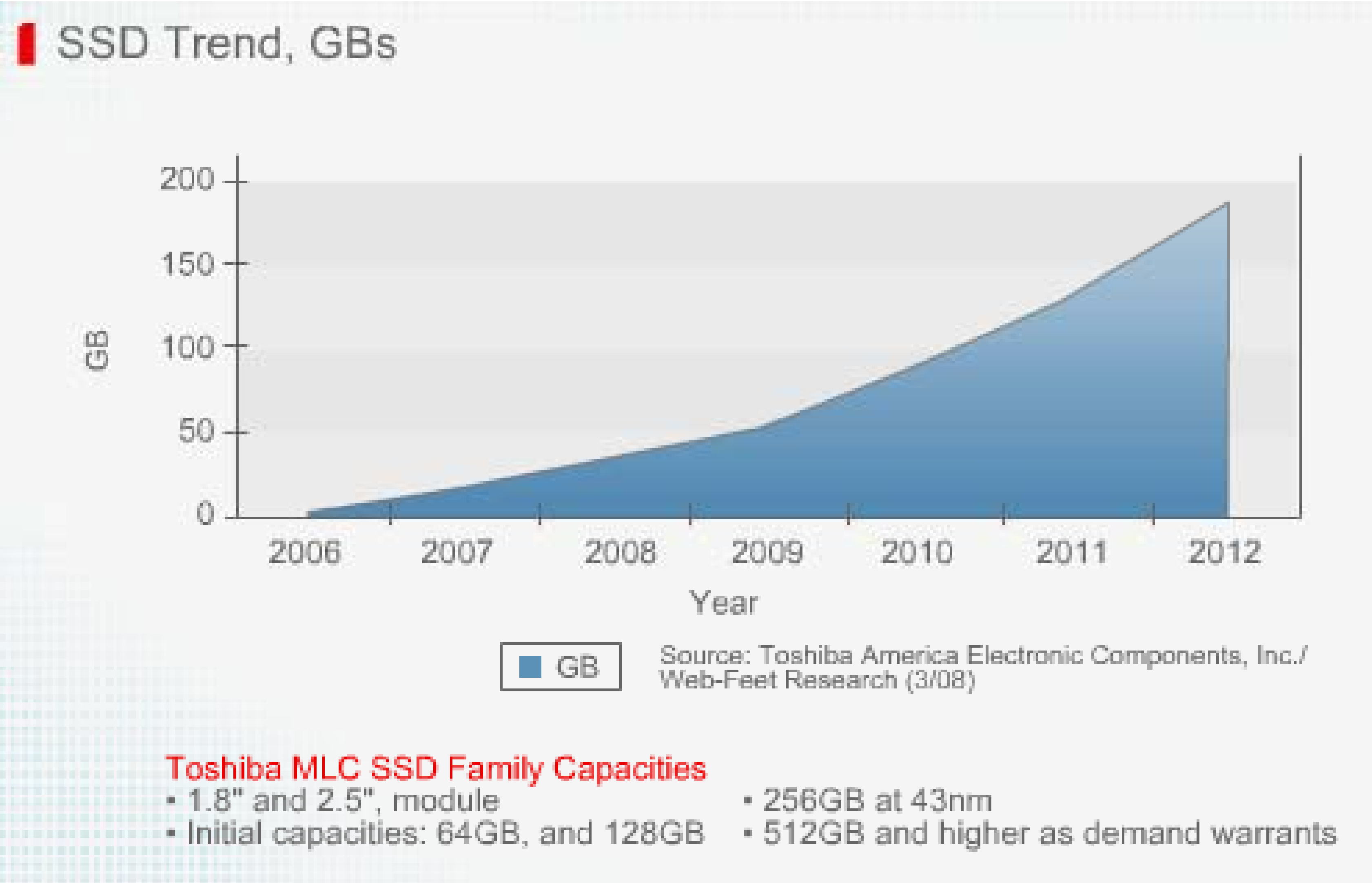


Average HDD and SSD prices in USD per gigabyte
HDD
SSD
$60
$45
$30
$15
$0
Prediction
$56.30/GB
$40/GB
$1/GB
$0.054/GB
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Data sources: Mkomo.com, Gartner, and Pingdom (December 2011)
www.pingdom.com

Evolución de Precios HDD vs HDD
Introducción de nuevas Tecnologías

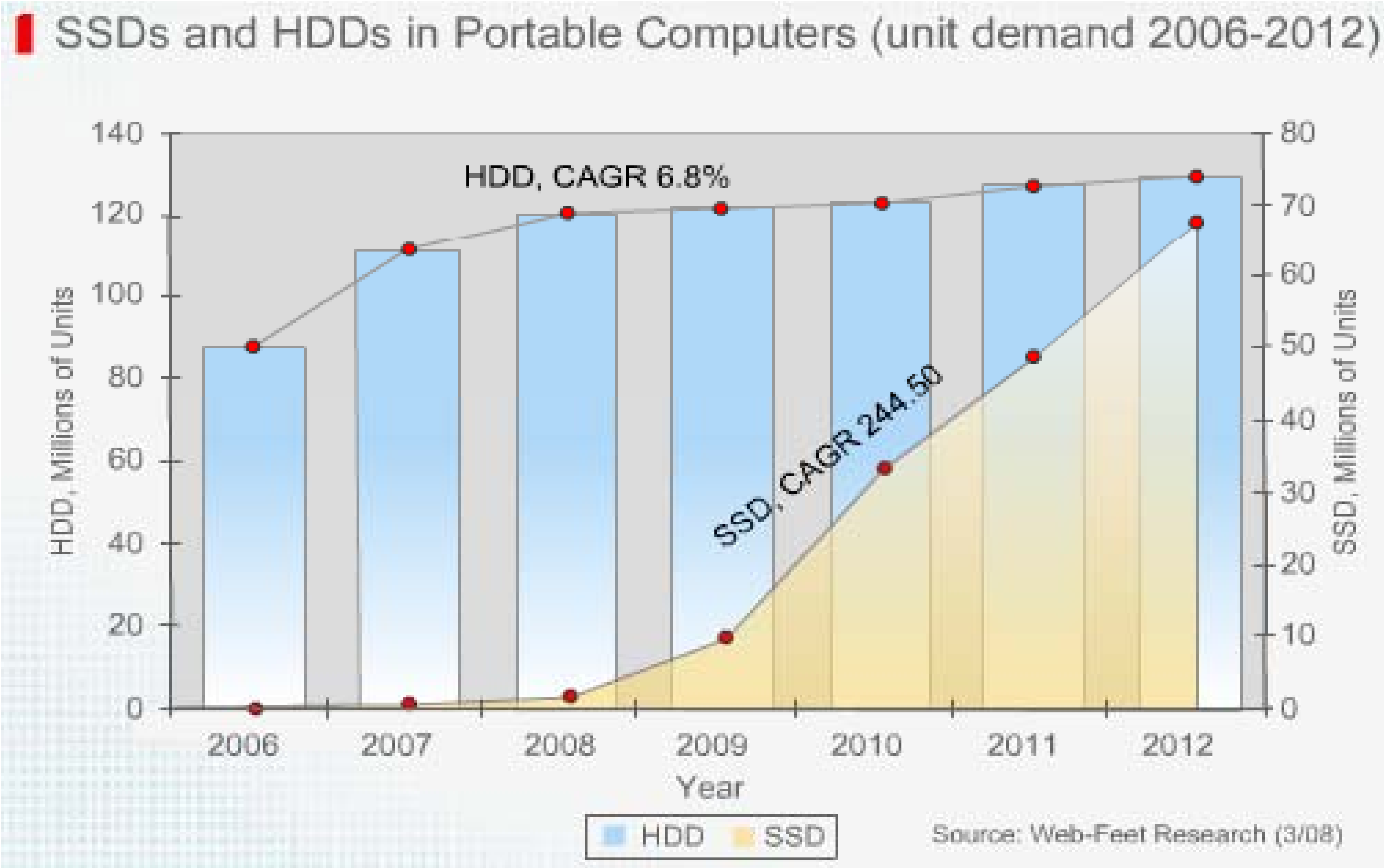


SSD market Trends



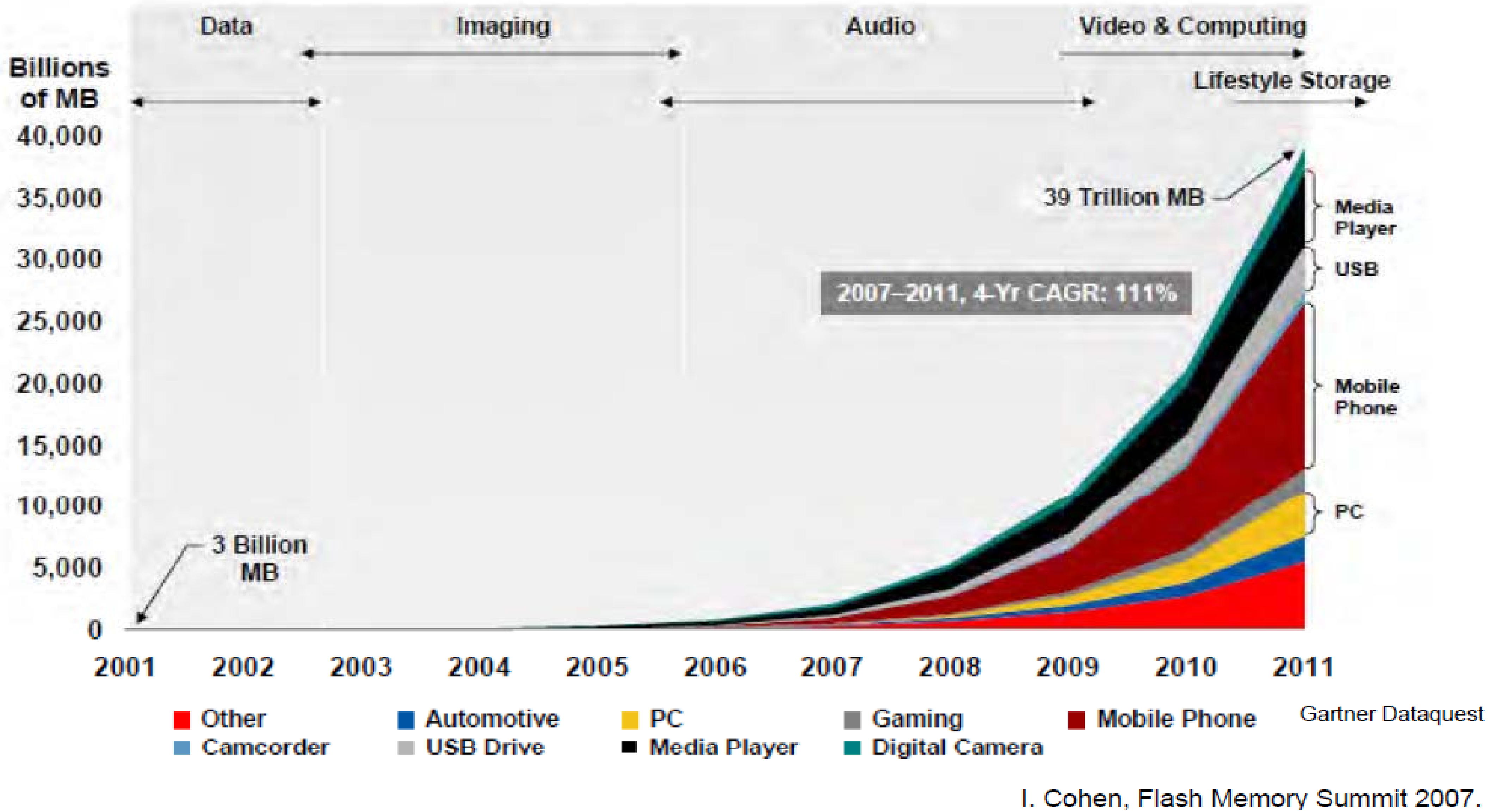
(Source -Toshiba, 08)

SSD market Trends



(Source -Toshiba, 08)

SSD market Trends



(Source - Ken Takeuchi INRET, 08)

NAND Flash Internals

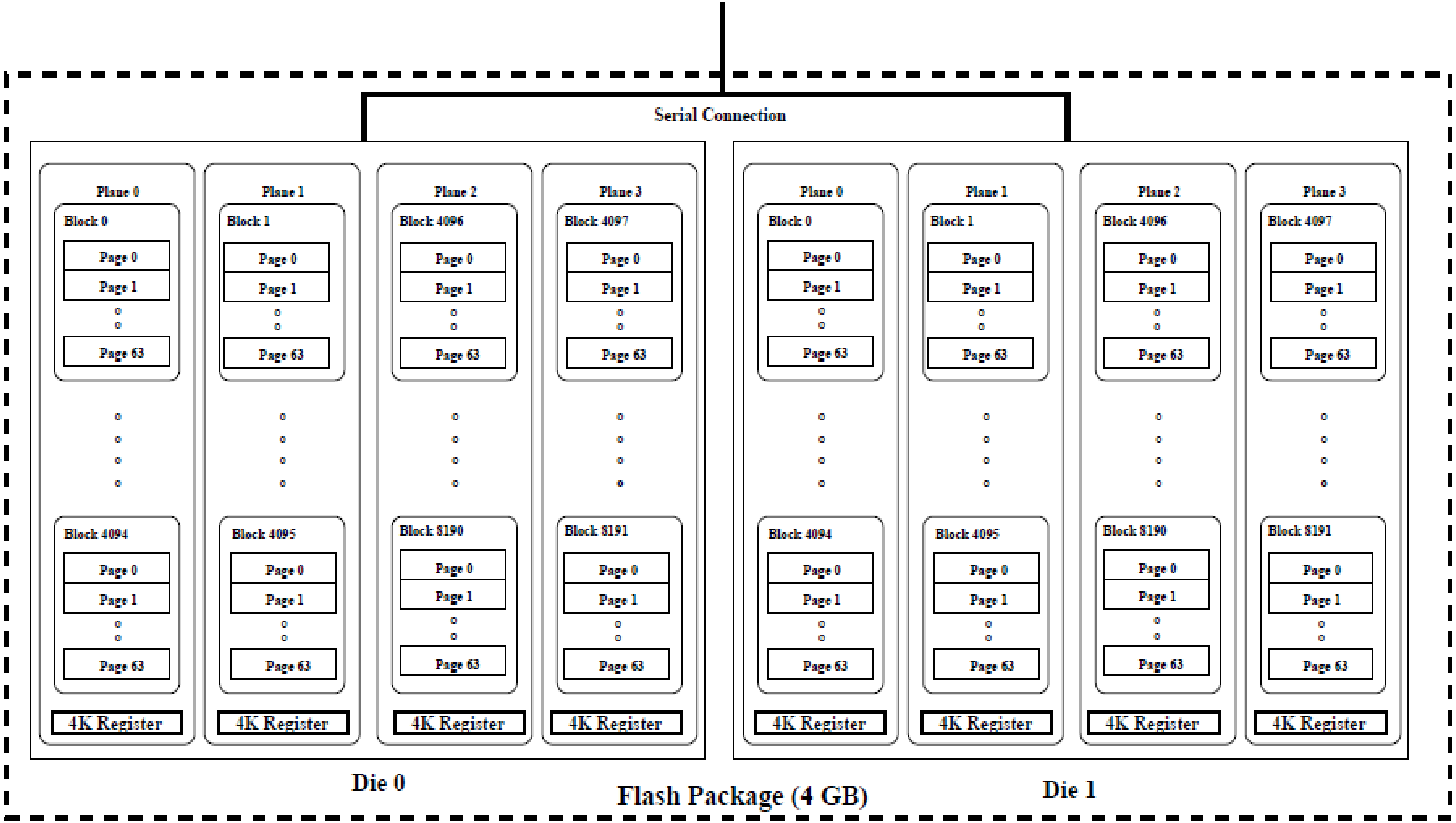


Figure 1: Samsung 4GB flash internals

(Source – SSD USENIX, 08)

NAND Flash Internals – Key points

- 4GB package consisting of 2GB dies, share 8-bit serial I/O bus and common control signals.

- Two dies have separate chip enable and ready/busy signals – One of them can accept commands while the other is carrying out another operation.

- Two plane-commands can be executed on either plane 0 & 1 or 2 & 3.

(Source – SSD USENIX, 08)

NAND Flash Internals – Key points

- **Each page includes 128 byte region to store meta data (Identification and error detection information).**
- **Data read/write at the granularity of flash pages, thru 4KB data register.**
- **Erase at block level.**
- **Each block can be erased only finite number of time 100K for SLC.**

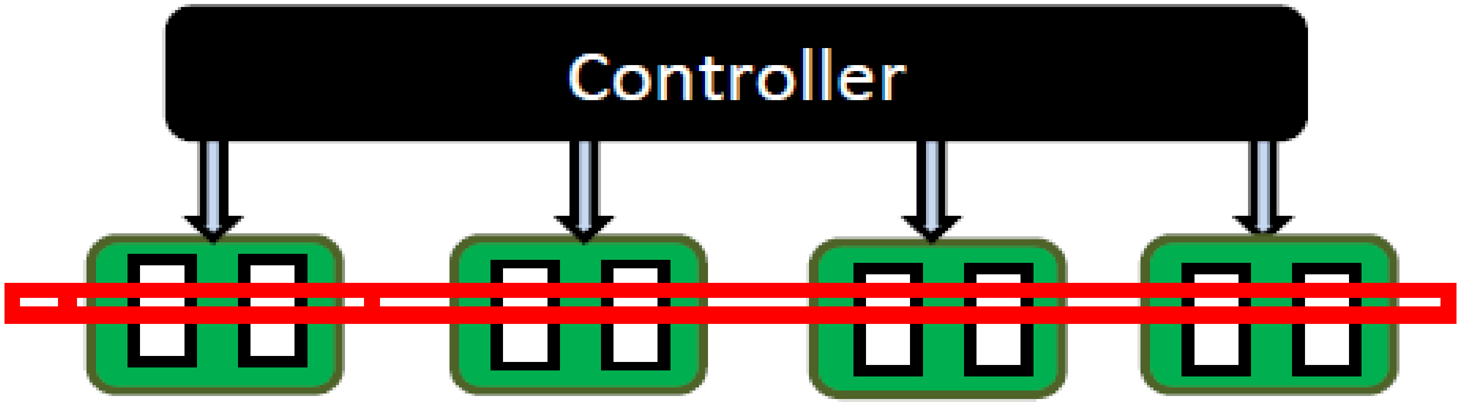
(Source – SSD USENIX, 08)

Limited Serial Bandwidth

Exploiting parallelism: Interleaving

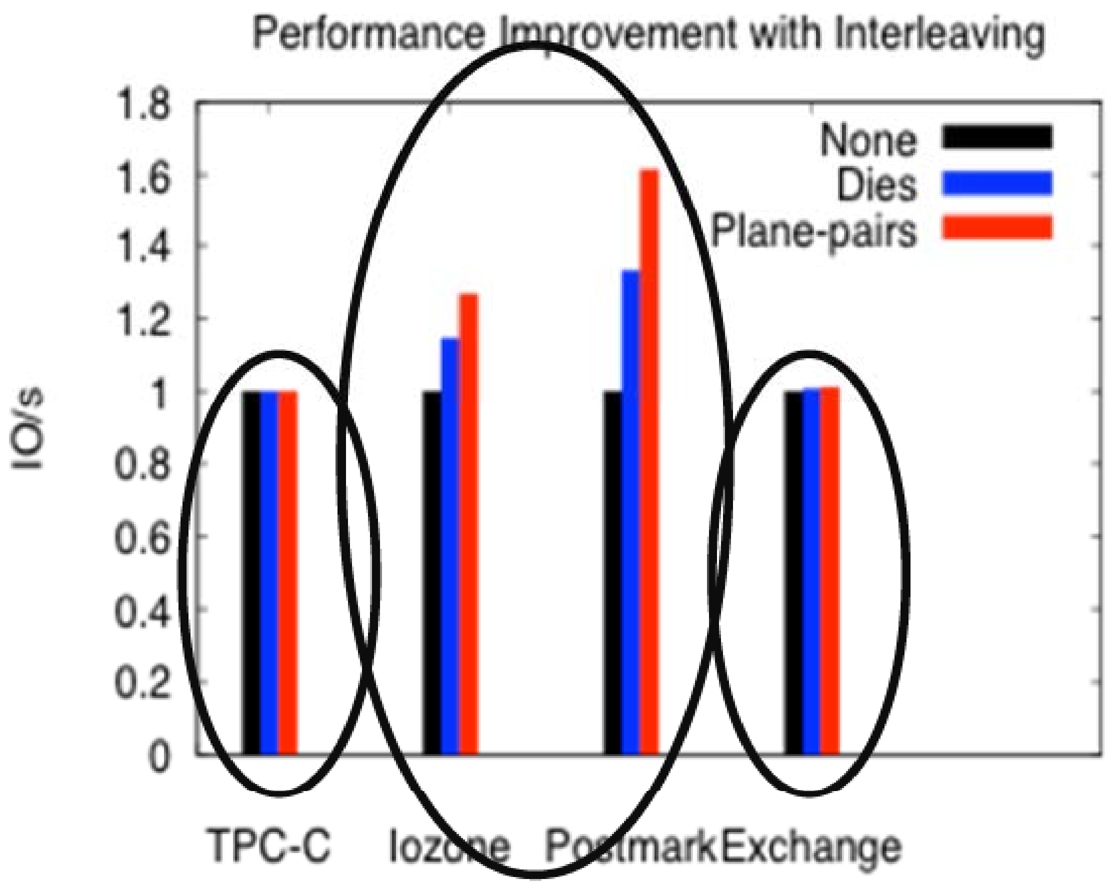
Inherent parallelism : multiple packages, dies, planes

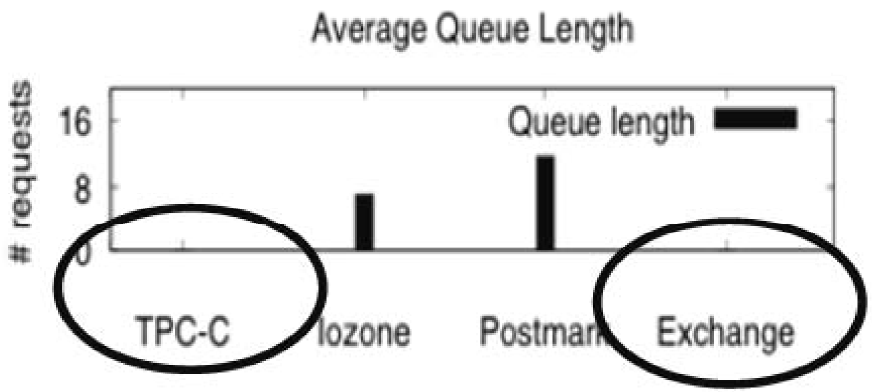
Stripping across and within packages



(Source – SSD USENIX, 08)

Interleaving Within Package

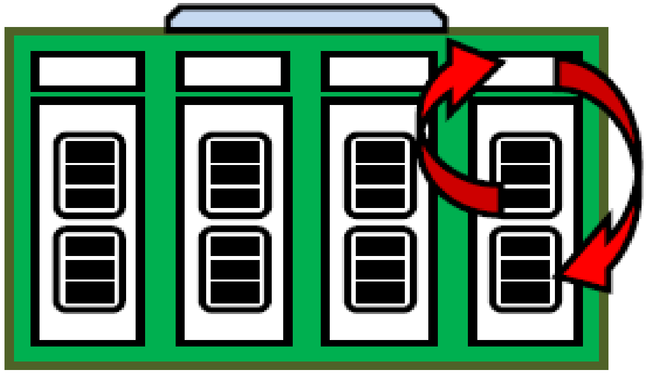




(Source – SSD USENIX, 08)

Copy-back

Copy-back : copy pages within a flash package Cleaning and wear-leveling



Workload	Improvement with Copy-back	Efficiency
TPC-C	40%	70%
Iozone	0%	100%
Postmark	0%	100%

(Source – SSD USENIX, 08)

Concluding remark

"There have been few times in the history of computing when a new technology becomes pivotal to completely changing the PC platform and user experience, Solid State Drive have this capability."

- Gordon Moore.

Almacenamiento Magnético en Cintas

Visión General

Repaso:
Historia del almacenamiento magnético

- ***Sistemas magnéticos***

- La historia del almacenamiento magnético, se remonta a 1949, cuando un grupo ingenieros y científicos de IBM, empezaron a desarrollar un nuevo dispositivo de almacenamiento, que revolucionaria la industria.

- En 1952, IBM anunció su primer dispositivo de almacenamiento magnético, la *IBM 726* que fue la primera cinta magnética, junto con la *IBM 701*, que fue el primer computador para aplicaciones científicas.

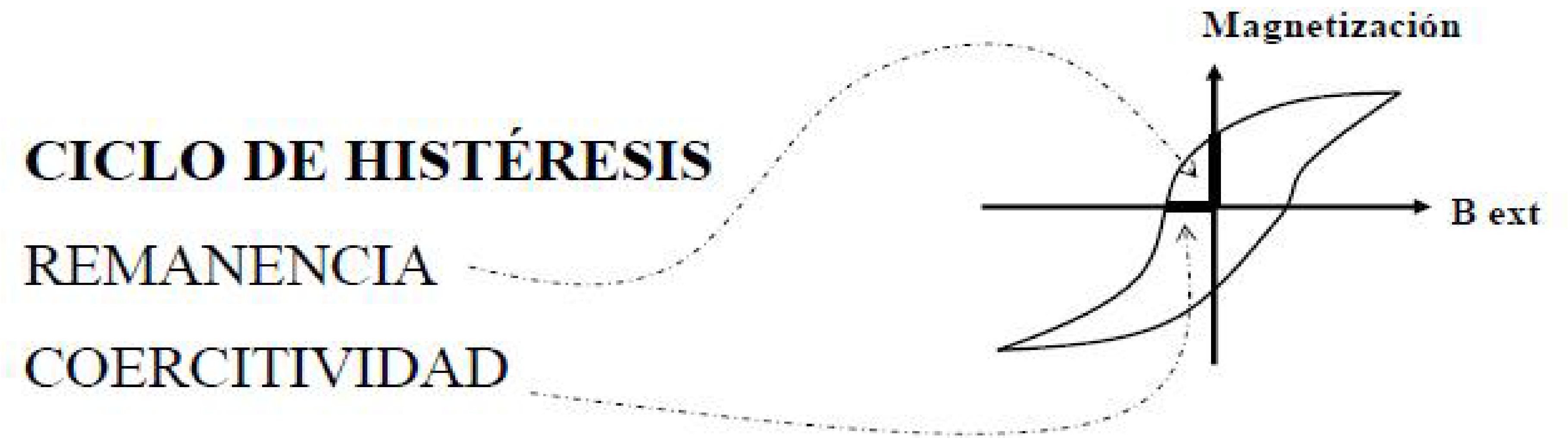
Repaso:
Historia del almacenamiento magnético

IBM 726

IBM 701

Repaso:
Uso de campos magnéticos para almacenar datos

- **Si representamos el valor del campo magnético en función del valor de la corriente que circula, tenemos el llamado *ciclo de histéresis***



MÓDULO 4
Principios básicos

- **Características importantes:**
  - Capacidad
  - Coste
  - Densidad de información
  - Velocidad de Transferencia
  - Tiempo de acceso
  - Otros: Fiabilidad, durabilidad…

- **Soporte Magnético**

- **Cabezal de lectura y de escritura**

Densidad, capacidad y coste

- **Es mejor cuanto:**
  - Mayor Capacidad (cantidad de información máxima que podemos almacenar)
  - Menor Coste (estático o fabricación + dinámico o explotación)
  - Mas alta sea la relación Capacidad/Coste

- **Densidad de Información:**
  - Cantidad de información por unidad de volumen (y consecuentemente área, longitud)

- **Mayor densidad de información suele implicar:**
  - Mayor Capacidad
  - Menor Coste
  - Mejor relación Capacidad/Coste

- **➔ Siempre se buscará aumentar la densidad de información** **(manteniendo el resto de especificaciones de fiabilidad y durabilidad de la información almacenada)**

Densidad, velocidad de transferencia y tiempo de acceso

- **Es mejor cuanto:**
  - Mayor Velocidad de transferencia (cantidad de información por unidad de tiempo que podemos leer y/o escribir)
  - Menor Tiempo de Acceso (tiempo que se tarda en alcanzar la posición de elemento buscado y estar en condiciones de leer)

- **Mayor Densidad de Información implica mayor cercanía espacial de los bits y por lo tanto … ➔**
  - Mayor velocidad de transferencia potencial
  - Menor tiempo de acceso
  - Menor coste energético

- **➔ <u>Siempre se buscará aumentar la densidad de información</u> (manteniendo el resto de especificaciones de fiabilidad y durabilidad de la información almacenada)**

Densidad de información en el medio magnético

- **Depende determinantemente de la tecnología del soporte magnético y las cabezas de lectura y de escritura (estando ambas íntimamente imbricadas)**

- **La información se almacena codificada en un "dominio magnético" cuyas características fundamentales son**
  - la dimensión espacial (superficie o volumen que ocupan)
  - la intensidad magnética
  - la orientación espacial

- **El dominio magnético ➔**
  - Es "**creado**" sobre el medio magnético mediante el "**cabezal grabador**"
  - Es "**leido**" mediante el "**cabezal lector**"
  - En algunos casos es "**destruido**" o "**borrado**" por el "**cabezal de borrado**", especialmente frecuentes en cintas de almacenamiento magnético (donde el peso no es un factor limitante y puede evitar la necesidad de "formateo previo" del medio a usar)

- **Las tecnología de las cabezas grabadoras y lectoras condicionan de forma determinante el las características del dominio magnético, y por tanto, las de la densidad de almacenamiento**

Medios de almacenamiento magnético:

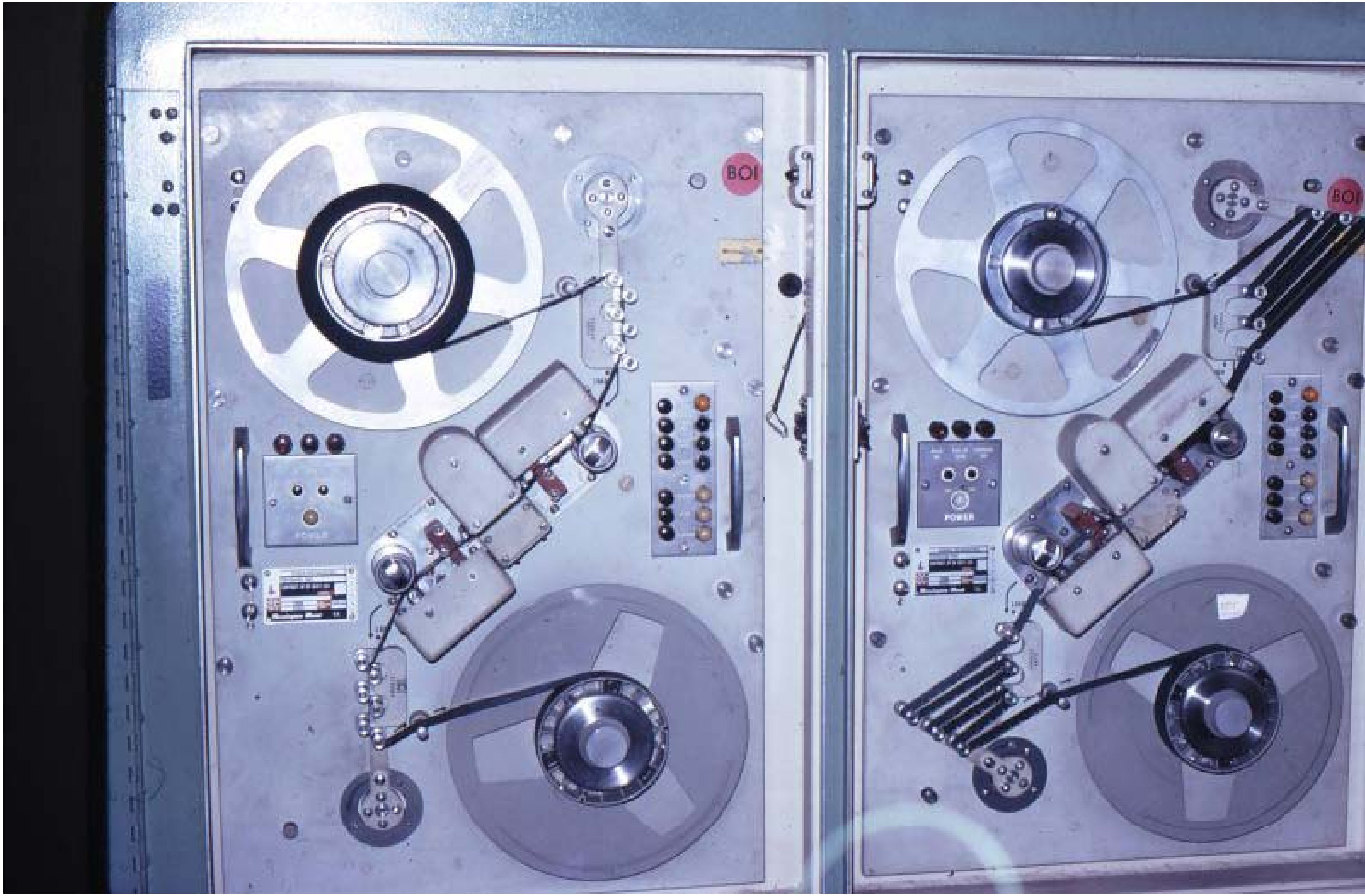
- **Dos tipos han dominado históricamente el almacenamiento magnético puro:**
  - **Cintas magnéticas**
    - **Grabación longitudinal** (muy robusto)
      - Derivado de los sistemas de grabación de audio analógico
      - Cabezal único (una sola pista) uni-y/o-bi-direccional
      - Cabezal múltiple (e.g. 9 pistas) uni-y/o-bi-direccional
    - **Grabación helicoidal** (más densidad pero más frágil)
      - Derivado de los sistemas de grabación de video analógico (VCR) y audio digital (DAT).
      - Cabezal principal rotatorio único o múltiple.
  - **Discos magnéticos**

Medios de almacenamiento magnético:

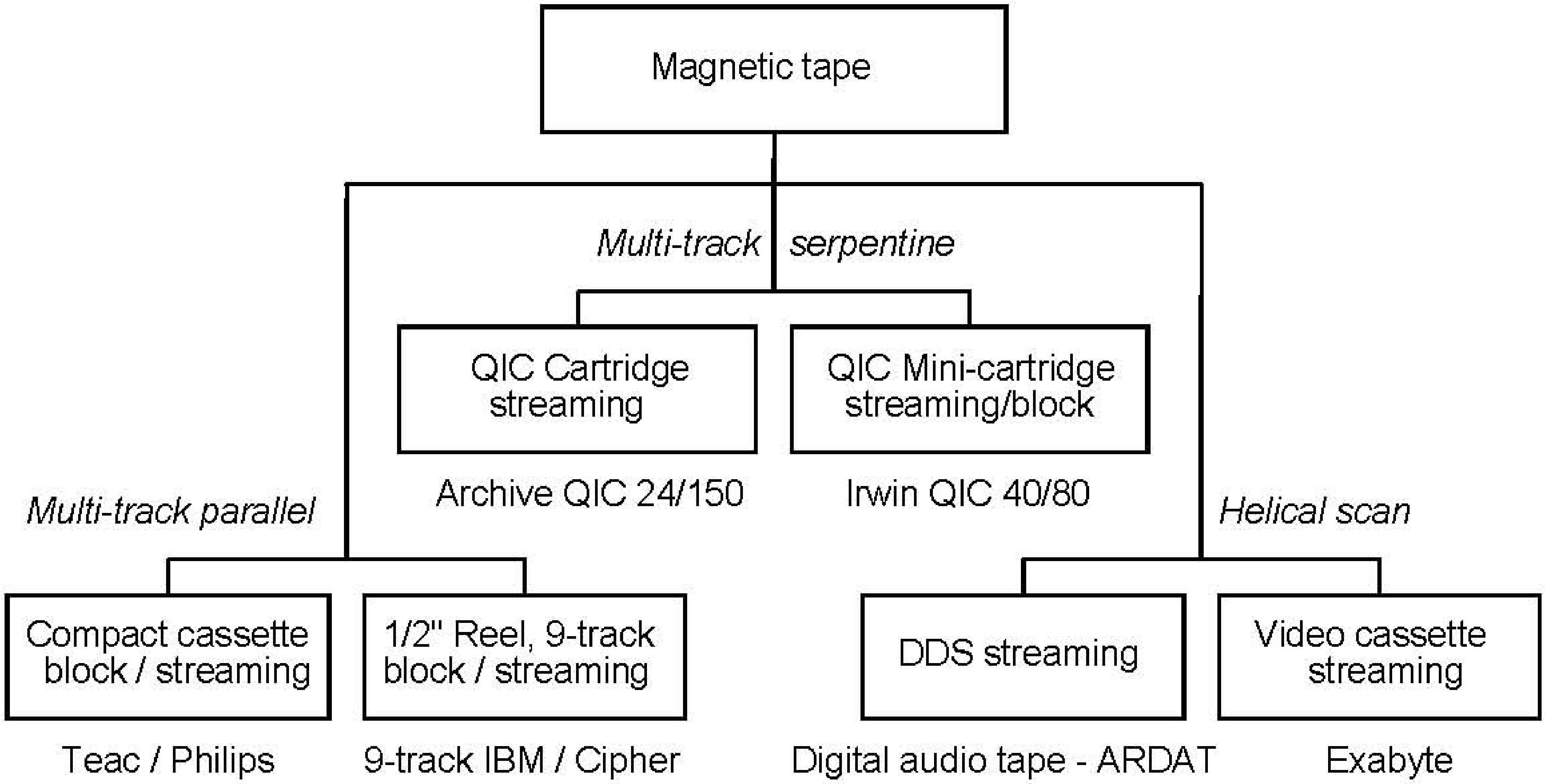
- **Dos tipos han dominado históricamente el almacenamiento magnético puro:**
  - **Cintas magnéticas**
    - Gran capacidad de almacenamiento
    - Mínimo coste
    - Usado típicamente para "**archivo**", "**copia de seguridad**" y distribución de software (históricamente).
    - Muy **lento en términos relativos**, debido al acceso secuencial a la información
  - **Discos magnéticos**
    - Buena capacidad de almacenamiento y velocidad de acceso
    - Coste tradicionalmente elevado, pero ha ido disminuyendo rápida y constantemente

La Visión de la Historia….Univac-1

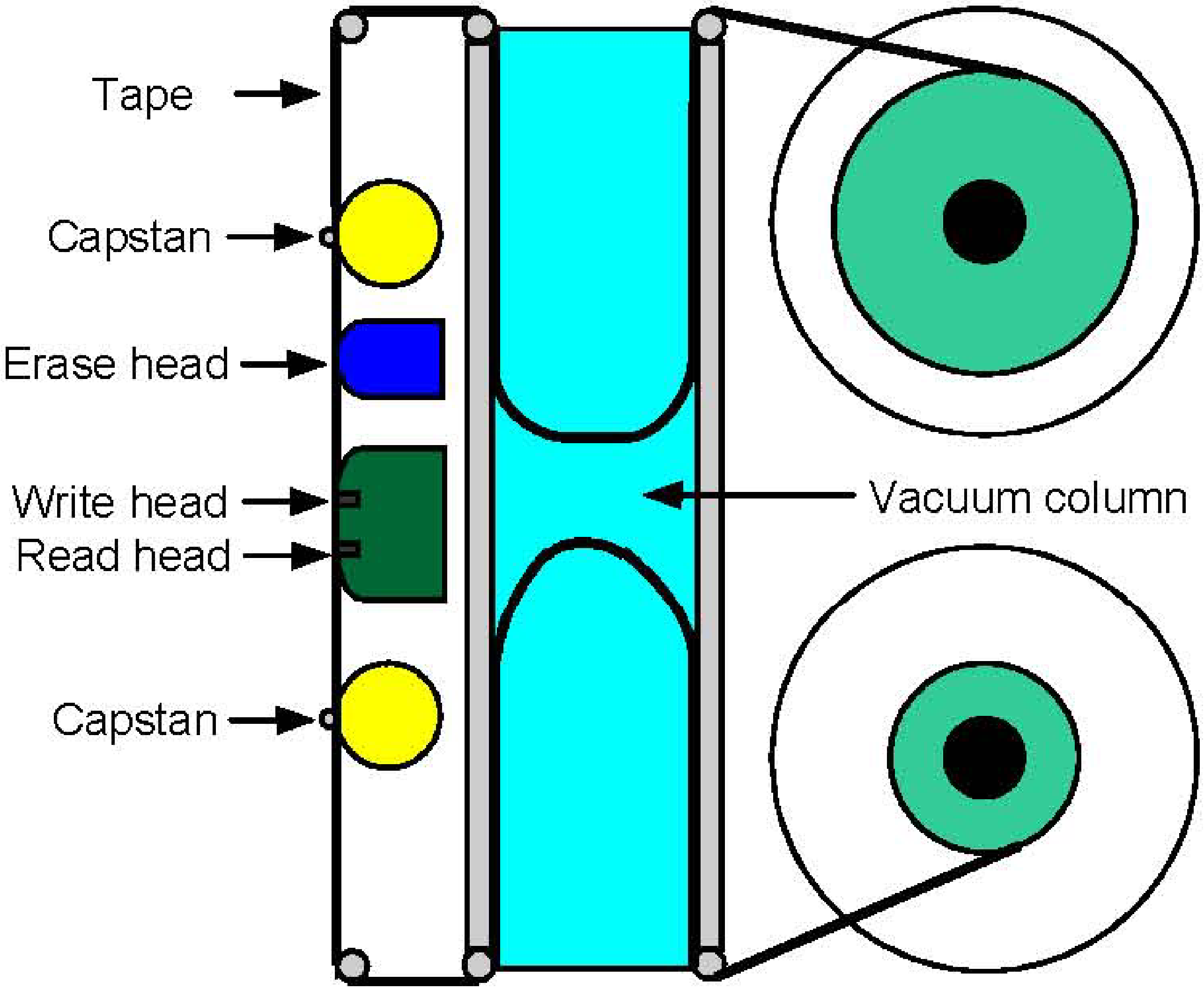
Cintas Magnéticas:



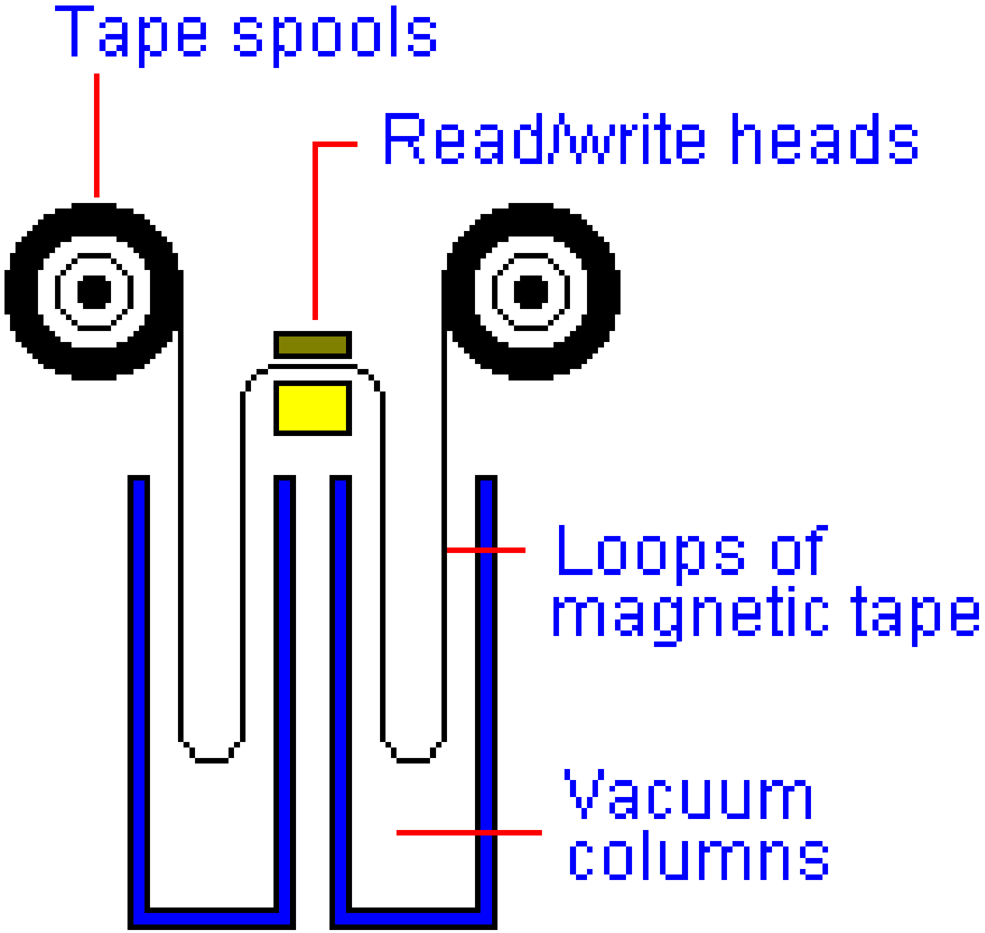
Tipos de Cintas Magnéticas:



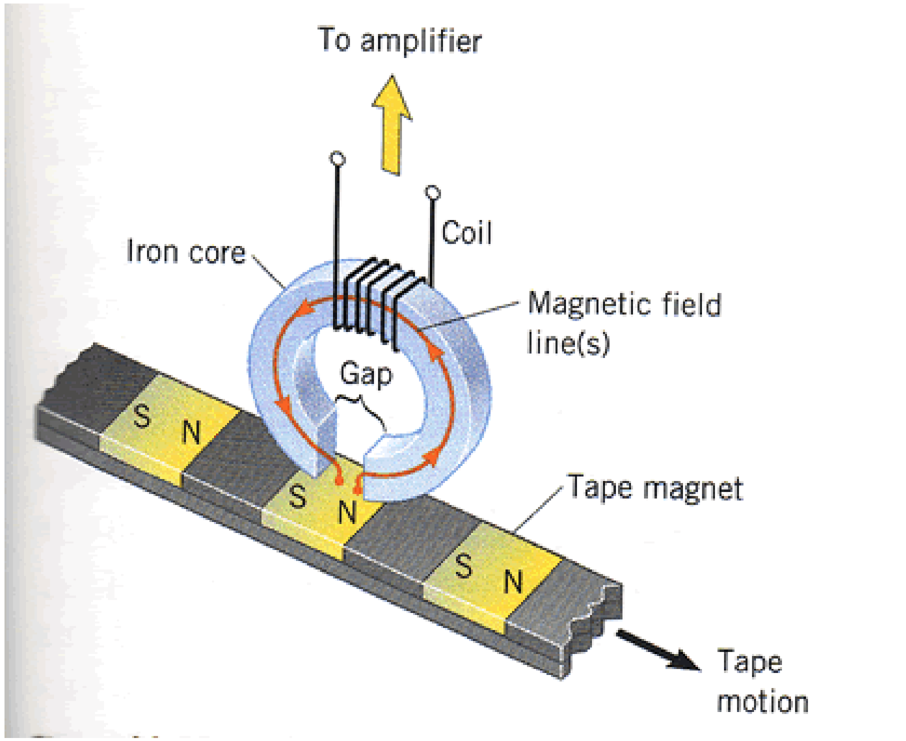
Cinta Magnética de Bobina Abierta:



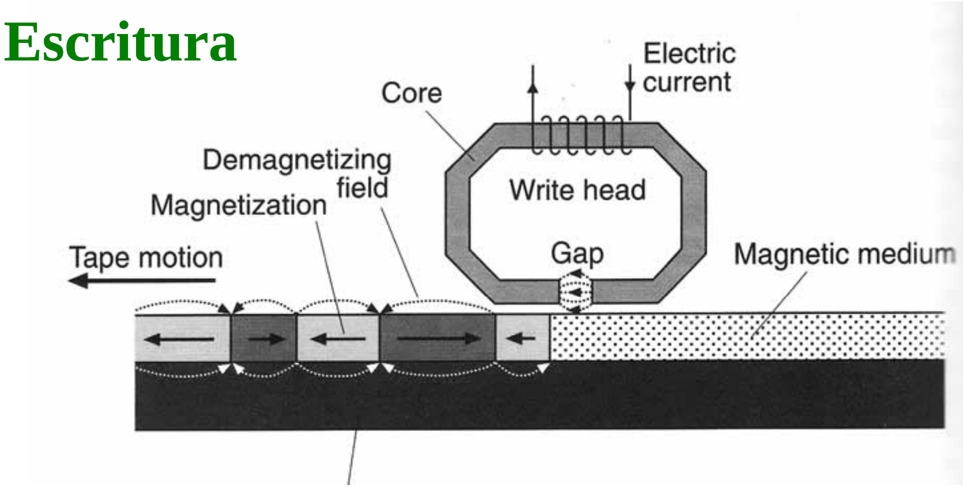
Cintas magnéticas



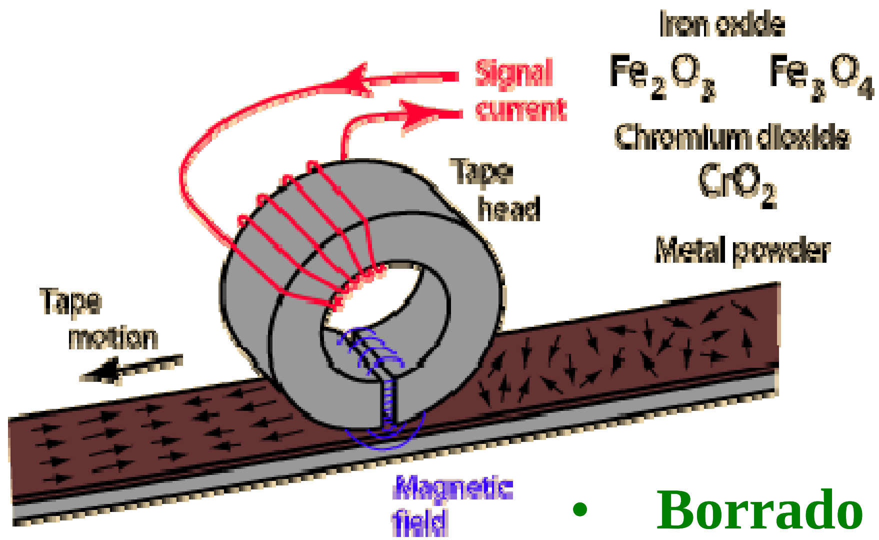
Cintas Magnéticas:
Cabezal de Escritura – Lectura -Borrado



- **Escritura**

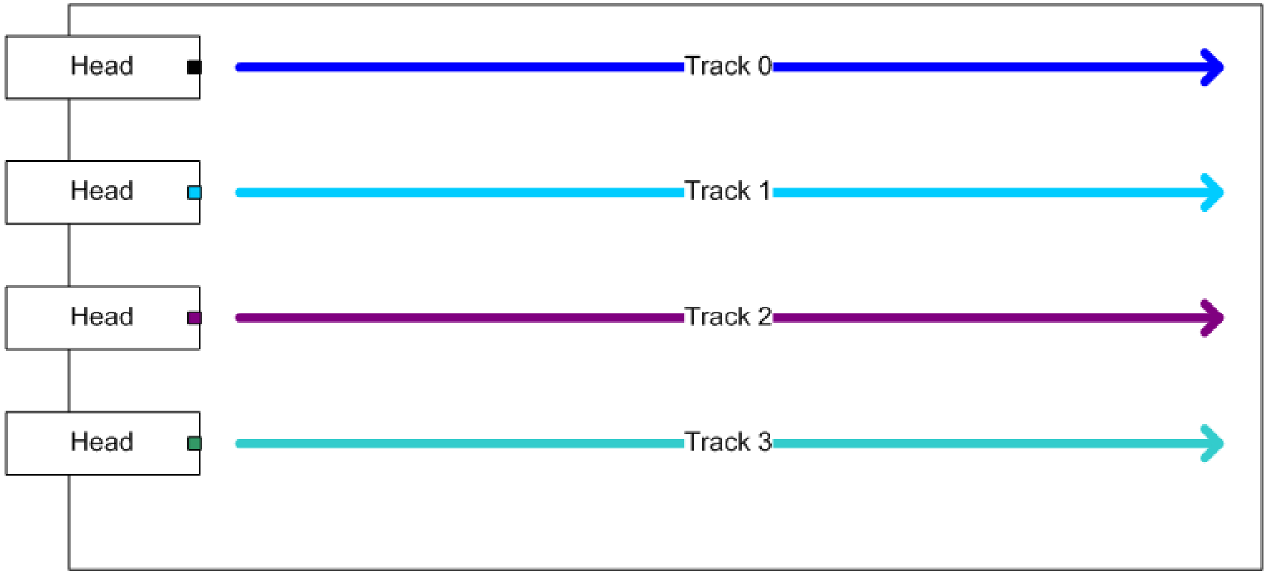


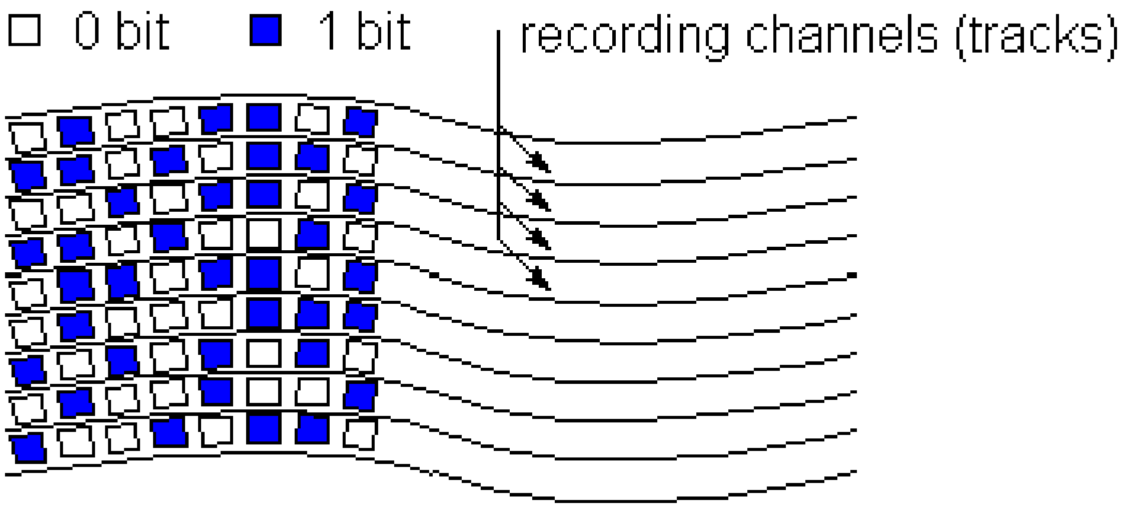
- **Lectura**

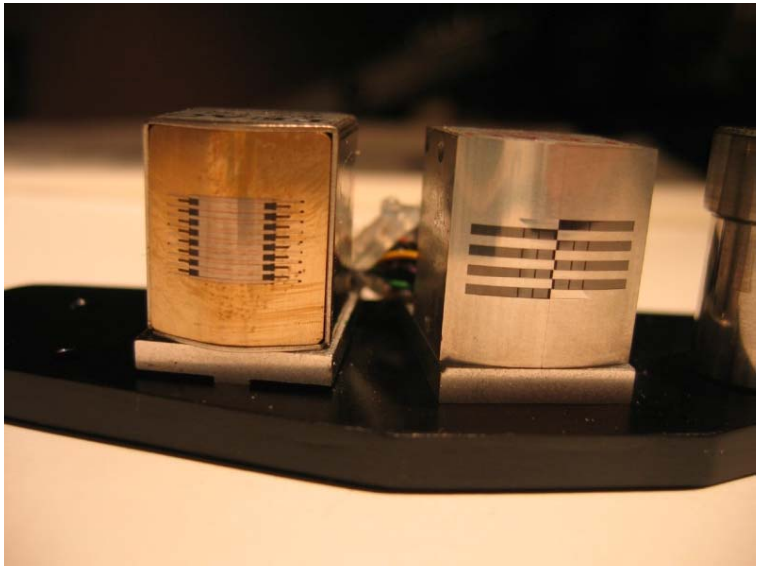


- **Borrado**

Cinta Magnética Formato de múltiples pistas:





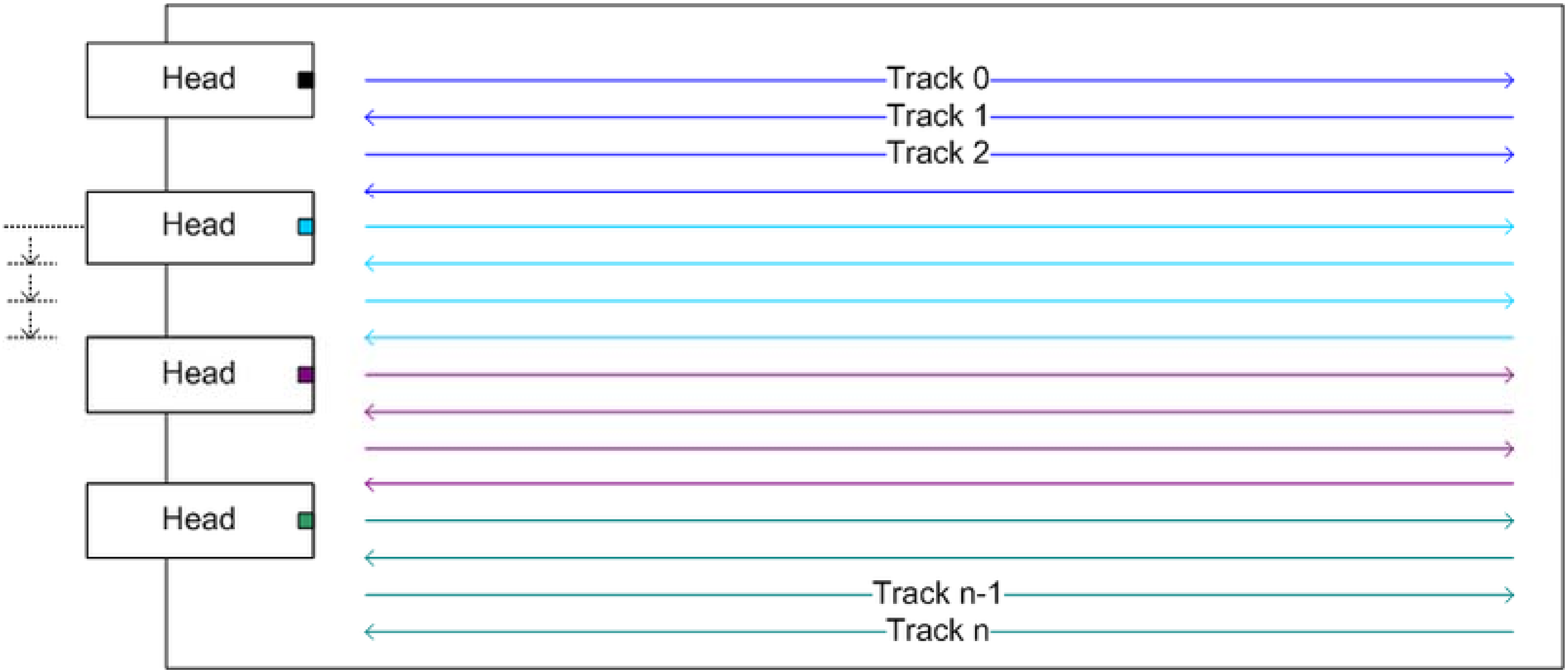


Ejemplo Especificaciones bobina abierta (reel-to-reel):

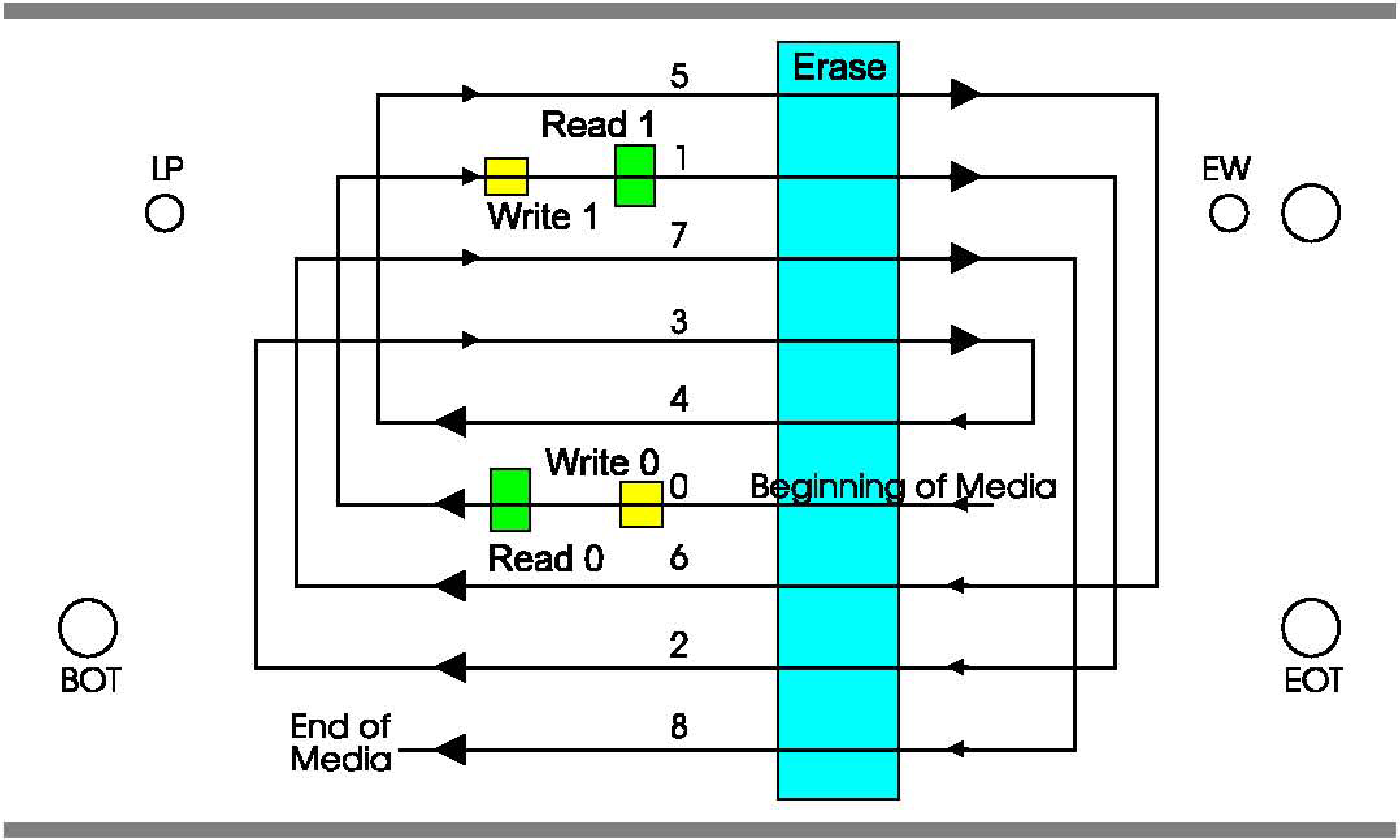
Table 1. Typical specifications of IBM reel-to-reel tape drives.

IBM Product No.	726	3420	3480
FCS (First customer shipment)	1953	1973	1985
Linear Density (BPI)	100	6250	38,000
Number of Tracks	7	9	18
Reel Capacity (MB)	2.2	156	200
Data Rate (KBytes/sec)	75	1250	3000
Recording Code	NRZI	GCR(0,2)	GCR(0,3)
Tape Transport	Vacuum	Vacuum	Cartridge

Cinta Magnética Formato bidireccional (serpentina):



Cinta Magnética Formato bidireccional (serpentina):

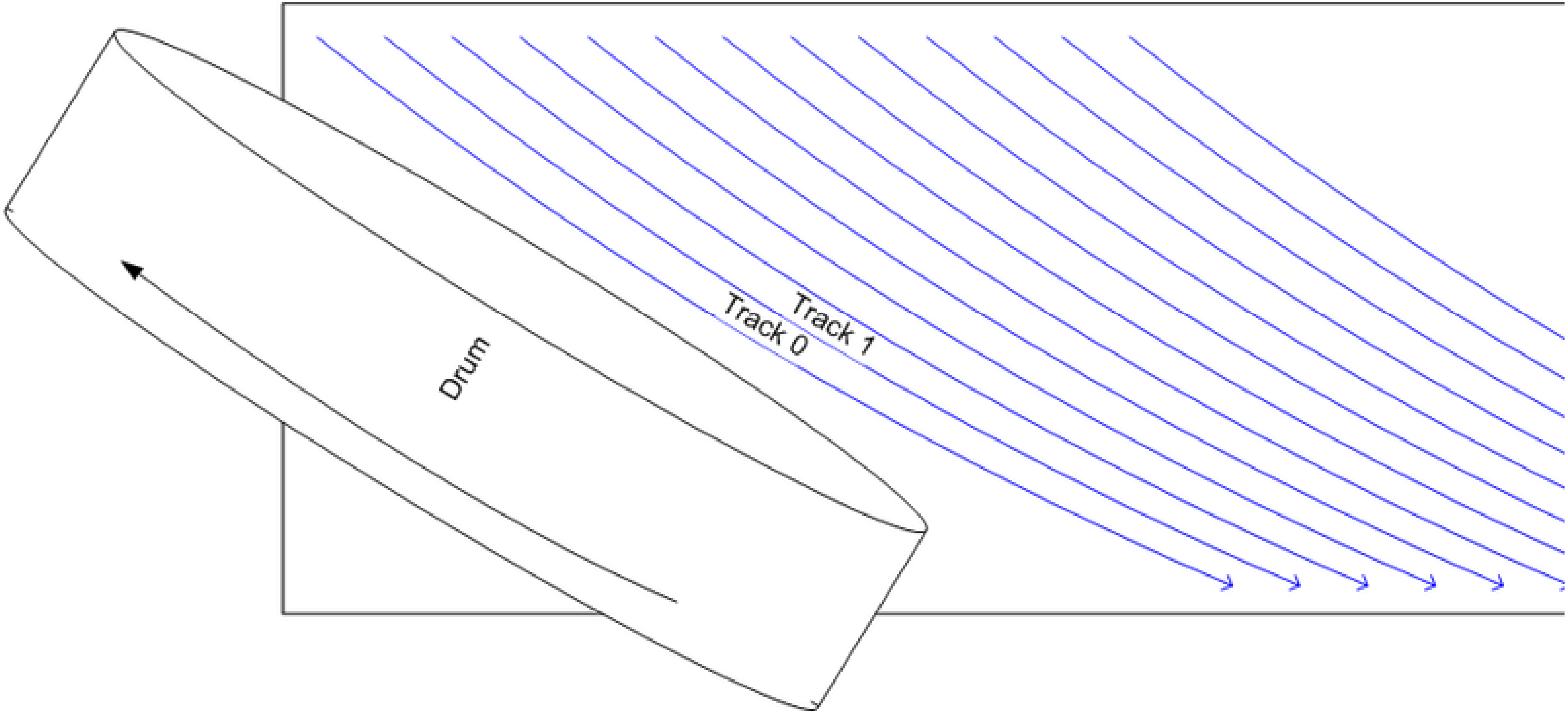


Ejemplo especificaciones:

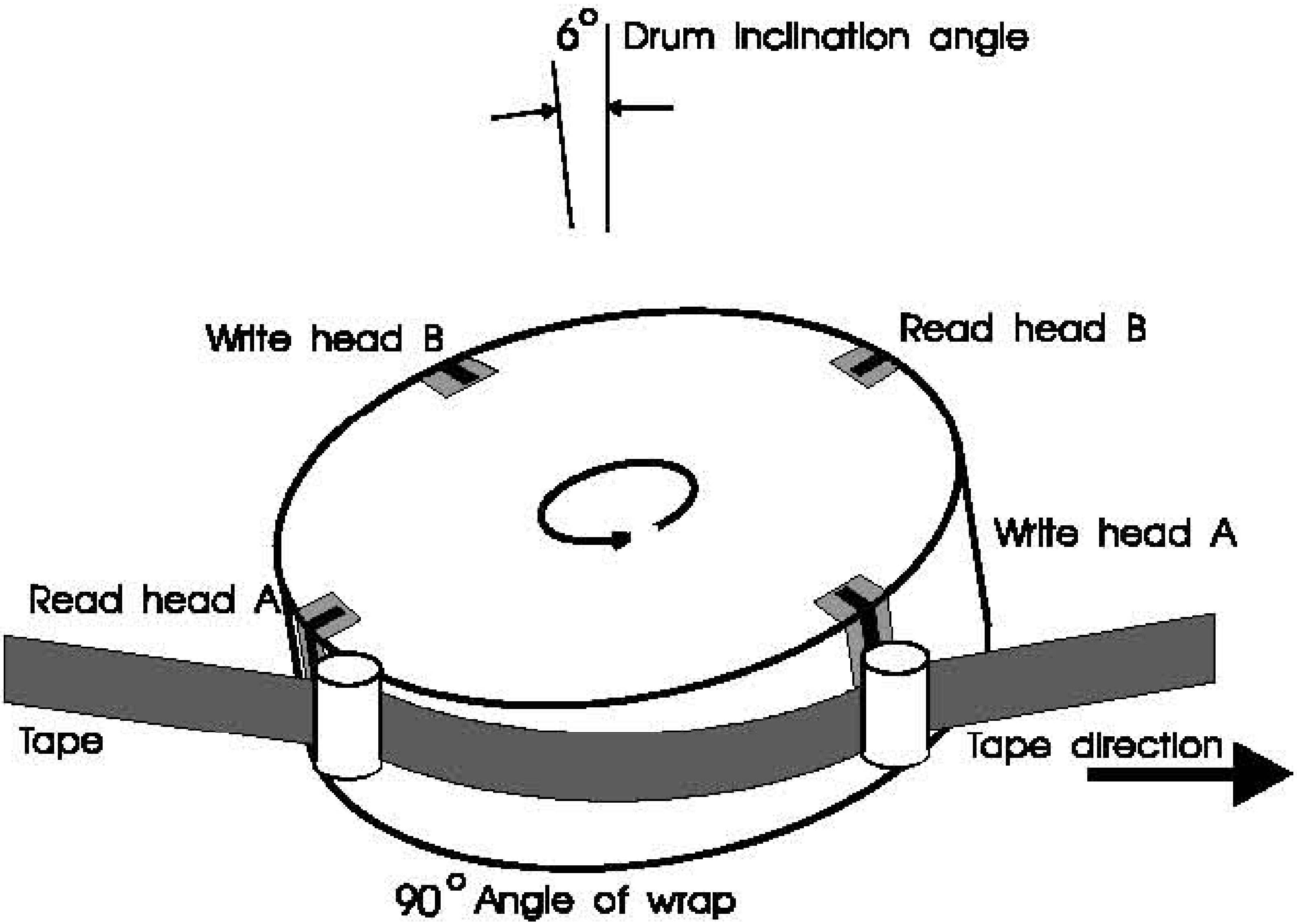
Table 2. QIC tape standards.

	QIC-24	QIC-150	QIC-525	QIC-1350
Capacity (formatted) MB	45 or 60	125 or 150	320 or 525	1.35 GB
Track Format	9	18	26	30
Flux Density	10,000 ftpi	12,500 ftpi	20,000 ftpi	38,750 ftpi
Data Density	8,000 bpi	10,000 bpi	16,000 ftpi	51,667 bpi
Tape Speed	90 ips	90 ips	120 ips	120 ips
Data Transfer Rate KBytes/Sec	90	112.5	240	600
Recording Code	GCR (0,2)	GCR (0,2)	GCR (0,2)	RLL(1,7)
Track Width (in)	0.0135	0.0056	0.0070	0.0070
Tape Length (ft)	450 or 600	600	600 or 1000	750
Soft Error Rate	1 in 10^8	1 in 10^8	1 in 10^8	1 in 10^8
Hard Error Rate	1 in 10^{10}	1 in 10^{10}	1 in 10^{10}	1 in 10^{10}

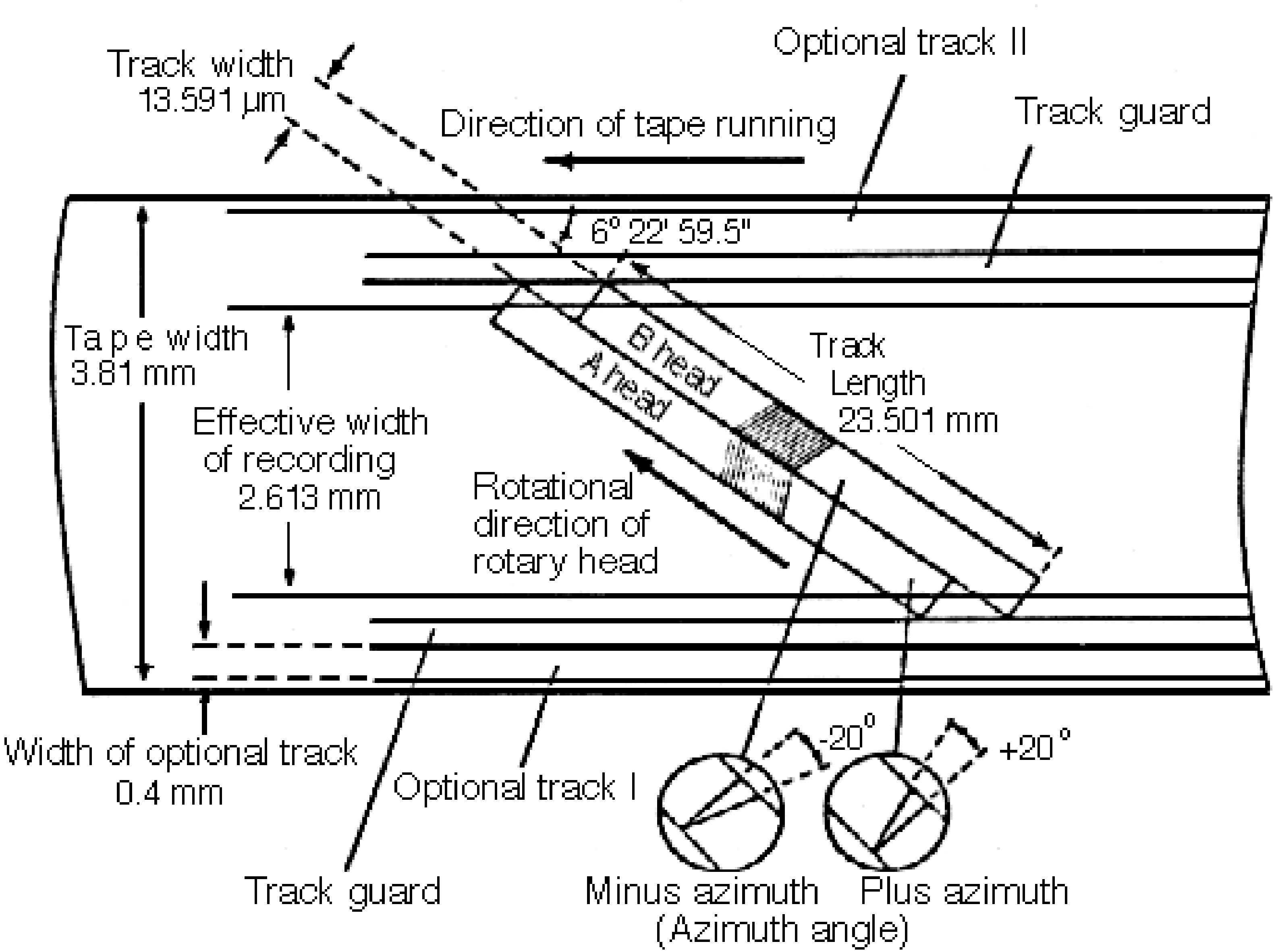
Cintas Magnéticas:
Cabezal de Escritura – Lectura helicoidal



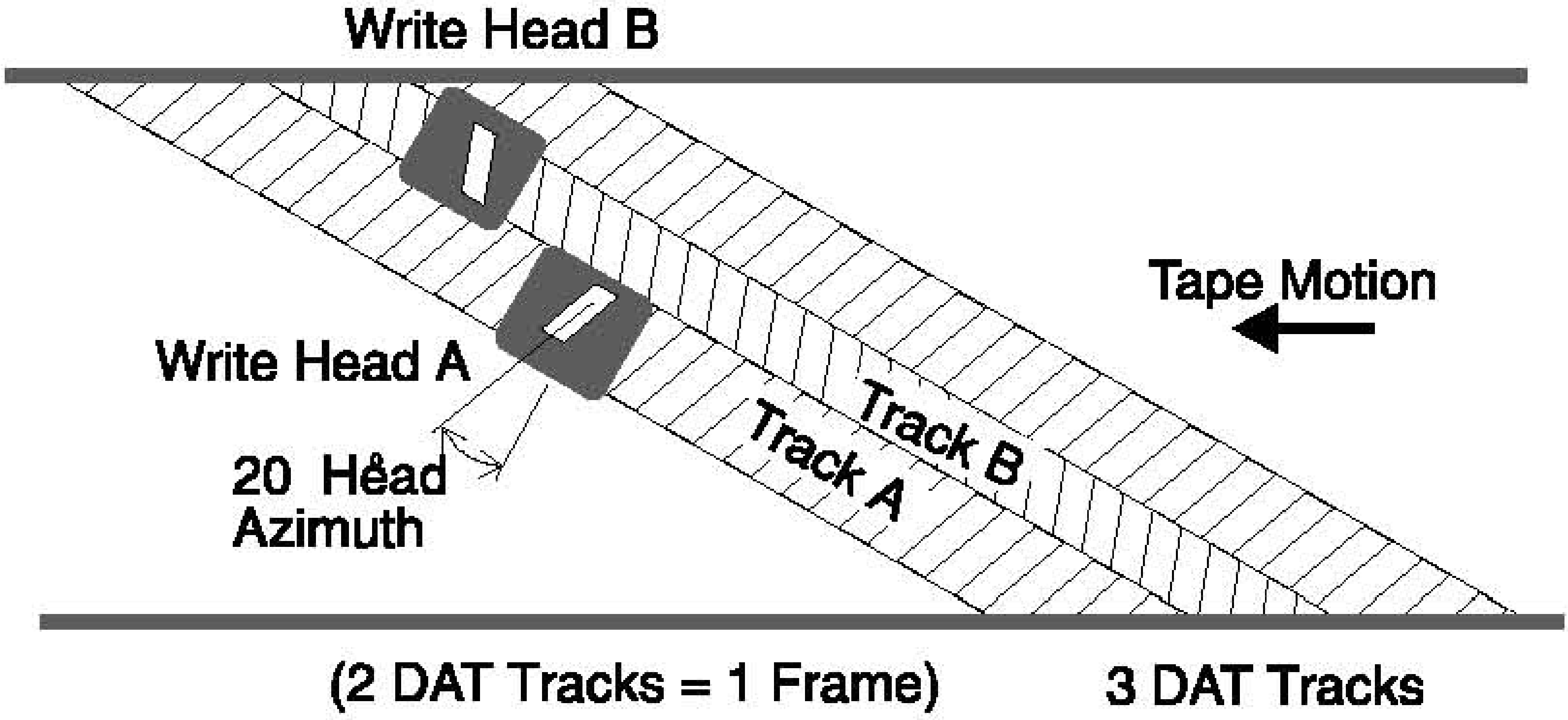
Cintas Magnéticas (DAT):
Detalle del Cabezal de Escritura – Lectura helicoidal



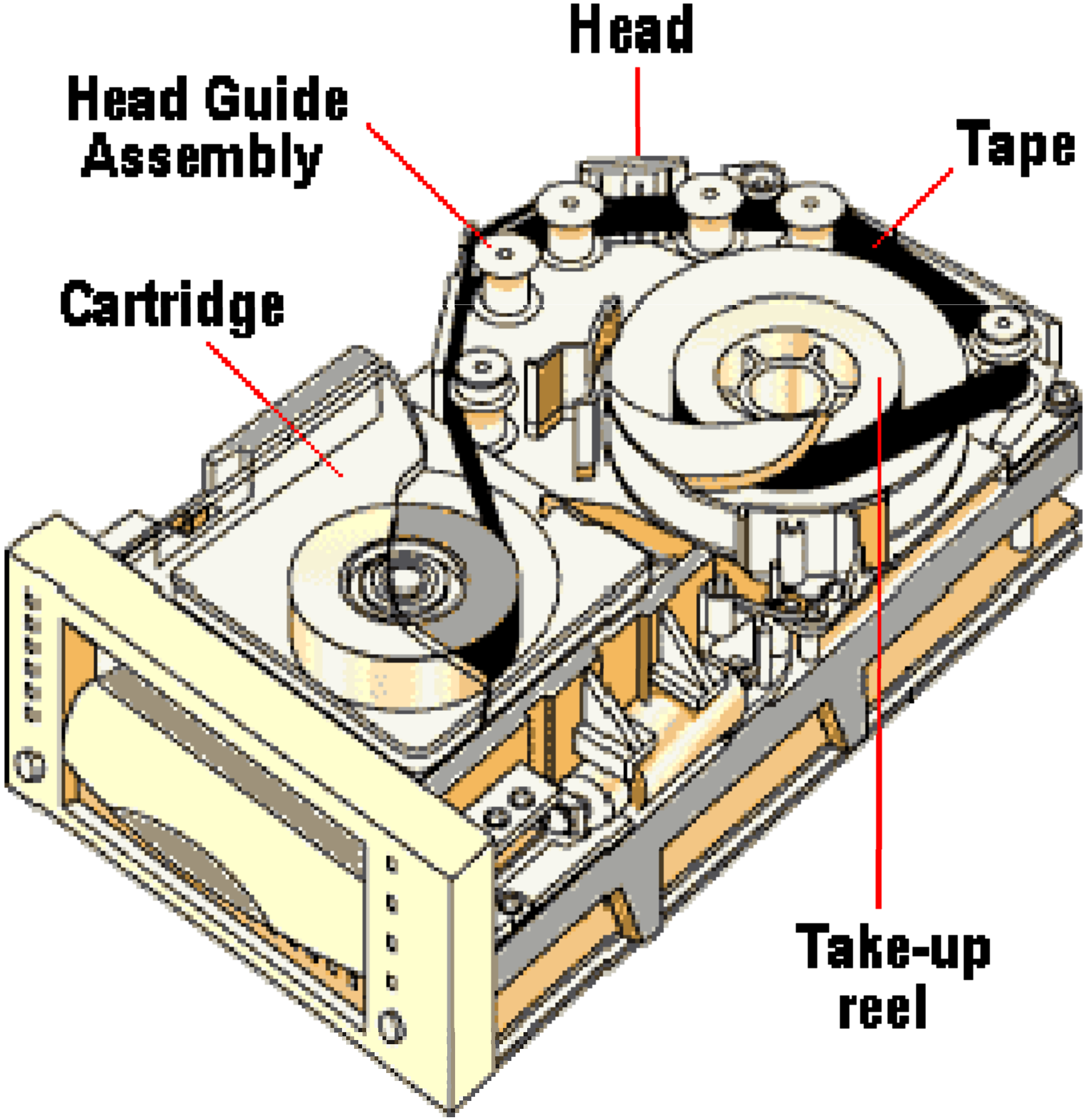
Cintas Magnéticas: Cabezal de Escritura – Lectura Helicoidal



Cintas Magnéticas (DAT): Detalle del formato helicoidal



Cintas Magnéticas:
Cabezal de Escritura – Lectura helicoidal



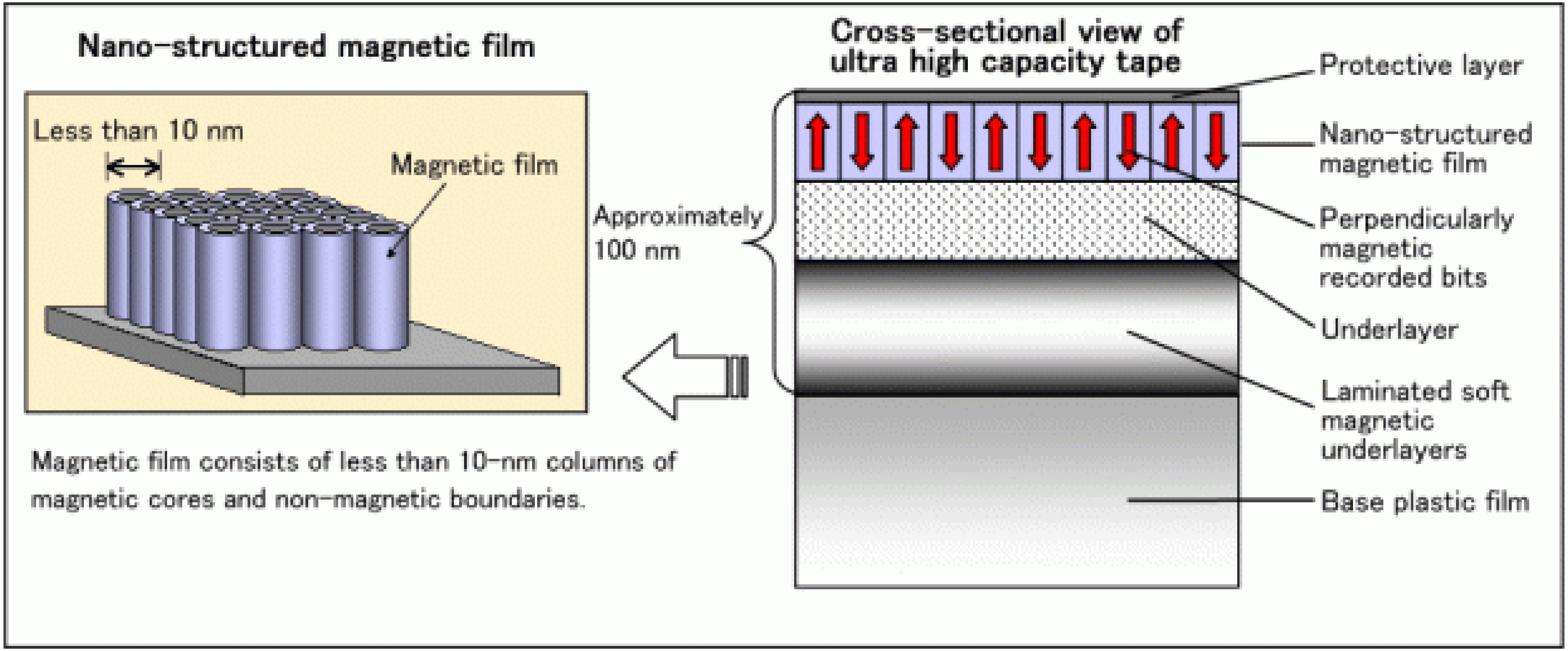

2J02BB
R-P
R-A
P-A

Ejemplo de especificaciones

Table 4. Typical specifications of a DDS DAT drive.

Product:	Archive Python 4330XT
Capacity:	1.3 GBytes with 60m tape.
Sustained transfer rate:	183 Kbytes/sec, sustained.
Average access time:	20 sec. seek time.
Small form factor:	3 1/2"
Standard recording format:	ANSI DDS
Low cost:	Currently US$0.01 / Mbyte.
Interface:	SCSI-1 and SCSI-2
Media	4 mm. DAT Cartridge, 60/90m length.
Packing density	1869 tracks/in.
Areal density	114 Mbits/sq. in.
Uncorrectable error rate	Using ECC, 1 in 10^{15} bits.
Drum rotation speed:	2000 RPM
Tape speed:	0.32 in/sec.
Search/rewind speed	200 X normal speed
Head-to-tape speed:	123 in/sec, Helical scan (RDAT)

Cintas Magnéticas. Realizaciones Actuales: Cabezal de Escritura – Lectura de dominio transversal



- **Hasta 35 Tbytes por unidad de cinta, gracias a la utilización de las últimas tecnologías desarrolladas para discos magnéticos en cabezas y material de soporte magnético**

http://www.zurich.ibm.com/news/10/storage.html

http://www.flickr.com/photos/ibm_research_zurich/sets/72157623247462714/

IBM-35TBytes:

http://www.flickr.com/photos/ibm_research_zurich/4268767654/in/set-72157623247462714/lightbox/