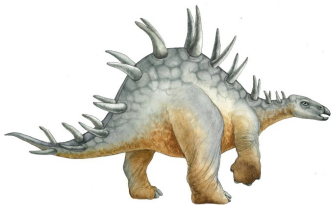
Capítulo 12: Sistemas de Almacenamiento Masivo

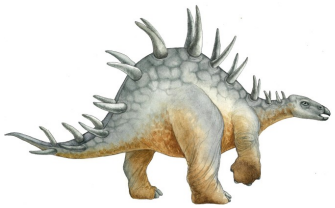
Capítulo 12: Sistemas de Almacenamiento Masivo

- Descripción de la estructura de almacenamiento masivo
- Estructura de disco
- Conexiones de disco
- Planificación de disco
- Manejo de disco
- Estructura de RAID
- Implementación de almacenamiento estable
- Dispositivos de almacenamiento terciario
- Detalles de rendimiento



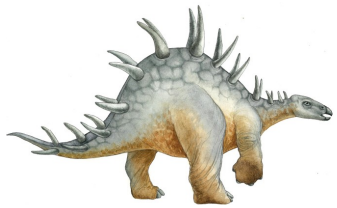
Objetivos

- Describir la estructura física del almacenamiento secundario y terciario y los efectos resultantes en el uso de los dispositivos
- Explicar las características de rendimiento de los dispositivos de almacenamiento masivo
- Discutir los servicios provistos para almacenamiento masivo, incluyendo RAID y HSM

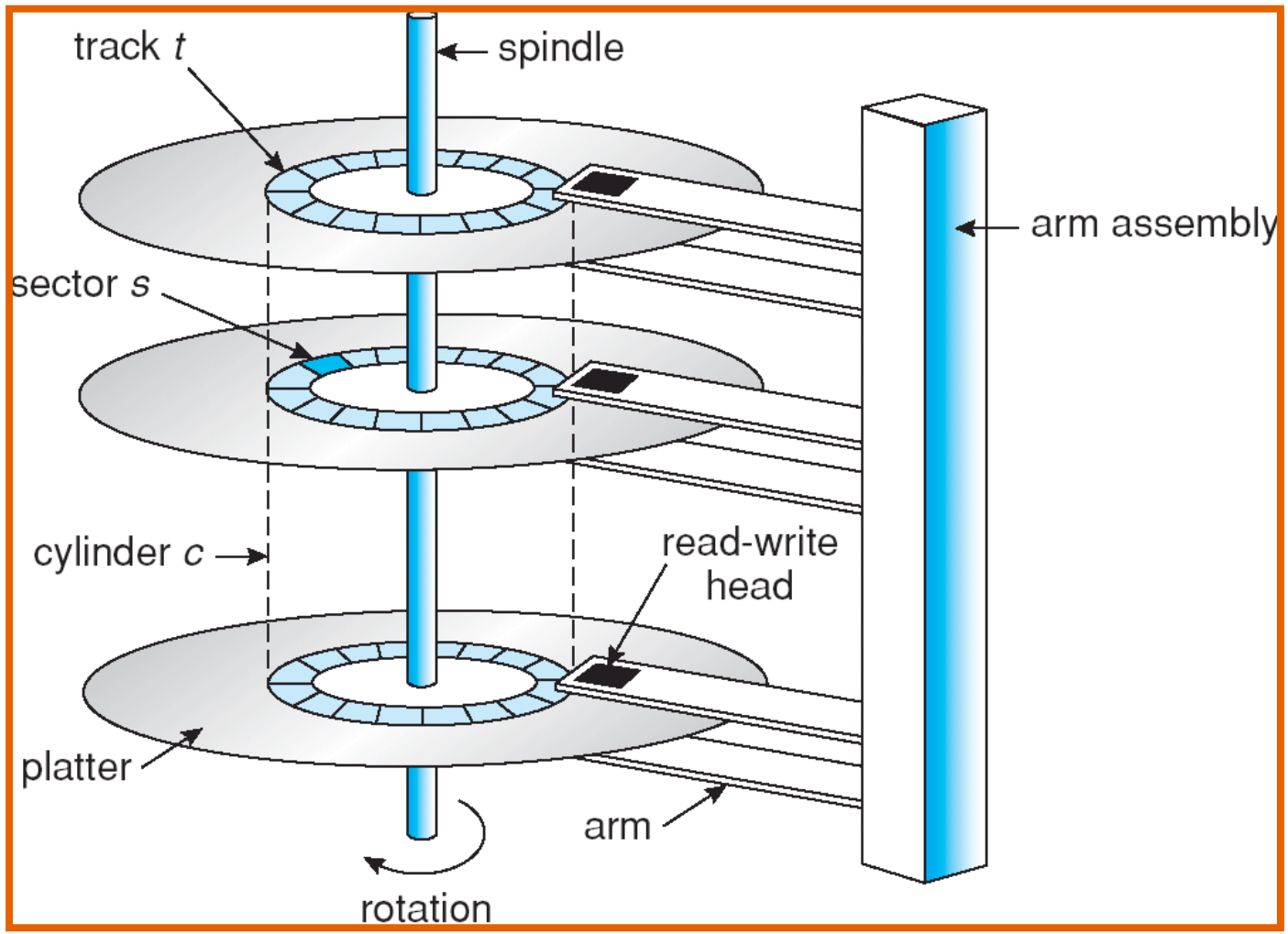


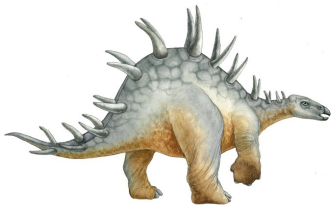
Descripción de Estructura de Almacenamiento Masivo

- Discos magnéticos ofrecen el grueso del almacenamiento masivo en computadoras modernas
  - Un drive típico rota de 60 a 200 veces por segundo
  - **Tasa de transferencia** es el ritmo al cuál fluyen datos entre el dispositivo y la comptuadora
  - **Tiempo de posicionamiento** (**tiempo de acceso-aleatorio**) es el tiempo que toma mover el brazo del disco al cilindro deseado (**tiempo de búsqueda**) y el tiempo para que el sector deseado rote y quede debajo de cabeza (**latencia rotacional**)
  - **Choque de cabeza** el resultado cuando la cabeza del disco hace contacto con la superficie del disco
    - Esto es muy malo
- Los discos pueden ser externos
- Dispositivo conectado a la computadora via **bus de E/S**
  - Los buses varían, incluyendo **EIDE, ATA, SATA, USB, Fibre Channel, SCSI**
  - **Controlador de host** en la computadora utiliza el bus para hablar con el **controlador de disco** o el **arreglo de discos**



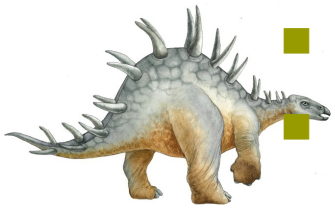
Mecanismo de disco de cabeza-móvil

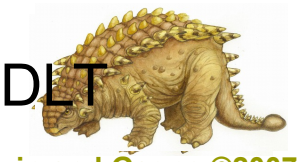




Descripción de Estructura de Almacenamiento Masivo

- Cinta magnética
  - Medio original de almacenamiento-secundario
  - Relativamente permanente y mantiene grandes cantidades de datos
  - Tiempo de acceso lento
  - Acceso aleatorio ~1000 veces más lento que el disco
  - Principalmente utilizado para respaldo, almacenamiento de datos utilizados con poca frecuencia, medio de transferencia entre sistemas
  - Se mantiene en *carrete* y se avanzan o retrasan sobre la cabeza de lectura-escritura
  - Una vez bajo la cabeza, ritmo de transferencia comparable al de disco
  - Almacenamiento típico 20-200GB
  - Tecnologías comunes: 4mm, 8mm, 19mm, LTO-2 y SDLT

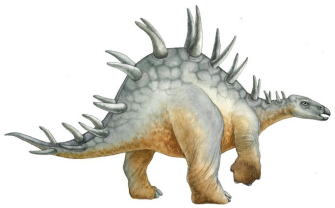




Estructura de disco

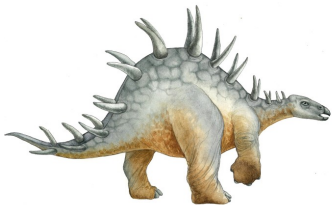
- Los discos son referidos como grandes arreglos 1-dimensionales de *bloques lógicos,* donde el bloque lógico es la unidad mínima de transferencia.

- El arreglo 1-dimensional de bloques lógicos se mapea secuencialmente en los sectores del disco.
  - El sector 0 es el primer sector del primer trac en el cilindro más externo.
  - El mapeo procede en orden a lo largo del trac, luego los demás tracks en el cilindro y finalmente en el resto de los cilindros del más externo, al más interno.



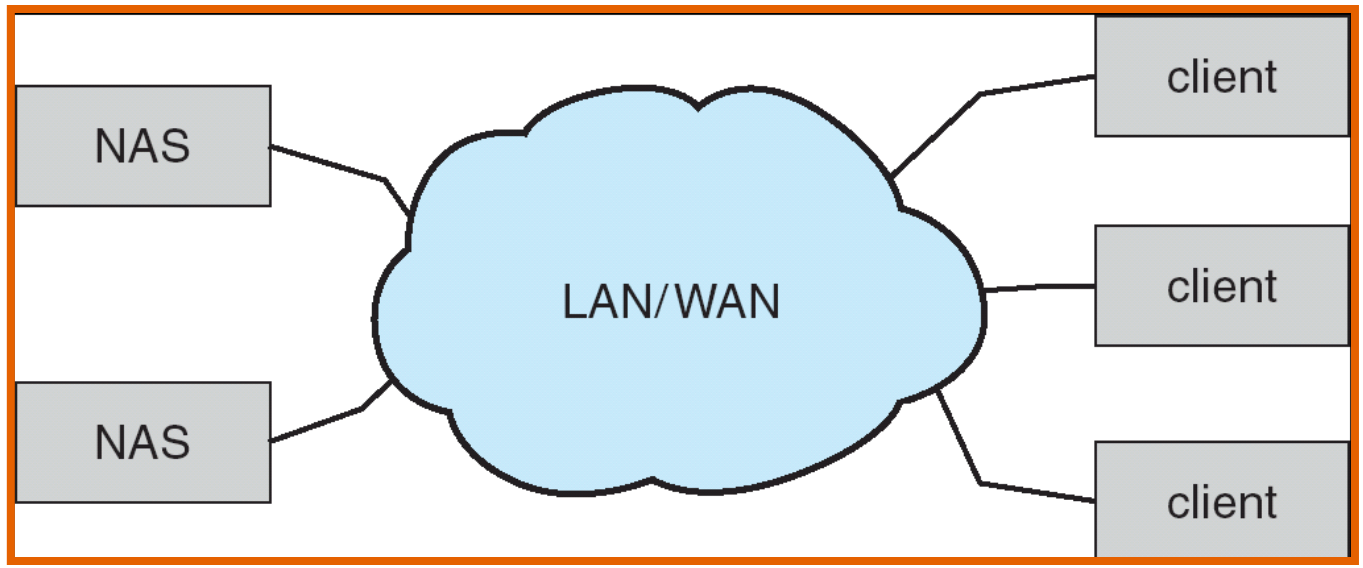
Conexión de disco

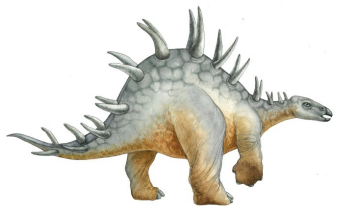
- Almacenamiento conectado al host se accede a través de puertos de E/S que hablan con busses de E/S
- SCSI es un bus en sí mismo, hasta 16 dispositivos en un sólo cable, **iniciador de SCSI** solicita la operación y los **objetivos SCSI** la realizan
  - Cada objetivo puede tener hasta 8 **unidades lógicas** (discos conectados al controlador)
  - FC es arquitectura serial de alta velocidad
  - Puede fabricarse con un espacio de direcciones de 24-bits – la base de **storage area networks** (**SAN**s): muchos hosts se conectan con varias unidades de almacenamiento
  - Puede ser un **ciclo arbitrado** (arbitrated loop, **FC-AL**) de 126 dispositivos



Almacenamiento conectado-en-red

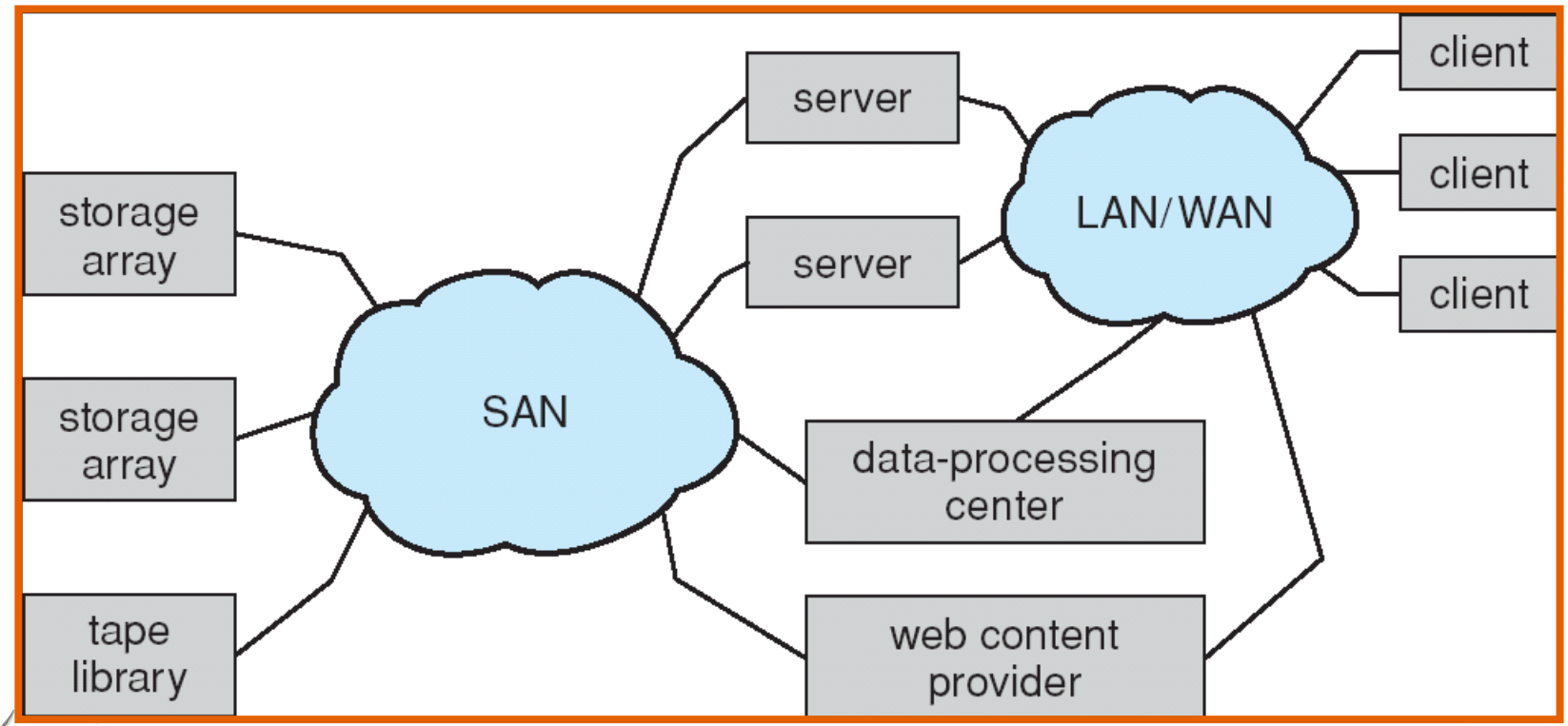
- Almacenamiento conectado-en-red (Network-attached storage, **NAS**) es almacenamiento disponible a través de una red en lugar de una conexión local (como un bus)
- NFS y CIFS son protocolos comunes
- Se implementan a través de llamadas a procedimientos remotos (RPCs) entre el host y el almacenamiento
- El nuevo protocolo iSCSI utiliza red IP para soportar el protocolo SCSI

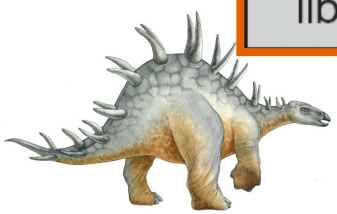




Storage Area Network (SAN)

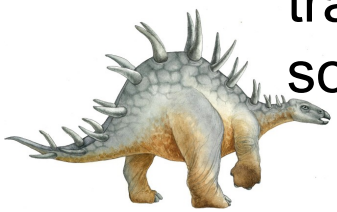
- Común en ambientes con gran almacenamiento (y cada vez más utilizado)
- Muchos hosts conectados a muchos arreglos de almacenamiento -- flexible





Planificador de disco

- El sistema operativo es responsable de utilizar el hardware eficientemente — en el caso de los discos, esto significa un tiempo de acceso rápido y ancho de banda.
- Tiempo de acceso tiene dos componentes mayores
  - *Tiempo de búsqueda* es el tiempo que toma al disco mover las cabezas al cilindro que contiene el sector deseado.
  - *Latencia rotacional* es el tiempo adicional esperando a que el disco rote y posicione el sector deseado a la cabeza del disco.
- Minimizar tiempo de búsqueda
- Tiempo de búsqueda $\approx$ distancia de búsqueda
- El ancho de banda del disco es el número total de bytes transferidos, divididos por el tiempo total entre la primera solicitud de servicio y el termino de la última transferencia.

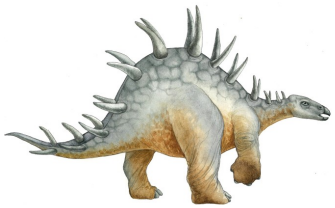


Planificador de disco

- Existen distintos algoritmos para organizar el servicio de solicitudes de E/S de disco.
- Vamos a ilustrarlos con una cola de solicitudes(0-199).

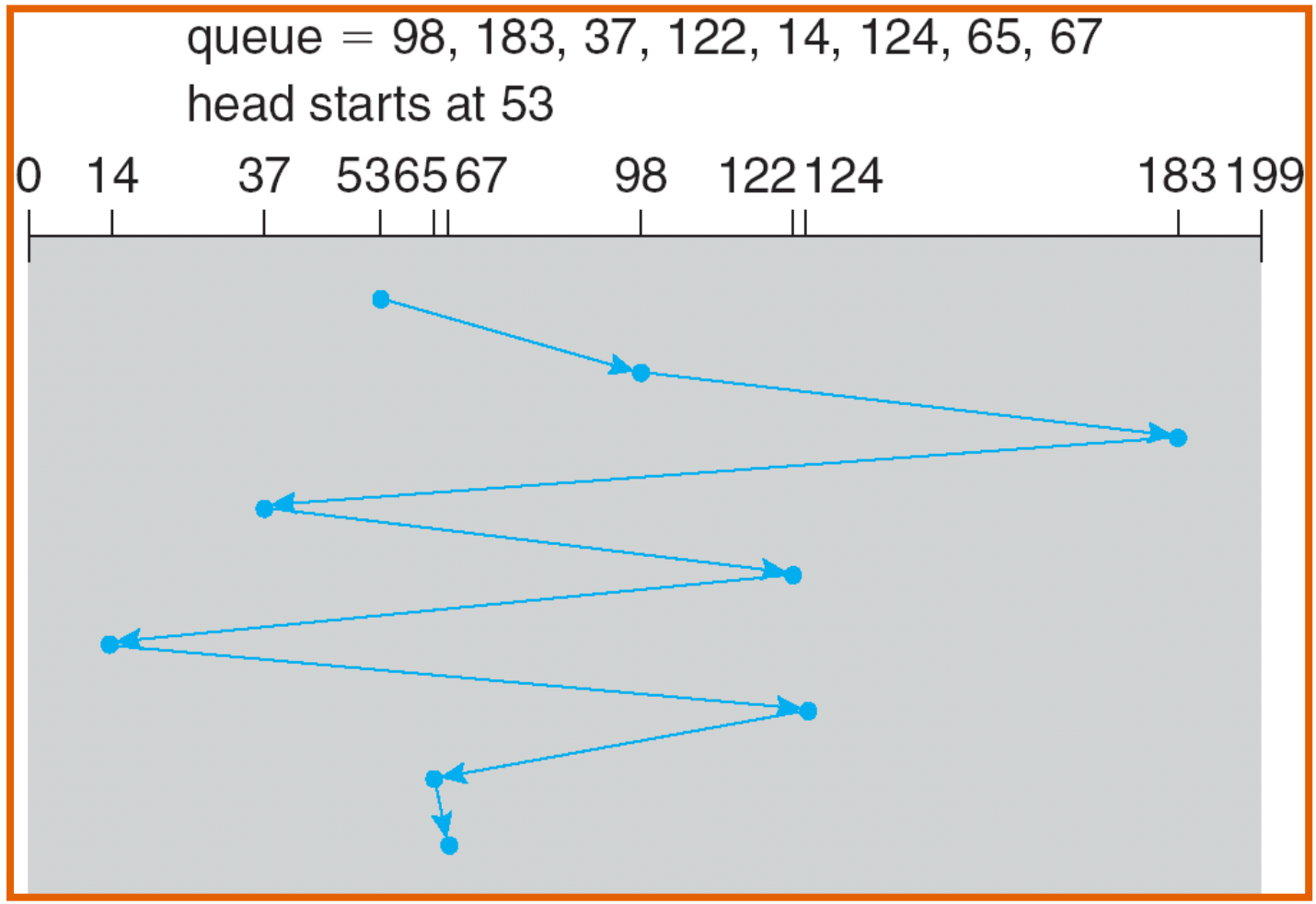
  98, 183, 37, 122, 14, 124, 65, 67

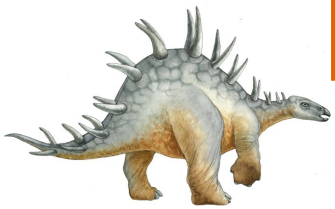
  Apuntador de cabeza: 53



FCFS

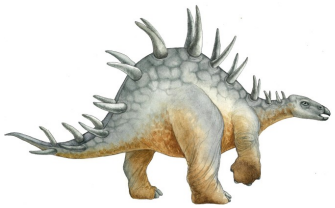
La ilustración muestra el movimiento total de la cabeza de 640 cilindros.



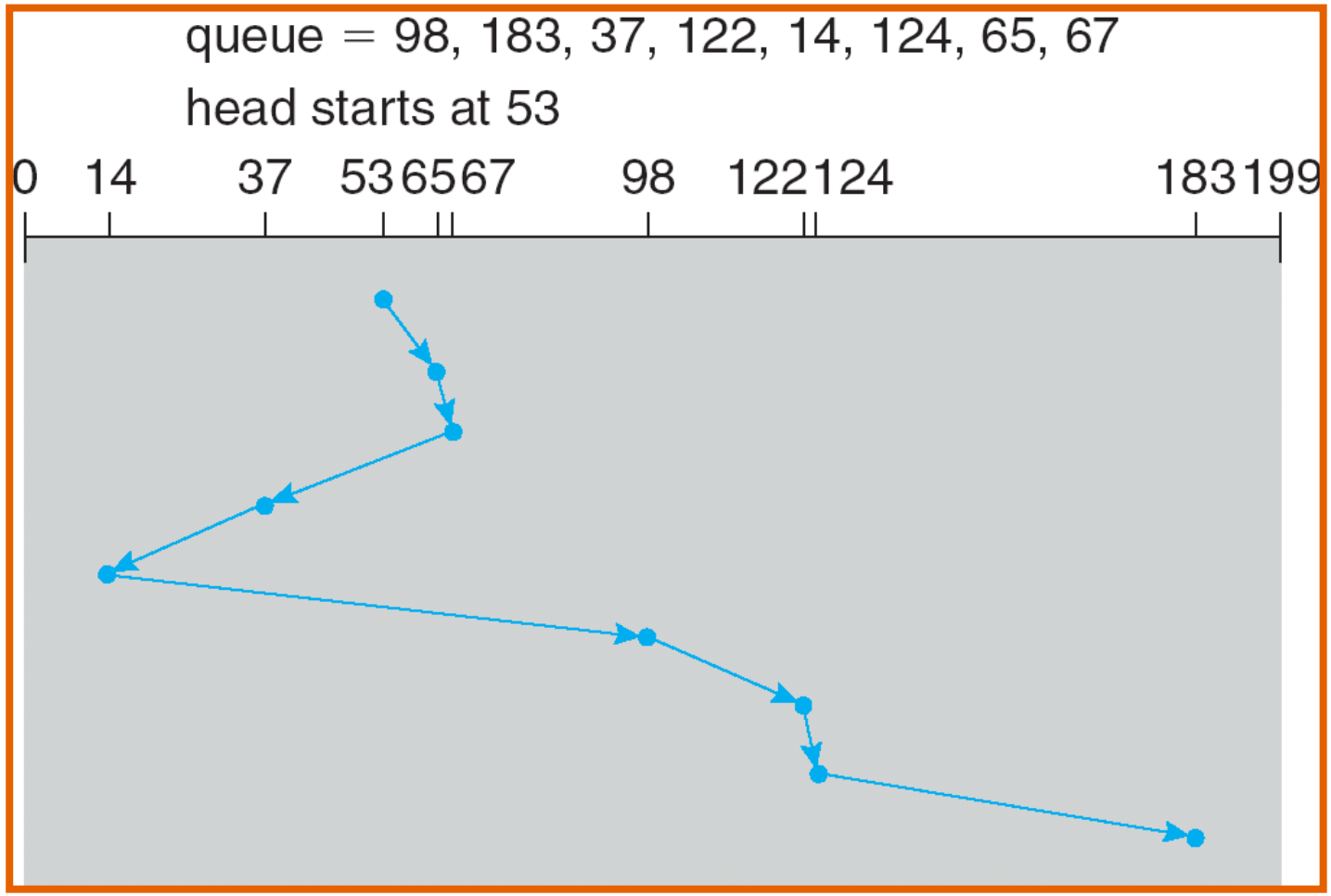


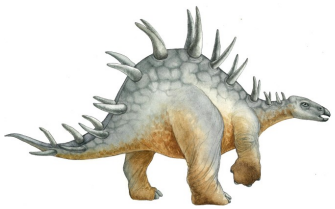
SSTF

- Selecciona la solicitud con el menor tiempo de búsqueda desde la posición actual de la cabeza.
- SSTF es una forma de planificador SJF; puede provocar hambruna para algunas solicitudes.
- La siguiente ilustración muestra el movimiento total de cabeza para 236 cilindros.



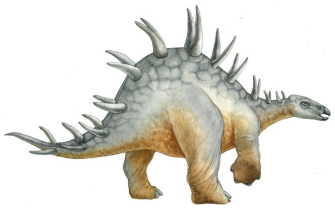
SSTF (Cont.)


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

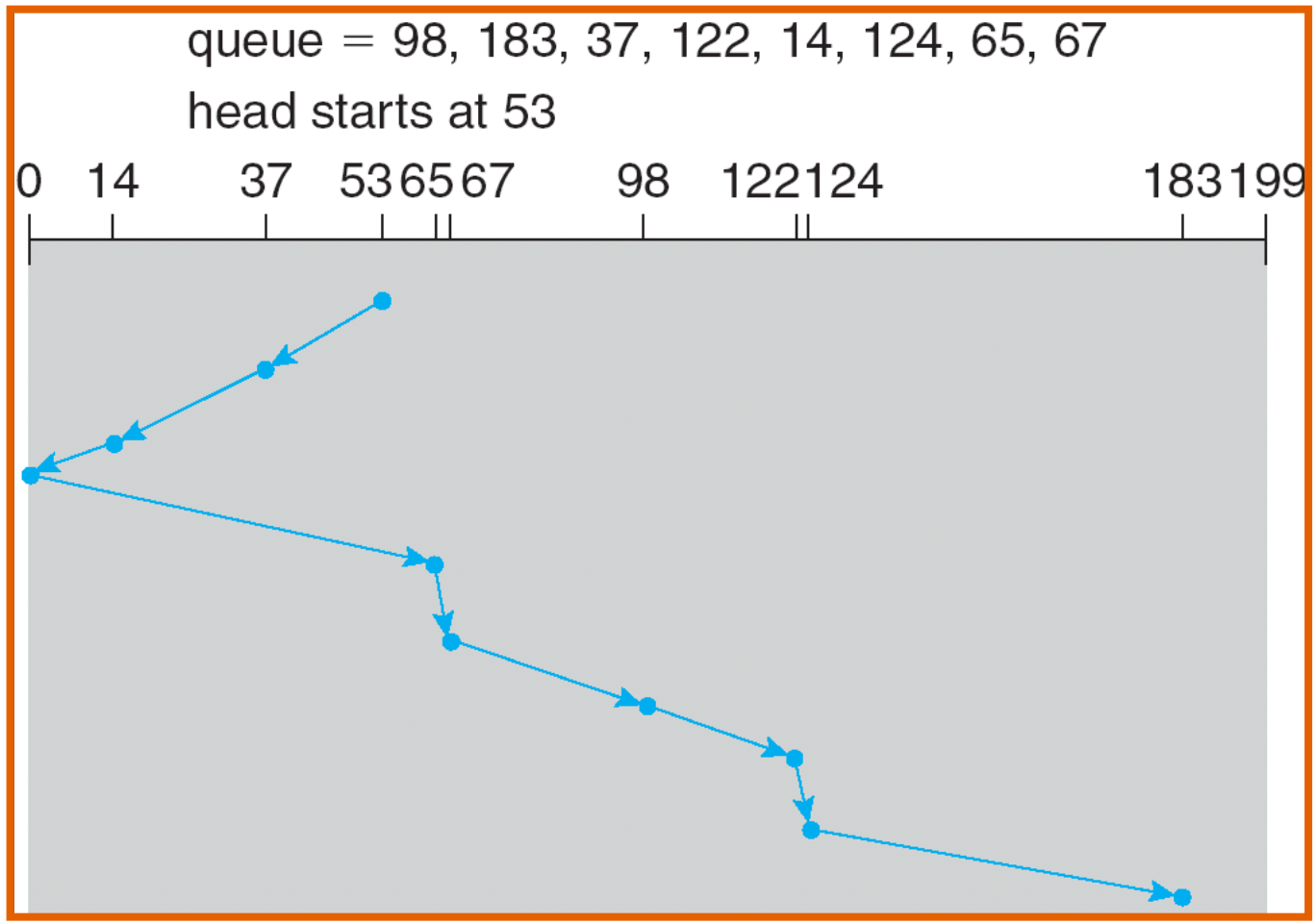


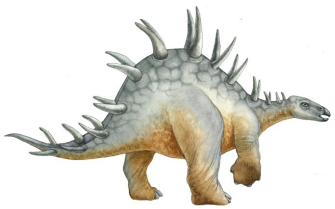
SCAN

- El brazo empieza en un extremo del disco y se mueve hacia el otro extremo, sirviendo solicitudes hasta que llega al otro extremo, donde se voltea el movimiento de la cabeza y sigue sirviendo.
- Algunas veces llamado *algoritmo de elevador*.
- La ilustración muestra el movimiento total de cabeza para 208 cilindros.



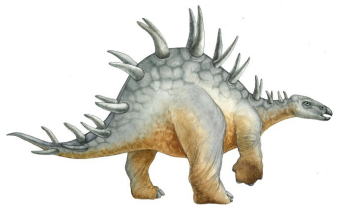
SCAN (Cont.)


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

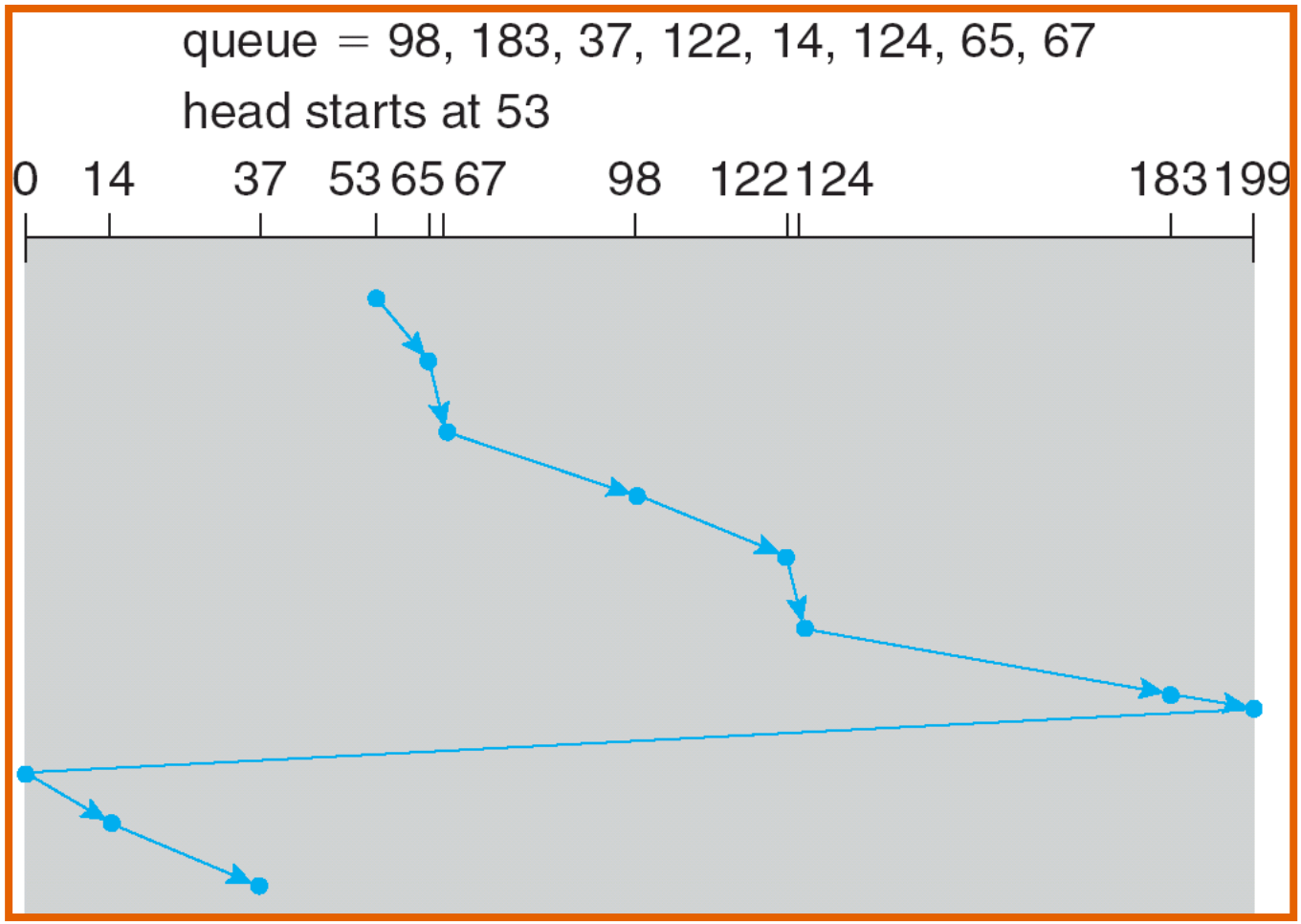


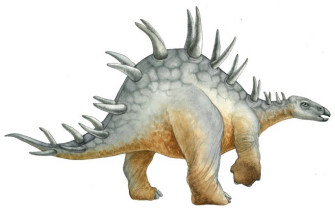
C-SCAN

- Provee un tiempo de espera más uniforme que SCAN.
- La cabeza se mueve de un extremo del disco al otro, sirviendo solicitudes conforme avanza. Cuando llega al otro extremo, sin embargo, regresa al inicio del disco, sin atender ninguna solicitud en el camino de regreso.
- Trata los cilindros como una lista circular que da la vuelta del último cilindro al primero.



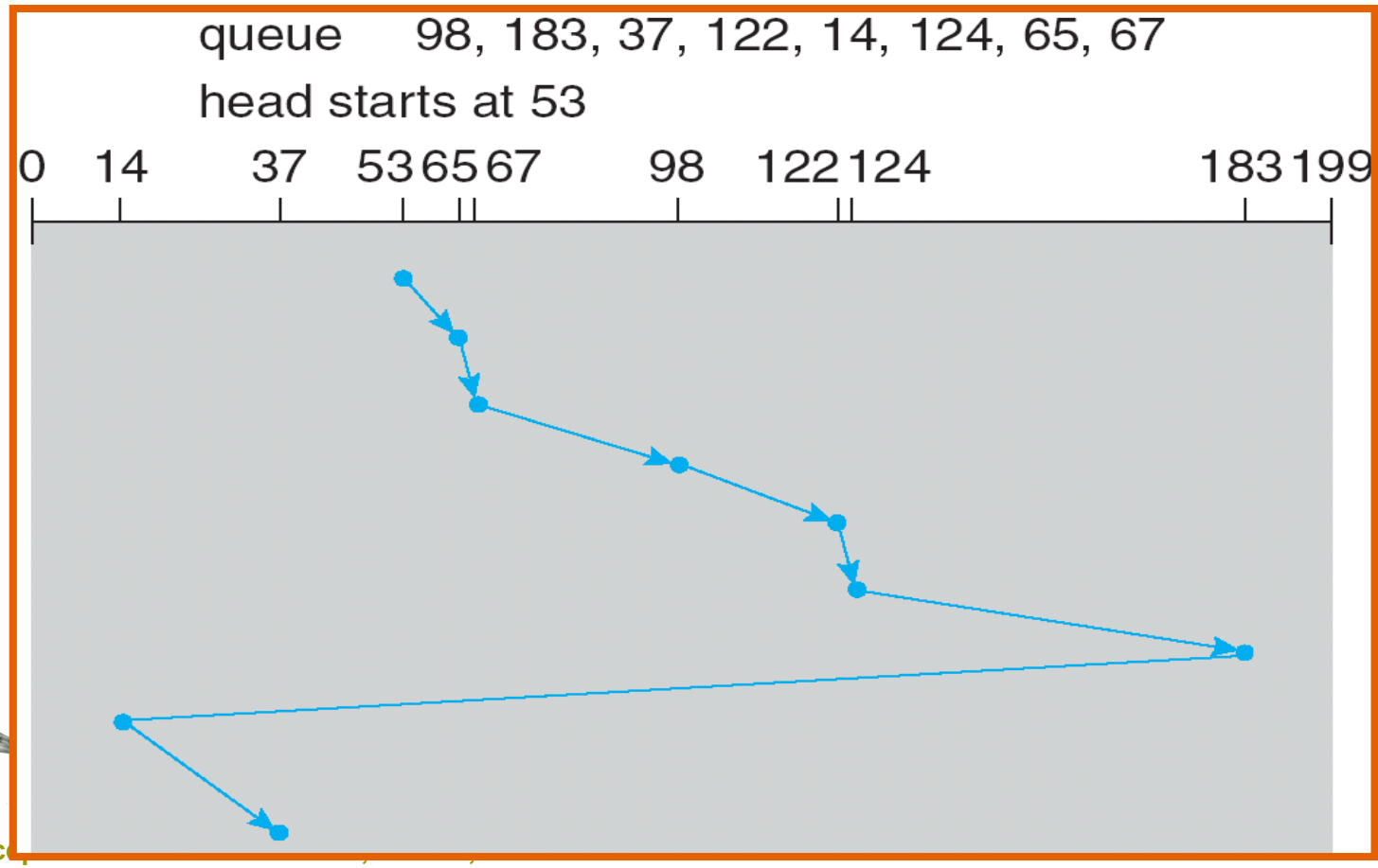
C-SCAN (Cont.)

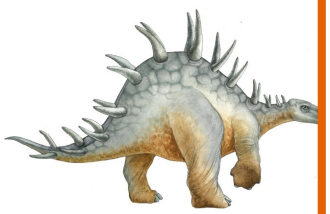

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199



C-LOOK

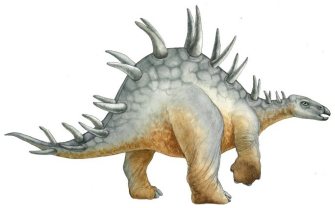
- Versión de C-SCAN
- El brazo sólo va tan lejos como la última solicitud en cada dirección, da reversa inmediatamente, sin antes ir hasta el final del disco.





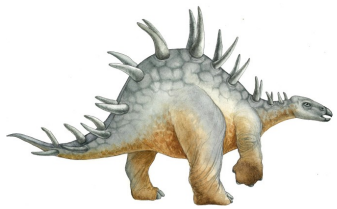
Seleccionado un planificador de disco

- SSTF es común y tiene un *atractivo natural*
- SCAN y C-SCAN ofrecen mejor rendimiento para sistemas que imponen una carga pesada en el disco.
- El rendimiento depende en el número y tipo de solicitudes.
- Solicitudes para servicio de disco pueden ser influenciadas por el método de asignación de archivos.
- El algoritmo de planificación del disco debe estar escrito como un módulo del sistema operativo, permitiendo cambiarlo si es necesario.
- Tanto SSTF o LOOK son un algoritmo razonable para utilizar por omisión.



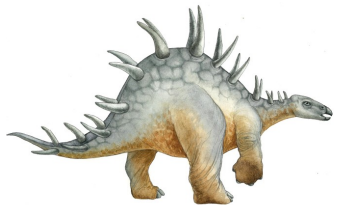
Estructura de RAID

- **RAID** – múltiples discos proveen **confiabilidad** y **redundancia**.
- RAID está dividido en seis niveles.

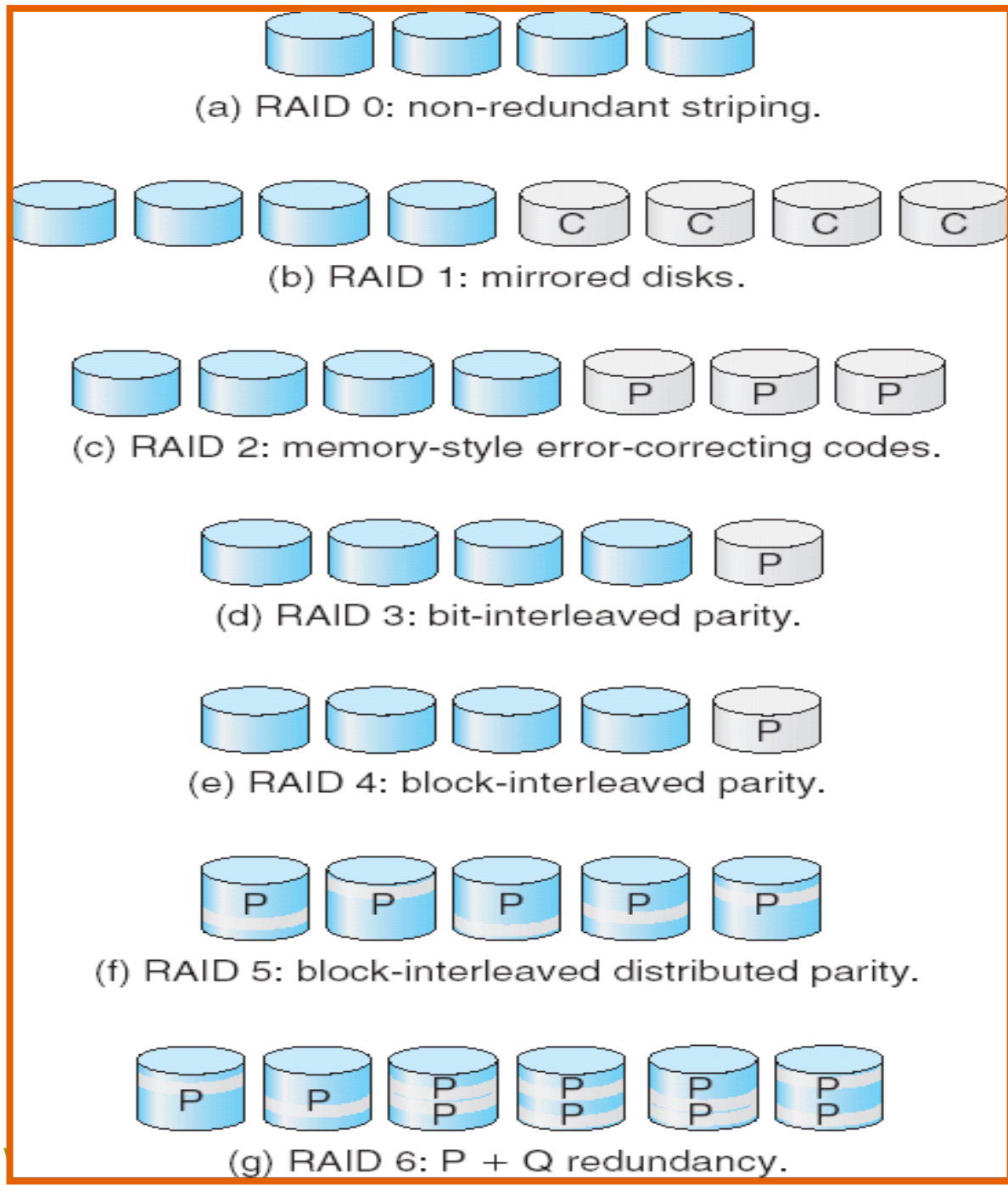


RAID (cont)

- Muchos avances en las técnicas de uso de disco involucran el uso de varios discos trabajando cooperativamente.
- Disk *striping* utiliza un grupo de discos como una unidad de almacenamiento.
- Los esquemas RAID mejoran el rendimiento y confiabilidad del sistema de almacenamiento, guardando datos redundantes.
  - *Mirroring* o *shadowing* mantiene duplicados de cada disco.
  - *Block interleaved parity* utiliza menos redundancia.



Niveles RAID



(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.

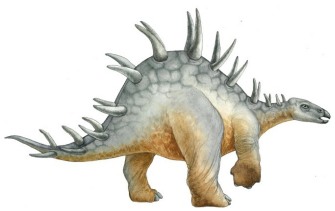
(c) RAID 2: memory-style error-correcting codes.

(d) RAID 3: bit-interleaved parity.

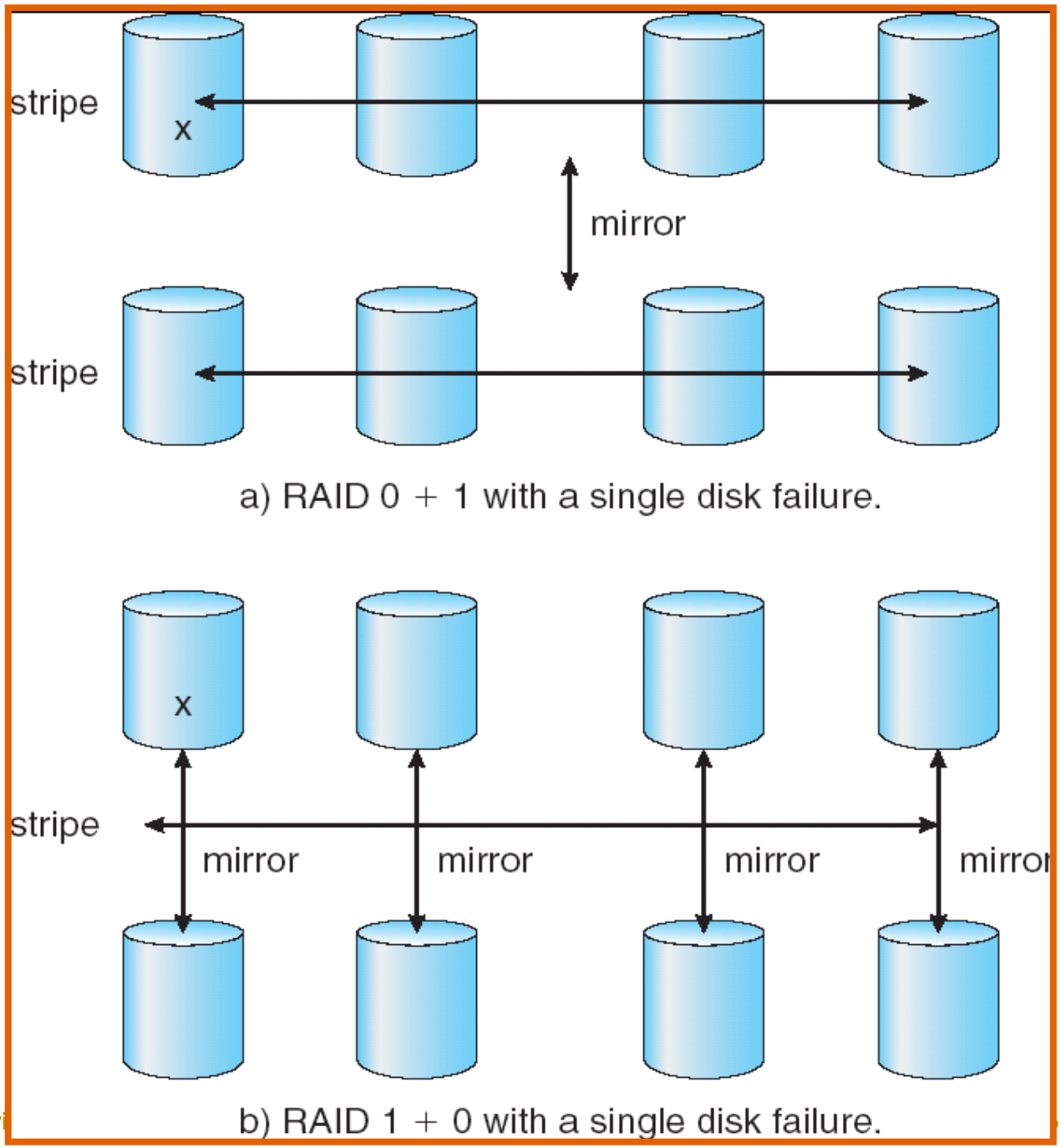
(e) RAID 4: block-interleaved parity.

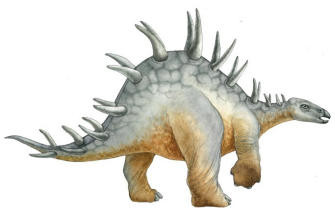
(f) RAID 5: block-interleaved distributed parity.

(g) RAID 6: P + Q redundancy.



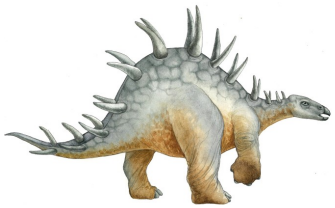
RAID (0 + 1) y (1 + 0)


stripe
x
mirror
stripe
a) RAID 0 + 1 with a single disk failure.
x
stripe
mirror
mirror
mirror
mirror
b) RAID 1 + 0 with a single disk failure.



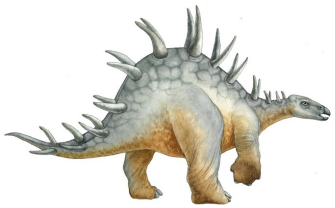
Implementación almacenamiento-estable

- El esquema de bitácora *write-ahead* requiere almacenamiento estable.
- Para implementar almacenamiento estable:
  - Replicar información en más de un medio de almacenamiento no-volátil con modos de fallo independientes.
  - Actualizar información de manera controlada para asegurar que podemos recuperar los datos después de cualquier falla durante la transferencia de datos o recuperación.



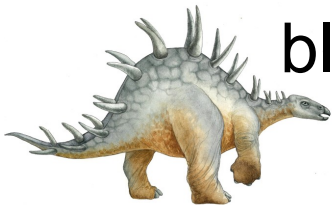
Cintas

- Comparada con un disco, la cinta es menos costosa y almacena más datos, pero el acceso aleatorio es más lento.
- La cinta es un medio económico para propósitos que no requieren un acceso aleatorio rápido, v.gr. copias de respaldo de datos o para mantener enormes volúmenes de datos.
- Instalaciones grandes de cinta típicamente utilizan cambiadores de cinta roboticos, que mueven cintas entre los drives y las ranuras en la biblioteca de cintas.
  - stacker – biblioteca que almacena algunas cintas
  - silo – biblioteca que almacena miles de cintas
- Un archivo que reside en disco, puede ser *archivado* en cinta para almacenamiento de bajo costo; la computadora puede regresarlo a disco para uso activo.



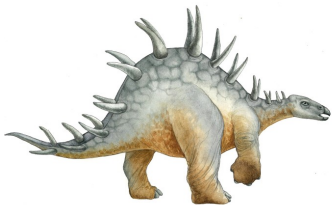
Drives de cinta

- Las operaciones básicas para un drive de cinta difieren de aquellas de un drive de disco.
- **localizar** posiciona la cinta en un bloque lógico específico, no un track entero (corresponde a **seek**).
- La operación **read position** regresa el número de bloque lógico donde está la cabeza de la cinta.
- La operación **space** habilita movimiento relativo.
- Drives de cinta son dispositivos de “sólo añadir”; actualizar un bloque en el medio de una cinta también borra efectivamente todo lo que sigue a ese bloque.
- Una marca EOT se pone después de escribir un bloque.



Manejo de Almacenamiento Jerárquico (HSM)

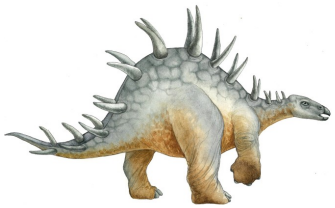
- Un sistema de almacenamiento jerárquico extiende la jerarquía de almacenamiento más allá de memoria principal y almacenamiento secundario para incorporar almacenamiento terciario — usualmente implementado a través de un conjunto de cintas y discos movibles.
- Usualmente se incorpora el almacenamiento terciario extendiendo el sistema de archivos.
  - Archivos pequeños o utilizados frecuentemente se mantienen en el disco.
  - Arcivos grandes viejos o inactivos se mueven al terciario.
- HSM es usualmente encontrado en centros de súper-cómputo y otras instalaciones grandes y con volúmenes enormes de datos.



Velocidad

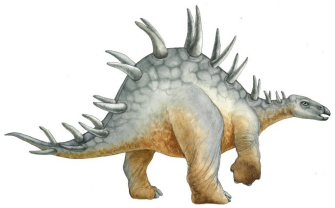
- Dos aspectos de velocidad en el almacenamiento terciario son el ancho de banda y latencia.

- El ancho de banda se mide en bytes por segundo.
  - Ancho de banda sostenido – tasa promedio de transferencia de datos durante una transferencia larga; # de bytes/tiempo de transferencia.
    Tasa de transferencia cuando el stream de datos está fluyendo.
  - Ancho de banda efectivo – promedio durante el tiempo de E/S completo, incluyendo **seek** o **locate**, y cambio de cartuchos.



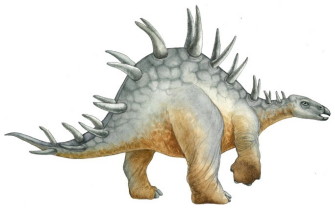
Velocidad

- Latencia de acceso – cantidad de tiempo requerida para localizar datos.
  - Tiempo de acceso para un disco – mover el brazo al cilindro seleccionado y esperar la latencia rotacional; < 35 milisegundos.
  - Acceso en cinta requiere rebobinar el carrete hasta que el bloque seleccionado alcanza la cabeza; decenas de cientos de segundos.
  - En general podemos decir que el tiempo de acceso aleatorio en un carrete de cinta es mil veces más lento que en un disco.
- El bajo costo del almacenamiento terciario es el resultado de tener muchos cartuchos baratos que comparten unos cuantos drives costosos.
- Una biblioteca movible es mejor si se destina al almacenamiento de datos utilizados con poca frecuencia, dado que solo puede atender un pequeño número de solicitudes de E/S por hora.



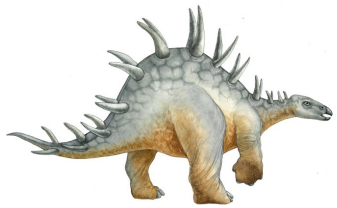
Confiabilidad

- Un drive de disco fijo es con seguridad más confiable que uno disco o cinta movible.
- Un cartucho óptico es probablemente más confiable que un disco o cinta magnética.
- Un choque de cabeza es un disco fijo generalmente destruye los datos, mientras que la falla de un drive de cinta o un disco óptico generalmente no daña los datos.

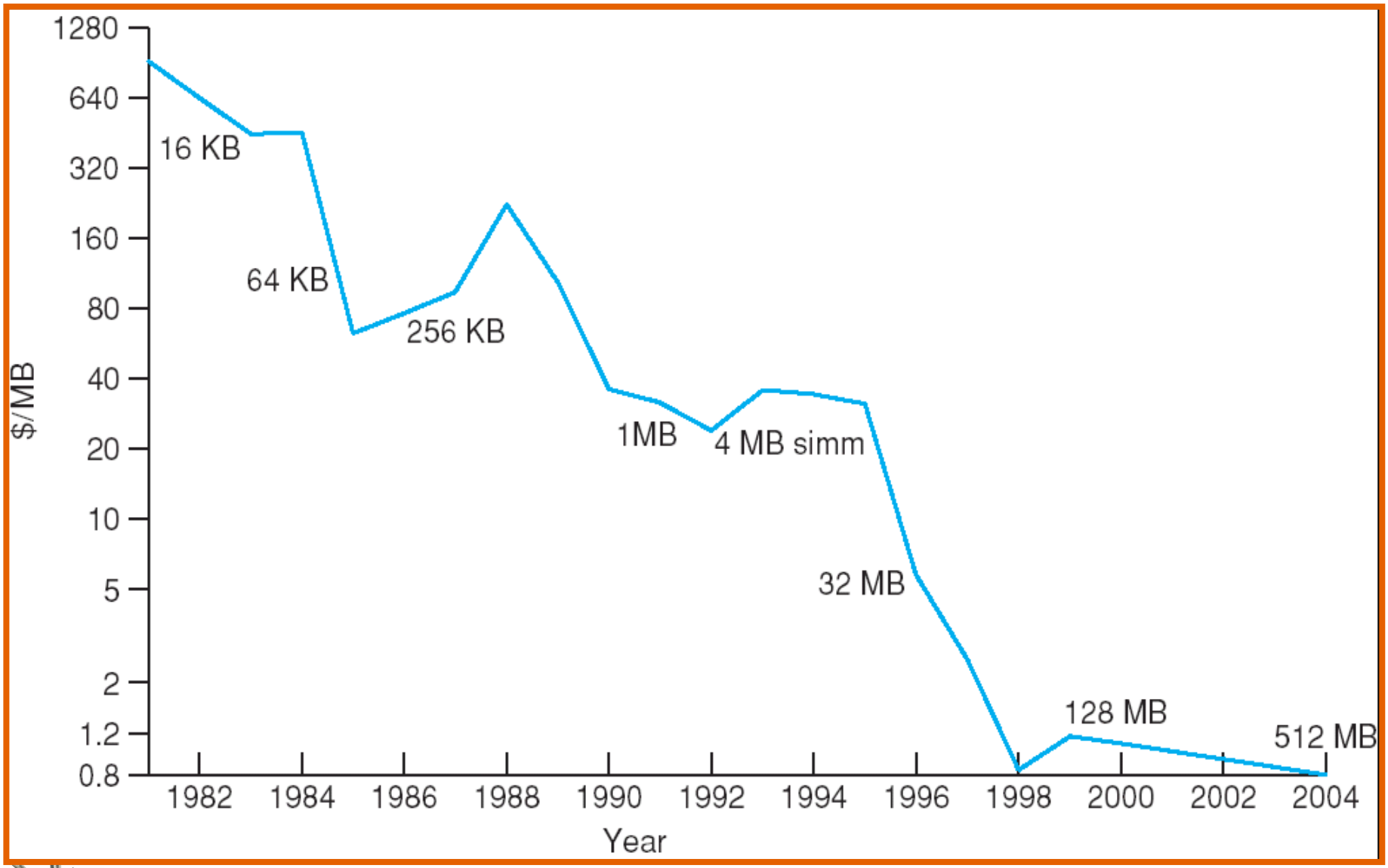


Costo

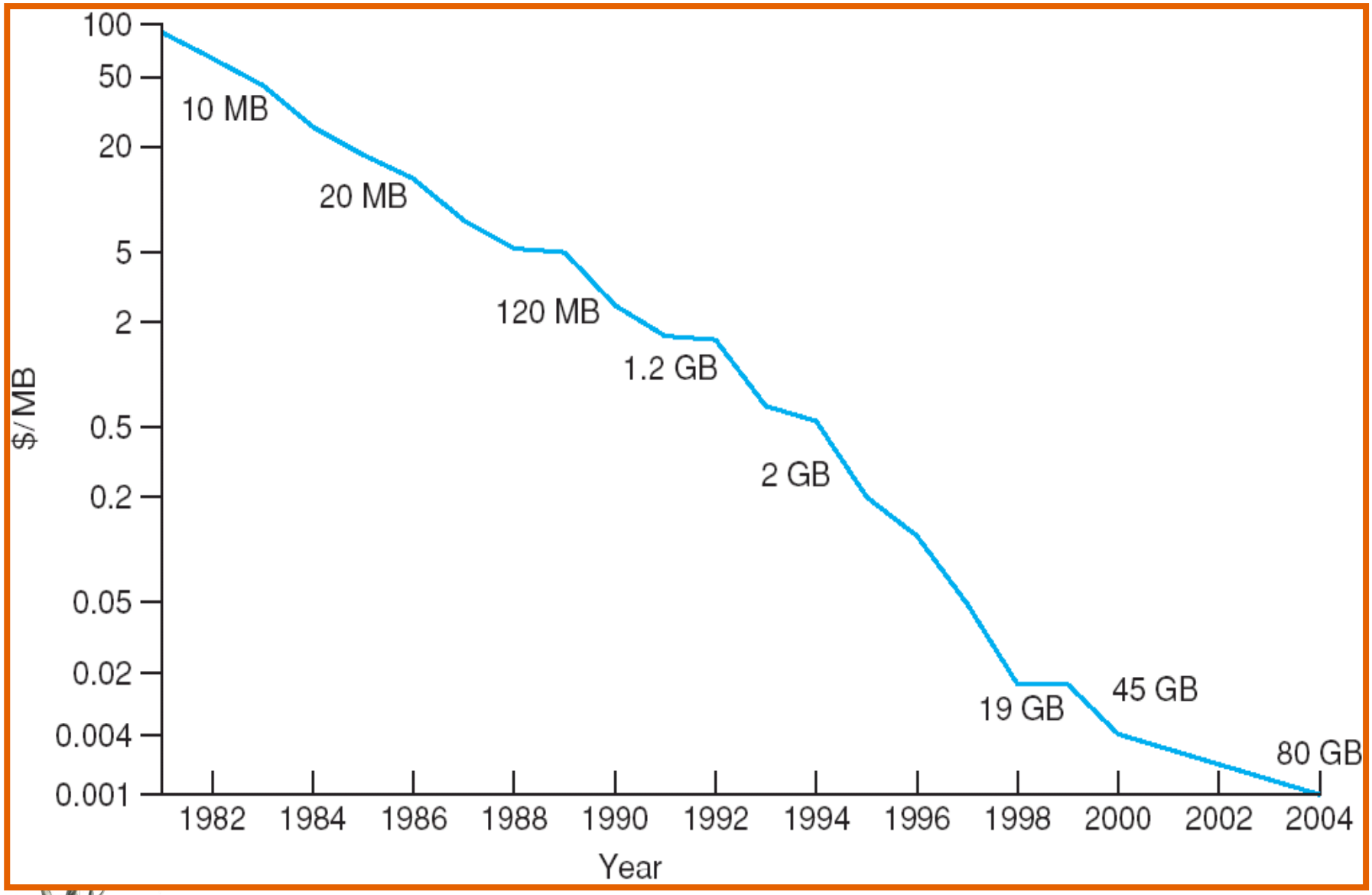
- Memoria principal es mucho más cara que el almacenamiento en disco
- El costo por megabyte del almacenamiento en disco duro compite con el de cintas magnéticas, si se utiliza una cinta por drive.
- Los drives de cinta y disco más baratos han tenido a lo largo de los años la misma capacidad.
- Almacenamiento terciario ofrece ahorros solo cuando el número de cartuchos es considerablemente más grande que el número de drives.



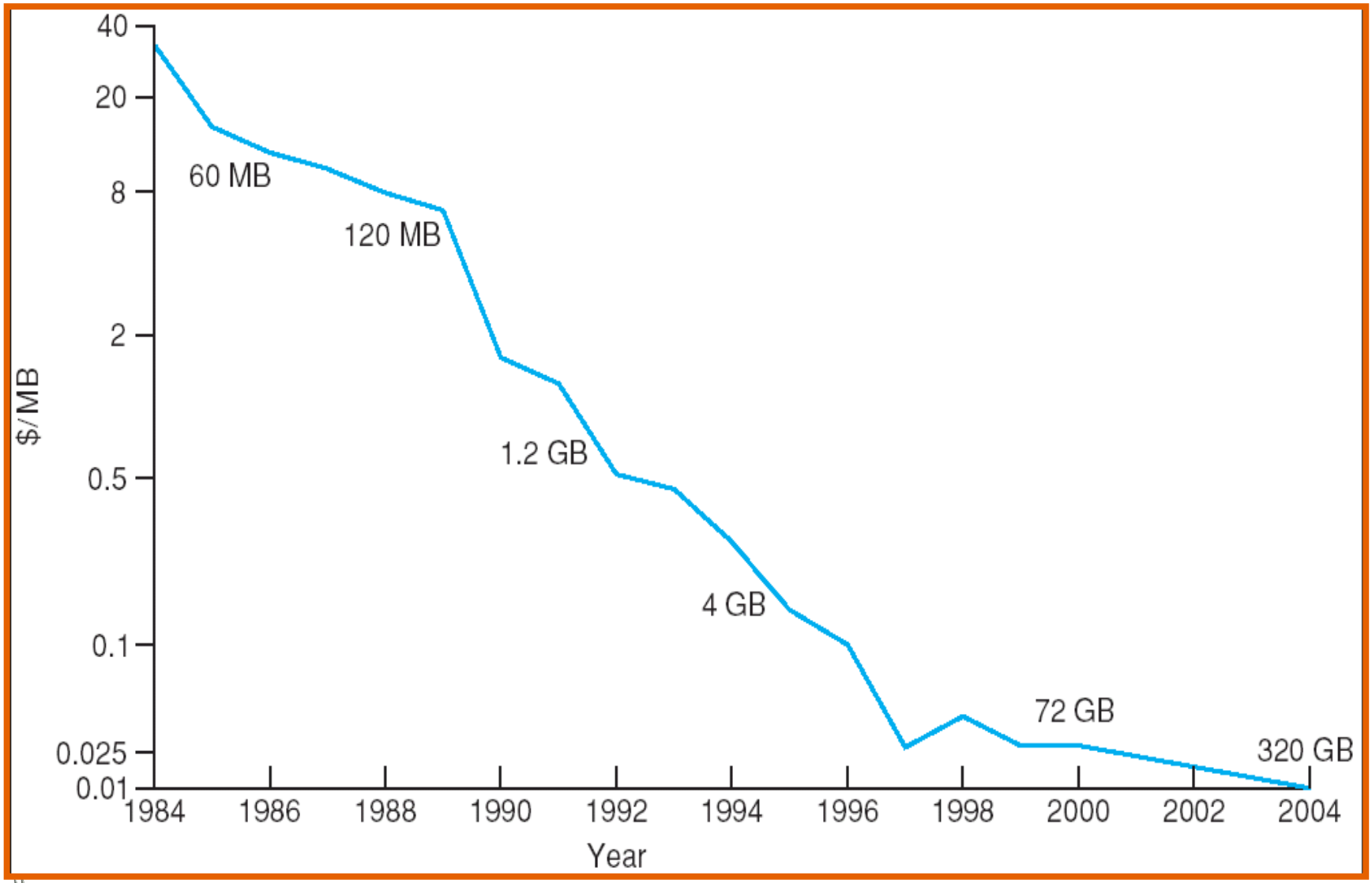
Precio p/Megabyte de DRAM, de 1981 a 2004

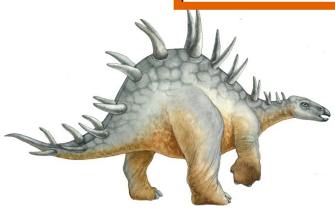


Precio p/Megabyte de DD Magnético, de 1981 a 2004



Precio p/Megabyte de Drive Cinta, de 1984 a 2000





Fin del Capítulo 12