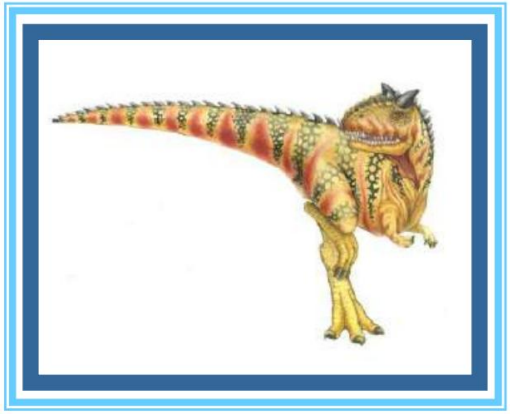
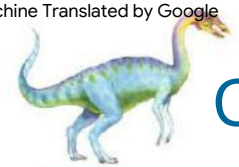
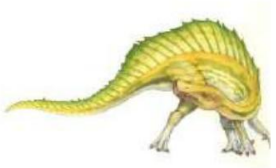
Capítulo 11: Almacenamiento masivo Sistemas

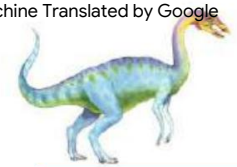




Capítulo 11: Sistemas de almacenamiento masivo

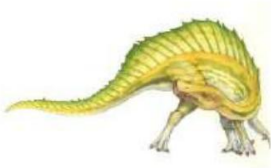
- Descripción general de la estructura de almacenamiento masivo
- Programación de HDD
- Programación de NVM
- Detección y corrección de errores
- Gestión de dispositivos de almacenamiento
- Gestión de espacio de intercambio
- Adjunto de almacenamiento
- Estructura RAID

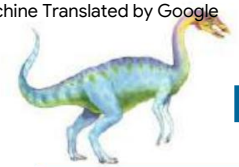




Objetivos

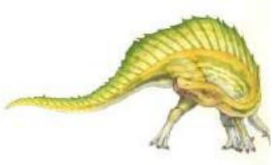
- Describir la estructura física de los dispositivos de almacenamiento secundario y el efecto de la estructura de un dispositivo en sus usos
- Explicar las características de rendimiento de los dispositivos de almacenamiento masivo.
- Evaluar algoritmos de programación de E/S
- Analizar los servicios del sistema operativo proporcionados para almacenamiento masivo, incluyendo RAID

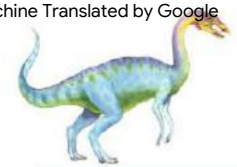




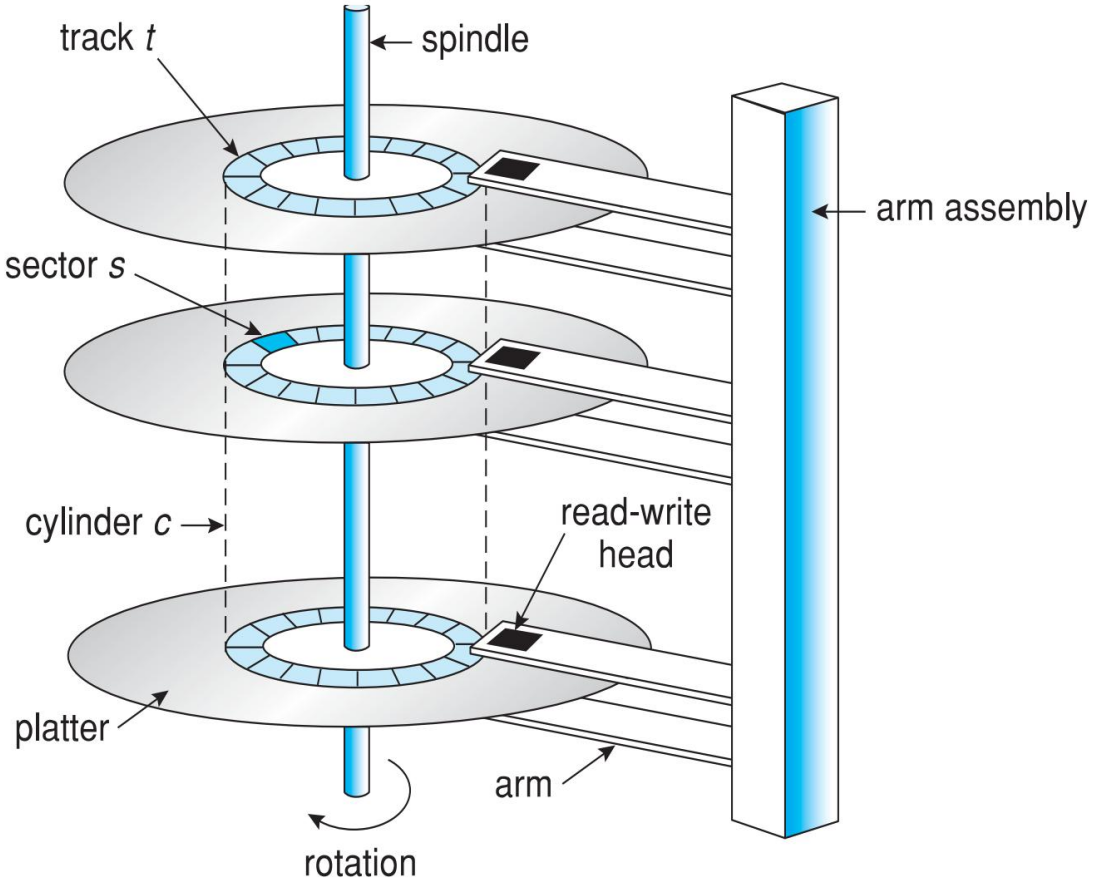
Descripción general de la estructura de almacenamiento masivo

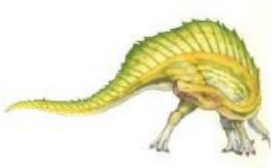
- La mayor parte del almacenamiento secundario de las computadoras modernas son unidades de disco duro. (HDD) y dispositivos de memoria no volátil (NVM)
- Los discos duros hacen girar platos de material recubierto magnéticamente bajo movimiento. cabezas de lectura y escritura
  - Las unidades giran entre 60 y 250 veces por segundo
  - La velocidad de transferencia es la velocidad a la que los datos fluyen entre la unidad y computadora
  - El tiempo de posicionamiento (tiempo de acceso aleatorio) es el tiempo para mover el brazo del disco al cilindro deseado (tiempo de búsqueda) y el tiempo para que el sector deseado gire debajo del cabezal del disco (latencia rotacional)
  - El choque del cabezal se produce cuando el cabezal del disco hace contacto con el disco. superficie - Eso es malo
- Los discos pueden ser extraíbles

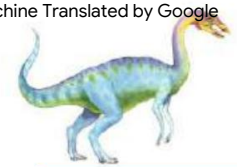




Mecanismo de disco de cabeza móvil



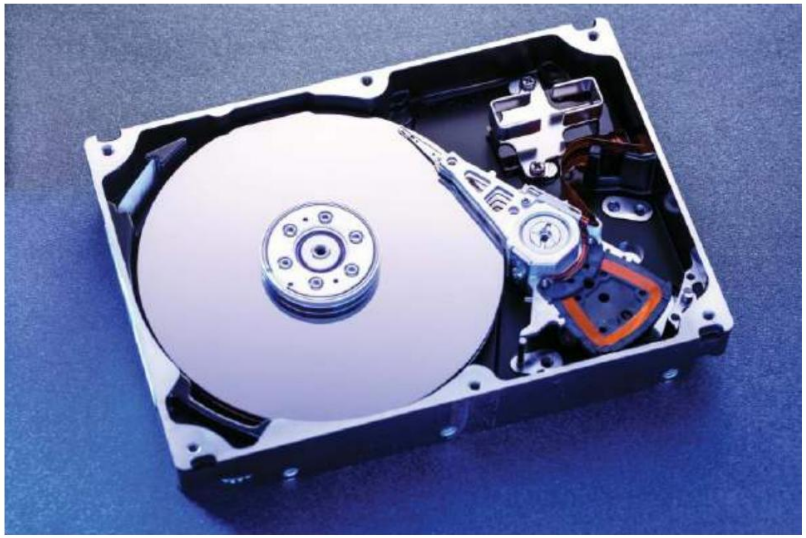


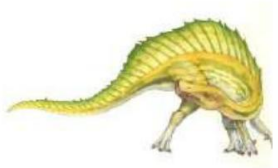


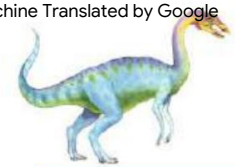
Discos Duros

- Los platos varían de 0,85” a 14” (históricamente)
  - Comúnmente 3,5”, 2,5” y 1,8”
- Rango de 30 GB a 3 TB por unidad
- Rendimiento
  - Tasa de transferencia – teórica – 6 Gb/seg
  - Tasa de transferencia efectiva – real – 1 Gb/s
  - Tiempo de búsqueda de 3 ms a 12 ms – 9 ms común para unidades de escritorio
  - Tiempo de búsqueda promedio medido o calculado en base a 1/3 de las pistas
  - Latencia basada en la velocidad del husillo

    1 / (RPM / 60) = 60 / RPM
  - Latencia promedio = ½ latencia

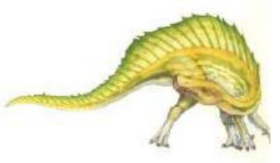


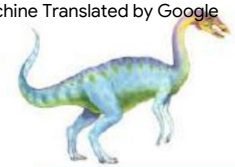




Rendimiento del disco duro

- Latencia de acceso = Tiempo promedio de acceso = tiempo promedio de búsqueda + latencia media
  - Para disco más rápido 3 ms + 2 ms = 5 ms
  - Para disco lento 9 ms + 5,56 ms = 14,56 ms
- Tiempo promedio de E/S = tiempo promedio de acceso + (cantidad a transferir / tasa de transferencia) + gastos generales del controlador
- Por ejemplo, para transferir un bloque de 4 KB en un disco de 7200 RPM con un tiempo de búsqueda promedio de 5 ms, velocidad de transferencia de 1 Gb/s con una sobrecarga de controlador de 0,1 ms =
  - 5 ms + 4,17 ms + 0,1 ms + tiempo de transferencia =
  - Tiempo de transferencia = 4 KB / 1 Gb/s * 8 Gb / GB * 1 GB / 10242 KB = 32 / (10242) = 0,031 ms
  - Tiempo promedio de E/S para bloque de 4 KB = 9,27 ms + 0,031 ms = 9,301 ms





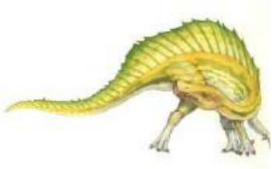
La primera unidad de disco comercial

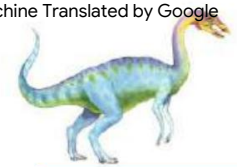
1956

La computadora IBM RAMDAC incluía el sistema de almacenamiento en disco IBM Modelo 350

5 millones de caracteres (7 bits)

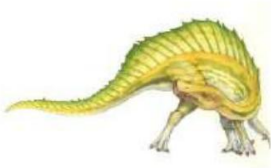
Platos de 50 x 24”
Tiempo de acceso = < 1 segundo

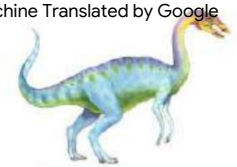




Dispositivos de memoria no volátil

- Si son unidades de disco, se denominan discos de estado sólido (SSD).
- Otras formas incluyen unidades USB (memoria USB, unidad flash), reemplazos de discos DRAM, montaje en superficie en placas base y almacenamiento principal en dispositivos como teléfonos inteligentes.
- Puede ser más confiable que los HDD
- Más caro por MB
- Quizás tengan una vida útil más corta; necesitan una gestión cuidadosa
- Menos capacidad
- Pero mucho más rápido
- Los buses pueden ser demasiado lentos -> conectarse directamente a PCI, por ejemplo
- Sin piezas móviles, por lo que no hay tiempo de búsqueda ni latencia de rotación



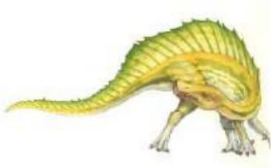


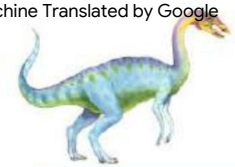
Dispositivos de memoria no volátil

- Tener características que presenten desafíos.

- Leer y escribir en incrementos de "página" (piense en el sector) pero no se puede sobrescribir en el lugar
  - Primero debe borrarse y los borrados se realizan en incrementos de "bloques" más grandes
  - Sólo se puede borrar una cantidad limitada número de veces antes de desgastarse – ~ 100.000
  - Vida útil medida en unidad escrituras por día (DWPD)

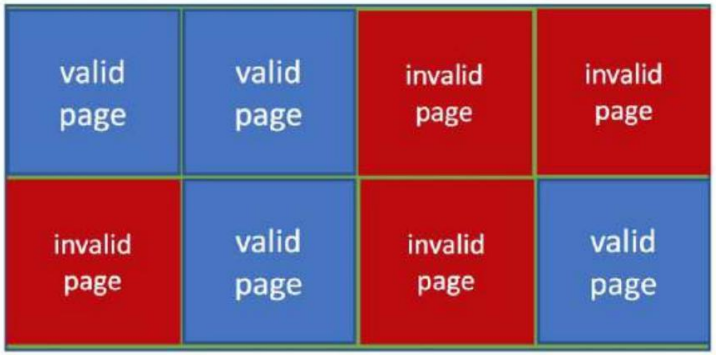
    Se espera que una unidad NAND de 1 TB con una clasificación de 5DWPD tenga 5 TB por día escritos dentro del período de garantía sin fallar



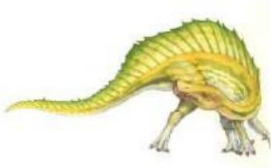


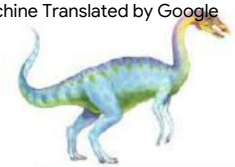
Algoritmos del controlador de flash NAND

- Sin sobrescritura, las páginas terminan con una combinación de datos válidos e inválidos.
- Para rastrear qué bloques lógicos son válidos, el controlador mantiene la memoria flash tabla de capa de traducción (FTL)
- También implementa la recolección de basura para liberar espacio en páginas no válidas.
- Asigna sobreaprovisionamiento para proporcionar espacio de trabajo para GC
- Cada celda tiene una vida útil, por lo que es necesario nivelar el desgaste para escribir por igual en todas células



Bloque NAND con páginas válidas e inválidas

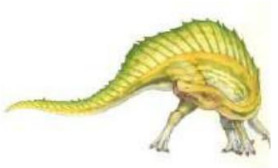


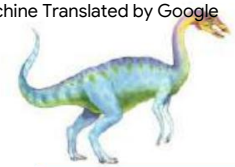


Memoria volatil

- DRAM utilizada frecuentemente como dispositivo de almacenamiento masivo
  - Técnicamente no es almacenamiento secundario porque es volátil, pero puede tener sistemas de archivos y usarse como almacenamiento secundario muy rápido.
- Unidades RAM (con muchos nombres, incluidos discos RAM) presentes sin formato dispositivos de bloqueo, comúnmente formateados por sistemas de archivos
- Las computadoras tienen buffer, almacenamiento en caché a través de RAM, entonces ¿por qué unidades de RAM?
  - Cachés/búferes asignados/administrados por el programador, sistema operativo, hardware
  - Unidades RAM bajo control del usuario
  - Se encuentra en todos los principales sistemas operativos.

    Linux /dev/ram, macOS diskutil para crearlos, Linux /tmp del tipo de sistema de archivos tmpfs
- Utilizado como almacenamiento temporal de alta velocidad.
  - Los programas podrían compartir datos masivos, rápidamente, leyendo/escribiendo en unidad de RAM



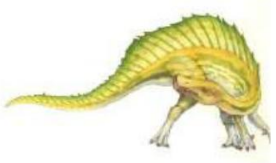


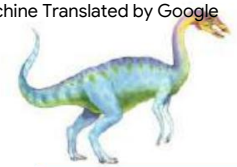
Cinta magnética

Magnetic tape was used as an early secondary-storage medium. Although it is nonvolatile and can hold large quantities of data, its access time is slow compared with that of main memory and drives. In addition, random access to magnetic tape is about a thousand times slower than random access to HDDs and about a hundred thousand times slower than random access to SSDs so tapes are not very useful for secondary storage. Tapes are used mainly for backup, for storage of infrequently used information, and as a medium for transferring information from one system to another.

A tape is kept in a spool and is wound or rewound past a read–write head. Moving to the correct spot on a tape can take minutes, but once positioned, tape drives can read and write data at speeds comparable to HDDs. Tape capacities vary greatly, depending on the particular kind of tape drive, with current capacities exceeding several terabytes. Some tapes have built-in compression that can more than double the effective storage. Tapes and their drivers are usually categorized by width, including 4, 8, and 19 millimeters and 1/4 and 1/2 inch. Some are named according to technology, such as LTO-6 (Figure 11.5) and SDLT.

Figure 11.5 An LTO-6 Tape drive with tape cartridge inserted.



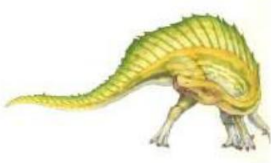


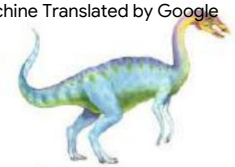
Accesorio de disco

- Almacenamiento conectado al host al que se accede a través de puertos de E/S que se comunican con buses de E/S
- Varios autobuses disponibles, incluido el accesorio de tecnología avanzada (ATA), serial ATA (SATA), eSATA, SCSI conectado en serie (SAS), bus serie universal (USB) y canal de fibra (FC).
- El más común es SATA
- Debido a que NVM es mucho más rápido que HDD, nueva interfaz rápida para NVM llamada NVM express (NVMe), que se conecta directamente al bus PCI
- Transferencias de datos en un bus realizadas por procesadores electrónicos especiales llamados controladores (o adaptadores de bus host, HBA)
  - Controlador de host en el extremo de la computadora del bus, controlador de dispositivo en extremo del dispositivo
  - La computadora coloca el comando en el controlador host, usando puertos de E/S asignados en memoria

    El controlador de host envía mensajes al controlador del dispositivo

    Datos transferidos a través de DMA entre el dispositivo y la computadora DRAM





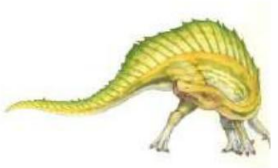
Mapeo de direcciones

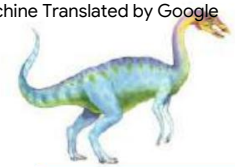
- Las unidades de disco se abordan como grandes matrices unidimensionales de lógica bloques, donde el bloque lógico es la unidad de transferencia más pequeña
  - El formateo de bajo nivel crea bloques lógicos en medios físicos.
- La matriz unidimensional de bloques lógicos se asigna secuencialmente a los sectores del disco.
  - El sector 0 es el primer sector de la primera vía en el extremo exterior.

    cilindro
  - El mapeo continúa en orden a través de esa pista, luego el resto de las pistas en ese cilindro y luego a través del resto de los cilindros desde el más externo al más interno.
  - La dirección lógica a física debería ser fácil

    Excepto sectores malos

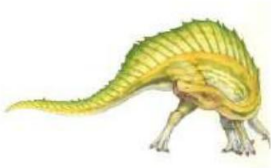
    Número no constante de sectores por pista mediante angular constante velocidad

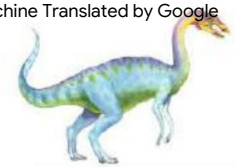




Programación de disco duro

- El sistema operativo es responsable de utilizar el hardware de manera eficiente.
  Para las unidades de disco, esto significa tener un tiempo de acceso y un ancho de banda de disco rápidos.
- Minimizar el tiempo de búsqueda
- Buscar tiempo buscar distancia
- El ancho de banda del disco es el número total de bytes transferidos, dividido por el tiempo total entre la primera solicitud de servicio y la finalización de la última transferencia

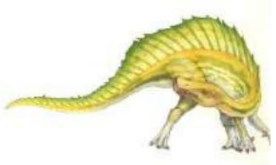


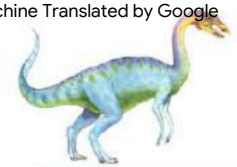


Programación de disco (cont.)

- Hay muchas fuentes de solicitud de E/S de disco.
  - SO
  - Procesos del sistema
  - Procesos de usuarios
- La solicitud de E/S incluye modo de entrada o salida, dirección de disco, memoria dirección, número de sectores a transferir
- El sistema operativo mantiene una cola de solicitudes, por disco o dispositivo
- El disco inactivo puede funcionar inmediatamente ante una solicitud de E/S, el disco ocupado significa trabajo debe hacer cola
  - Los algoritmos de optimización sólo tienen sentido cuando existe una cola
- En el pasado, el sistema operativo responsable de la gestión de colas, la programación de cabezales de unidades de disco
  - Ahora, integrados en los dispositivos de almacenamiento, controladores
  - Simplemente proporcione LBA y maneje la clasificación de solicitudes.

Algunos de los algoritmos que utilizan se describen a continuación



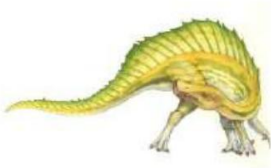


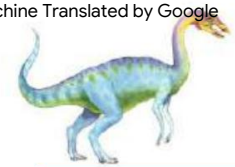
Programación de disco (cont.)

- Tenga en cuenta que los controladores de unidades tienen buffers pequeños y pueden administrar una cola de solicitudes de E/S (de “profundidad” variable)
- Existen varios algoritmos para programar el servicio de solicitudes de E/S de disco.
- El análisis es válido para uno o varios platos.
- Ilustramos algoritmos de programación con una cola de solicitudes (0-199)

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

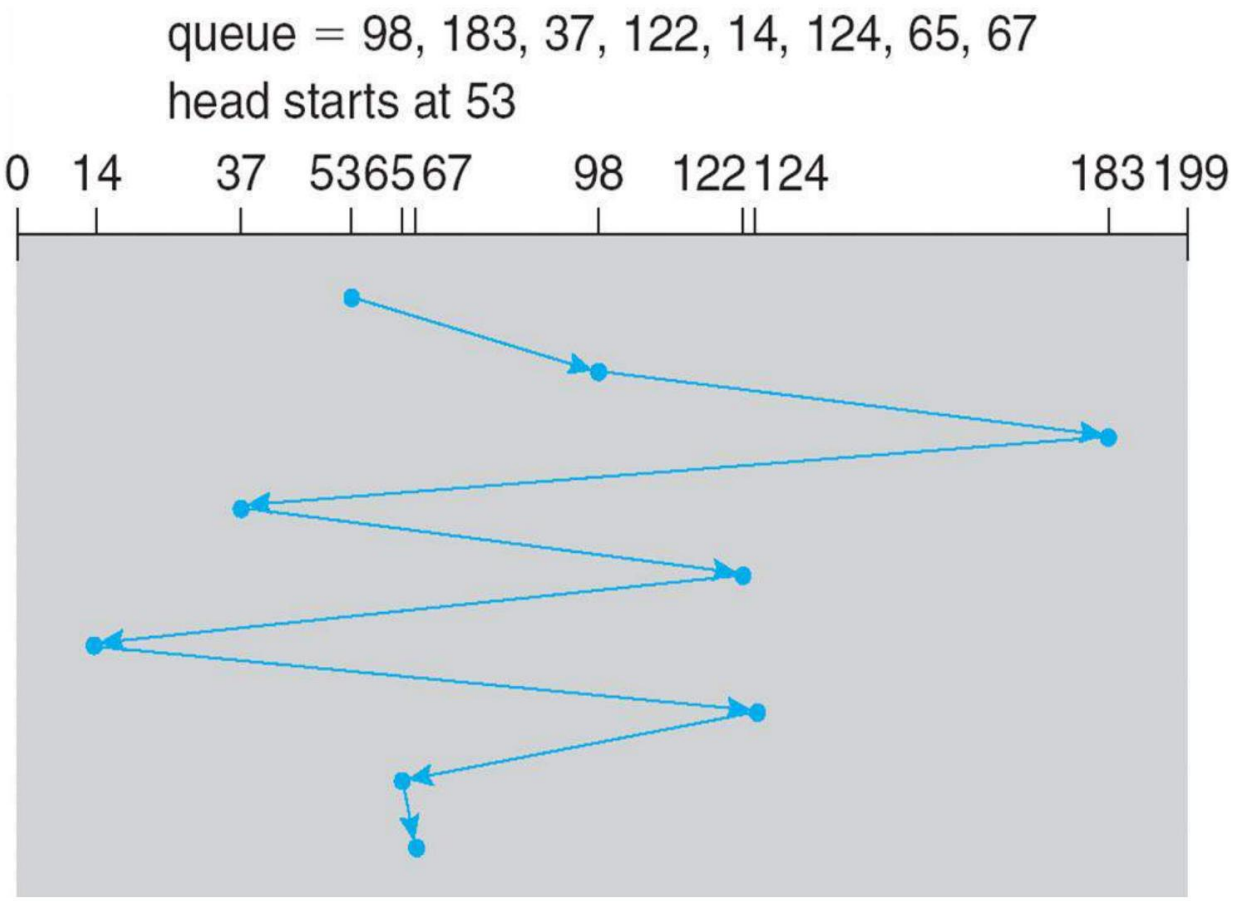
Puntero de cabeza 53

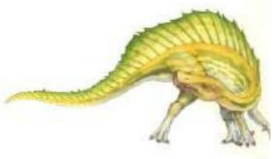


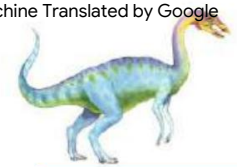


FCFS

La ilustración muestra el movimiento total del cabezal de 640 cilindros.

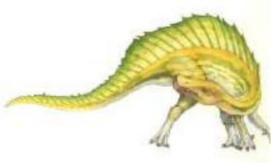


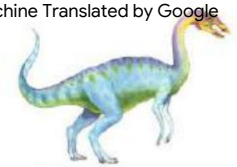




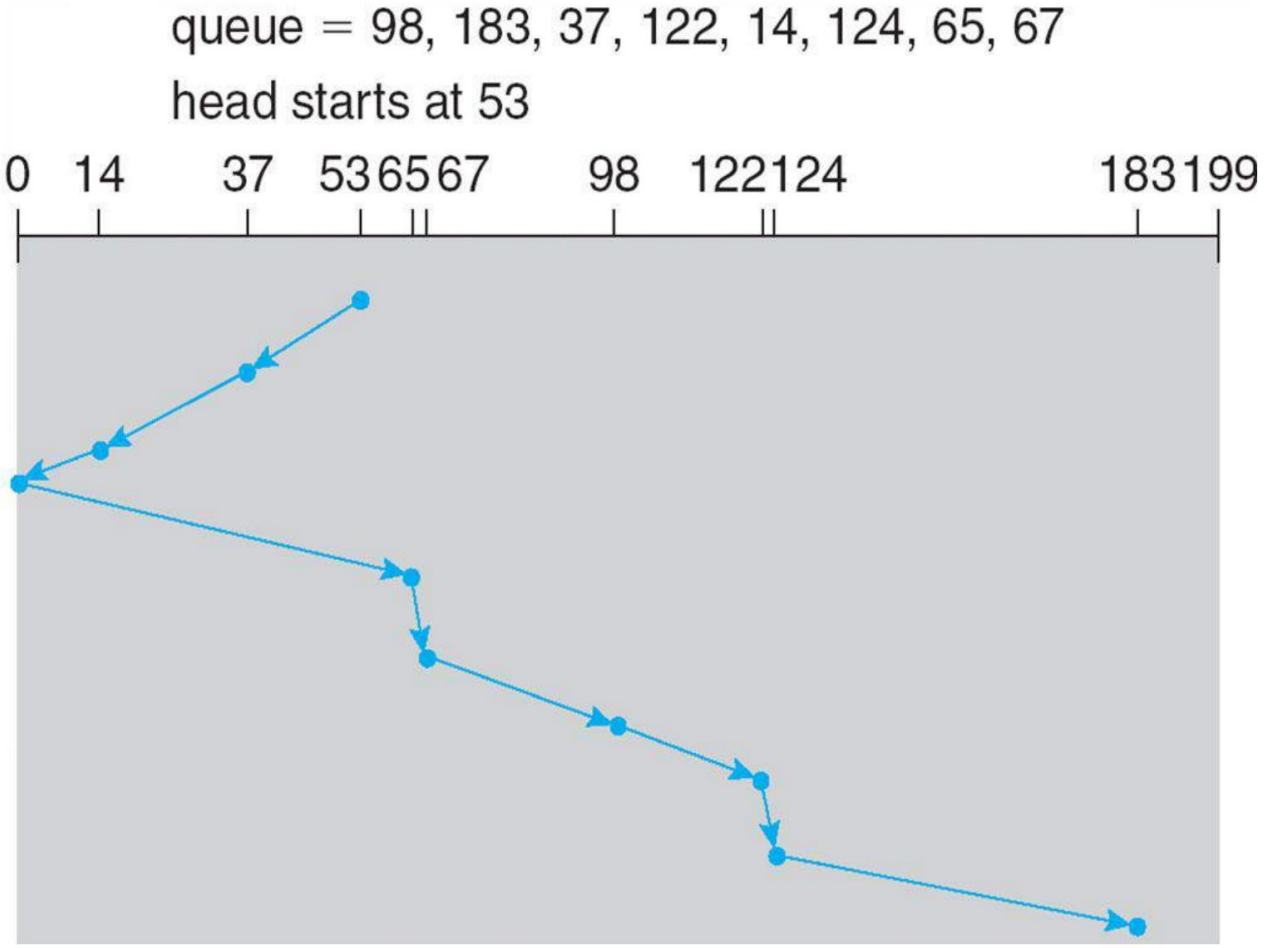
ESCANEAR

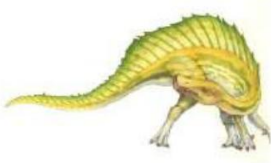
- El brazo del disco comienza en un extremo del disco y se mueve hacia el otro extremo, solicita servicio hasta llegar al otro extremo del disco, donde se invierte el movimiento del cabezal y continúa el servicio.
- Algoritmo SCAN A veces llamado algoritmo del ascensor.
- La ilustración muestra el movimiento total del cabezal de 208 cilindros.
- Pero tenga en cuenta que si las solicitudes son uniformemente densas, la mayor densidad en otros final del disco y los que esperan más tiempo

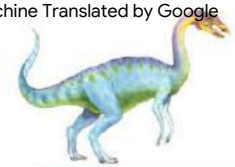




ESCANEAR (Cont.)

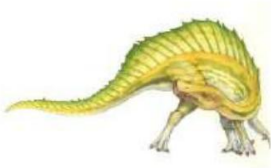


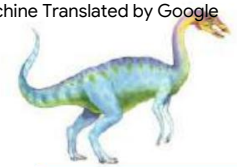




C-SCAN

- Proporciona un tiempo de espera más uniforme que SCAN
- El cabezal se mueve de un extremo del disco al otro, atendiendo las solicitudes a medida que avanza.
  - Sin embargo, cuando llega al otro extremo, regresa inmediatamente al principio del disco, sin atender ninguna solicitud en el viaje de regreso.
- Trata los cilindros como una lista circular que comienza desde el último cilindro al primero
- ¿ Número total de cilindros?

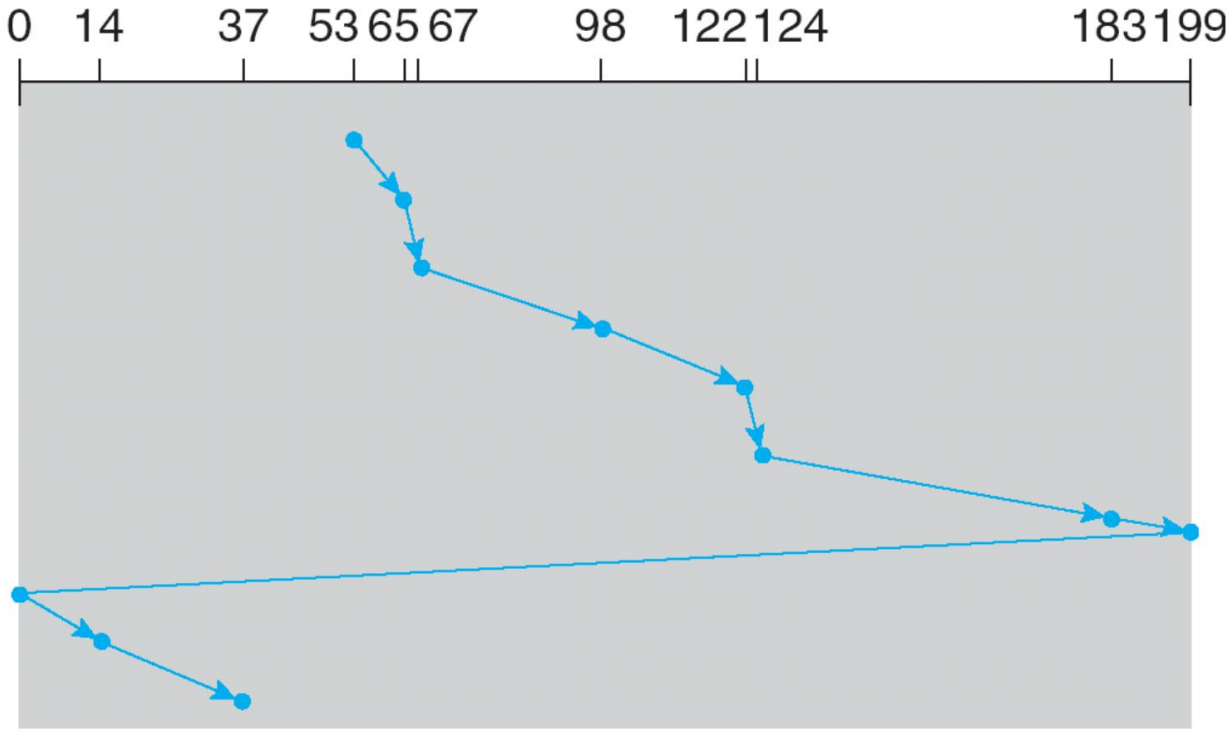


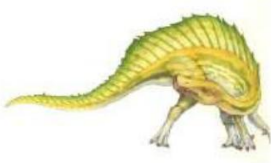


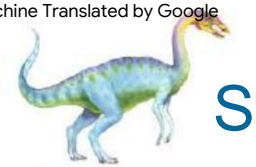
C-SCAN (Cont.)

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

head starts at 53







Seleccionar un algoritmo de programación de discos

- SSTF es común y tiene un atractivo natural.
- SCAN y C-SCAN funcionan mejor para sistemas que colocan una carga pesada en el disco
  - Menos hambre, pero aún es posible
- Para evitar el hambre, Linux implementa un programador de fechas límite.
  - Mantiene colas de lectura y escritura separadas, otorga prioridad de lectura

    Porque es más probable que los procesos se bloqueen en lectura que en escritura.
  - Implementa cuatro colas: 2 x lectura y 2 x escritura

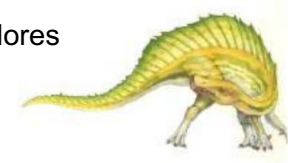
    1 cola de lectura y 1 cola de escritura ordenadas en orden LBA, esencialmente implementando C-SCAN

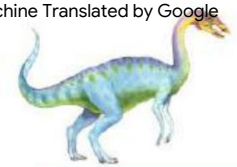
    1 cola de lectura y 1 cola de escritura ordenadas en orden FCFS

    Todas las solicitudes de E/S enviadas en lotes ordenadas en el orden de esa cola

    Después de cada lote, verifica si alguna solicitud en FCFS tiene una antigüedad mayor a la configurada (predeterminado 500 ms)

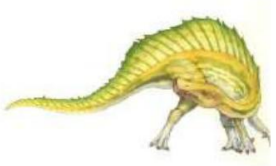
    – Si es así, la cola LBA que contiene esa solicitud se selecciona para el siguiente lote de E/S
- En RHEL 7 también están disponibles NOOP y el programador de colas completamente justo (CFQ) , los valores predeterminados varían según el dispositivo de almacenamiento.

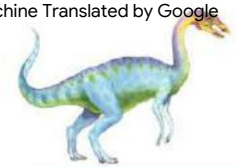




Programación NVM

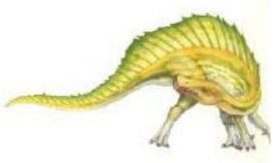
- Sin cabezales de disco ni latencia rotacional, pero aún hay margen de optimización
- En RHEL 7 se utiliza NOOP (sin programación), pero se utilizan solicitudes LBA adyacentes. están combinados
  - NVM es mejor en E/S aleatorias, HDD en secuencial
  - El rendimiento puede ser similar
  - Operaciones de entrada/salida por segundo (IOPS) mucho mayores con NVM (cientos de miles frente a cientos)
  - Pero la amplificación de escritura (una escritura, lo que provoca recolección de basura y muchas lecturas/escrituras) puede disminuir la ventaja de rendimiento.

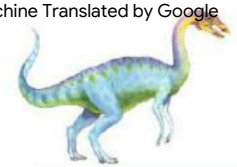




Detección y corrección de errores

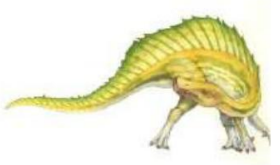
- Aspecto fundamental de muchas partes de la informática (memoria, redes, almacenamiento)
- La detección de errores determina si ha ocurrido un problema (por ejemplo un poco volteado)
  - Si se detecta, puede detener la operación.
  - Detección realizada frecuentemente mediante bit de paridad
- Paridad de una forma de suma de verificación : utiliza aritmética modular para calcular, almacenar y comparar valores de palabras de longitud fija.
  - Otro método de detección de errores común en redes es el cíclico. verificación de redundancia (CRC) que utiliza la función hash para detectar errores de múltiples bits
- El código de corrección de errores (ECC) no sólo detecta, sino que puede corregir algunos errores
  - Errores leves corregibles, errores graves detectados pero no corregidos

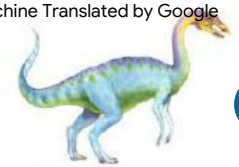




Gestión de dispositivos de almacenamiento

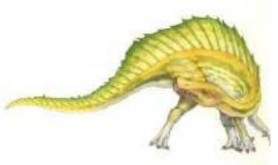
- Formateo de bajo nivel o formato físico : dividir un disco en Sectores que el controlador de disco puede leer y escribir.
  - Cada sector puede contener información de encabezado, más datos y más errores.
    código de corrección (ECC)
  - Generalmente 512 bytes de datos, pero se pueden seleccionar
- Para utilizar un disco para almacenar archivos, el sistema operativo aún necesita registrar sus propias estructuras de datos en el disco.
  - Particionar el disco en uno o más grupos de cilindros, cada uno tratado como un disco lógico.
  - Formateo lógico o "creación de un sistema de archivos"
  - Para aumentar la eficiencia, la mayoría de los sistemas de archivos agrupan bloques en clústeres.
    E/S de disco realizadas en bloques
    E/S de archivos realizada en clústeres

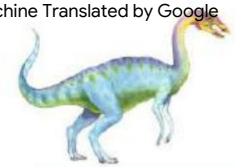




Gestión de dispositivos de almacenamiento (cont.)

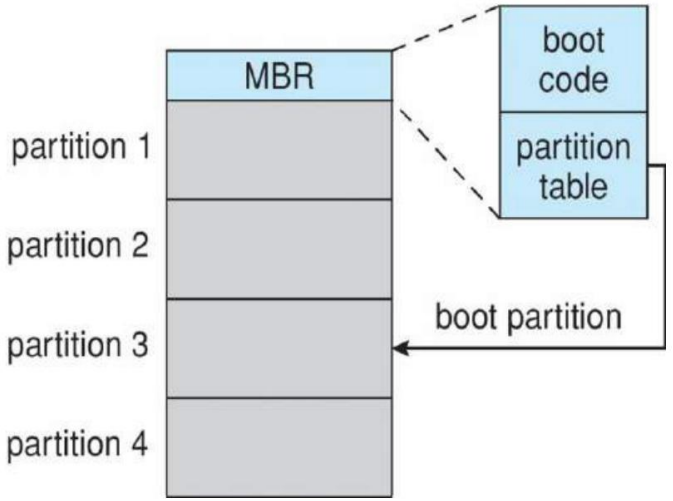
- La partición raíz contiene el sistema operativo, otras particiones pueden contener otras Oses, otros sistemas de archivos o sin formato
  - Montado en el momento del arranque
  - Otras particiones pueden montarse automática o manualmente
- En el momento del montaje, se verifica la coherencia del sistema de archivos.
  - ¿ Son correctos todos los metadatos?
    - Si no, arréglalo, vuelve a intentarlo.
    - En caso afirmativo, agregue a la tabla de montaje, permita el acceso
- El bloque de inicio puede apuntar al volumen de inicio o al conjunto de bloques del cargador de inicio que contienen suficiente código para saber cómo cargar el kernel desde el sistema de archivos.
  - O un programa de administración de arranque para el arranque de múltiples sistemas operativos.



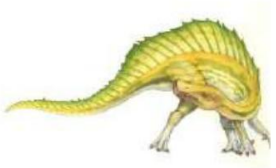


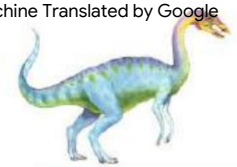
Gestión de almacenamiento de dispositivos (cont.)

- Acceso al disco sin procesar para aplicaciones que desean realizar su propia administración de bloques, manteniendo el sistema operativo fuera del camino (bases de datos, por ejemplo)
- El bloque de arranque inicializa el sistema.
  - El bootstrap se almacena en ROM, firmware
  - Programa de carga Bootstrap almacenado en bloques de arranque de la partición de arranque
- Métodos como el ahorro de sectores utilizado para manejar bloques defectuosos



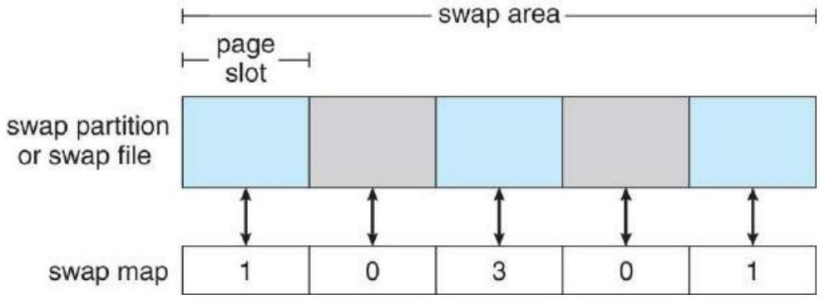
Arrancar desde almacenamiento secundario en Windows

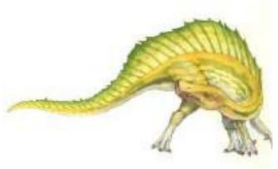


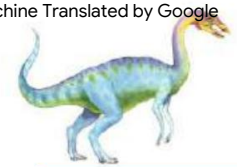


Gestión del espacio de intercambio

- Se utiliza para mover procesos completos (intercambio) o páginas (paginación), de DRAM al almacenamiento secundario cuando la DRAM no es lo suficientemente grande para todos los procesos
- El sistema operativo proporciona gestión del espacio de intercambio
  - El almacenamiento secundario es más lento que la DRAM, muy importante para optimizar el rendimiento.
  - Generalmente son posibles múltiples espacios de intercambio, lo que reduce la carga de E/S en cualquier dispositivo determinado.
  - Lo mejor es tener dispositivos dedicados
  - Puede estar en una partición sin formato o en un archivo dentro de un sistema de archivos (para mayor comodidad al agregarlo)
  - Estructuras de datos para intercambio en sistemas Linux:





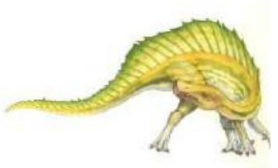


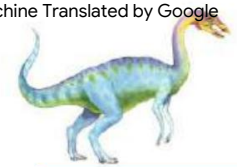
Accesorio de almacenamiento

- Las computadoras acceden al almacenamiento de tres maneras
  - adjunto al host
  - conectado a la red
  - nube
- Acceso al host conectado a través de puertos de E/S locales, utilizando una de varias tecnologías
  - Para conectar muchos dispositivos, utilice buses de almacenamiento como USB, firewire, rayo
  - Los sistemas de alta gama utilizan canal de fibra (FC)

    Arquitectura serial de alta velocidad usando cables de fibra o cobre

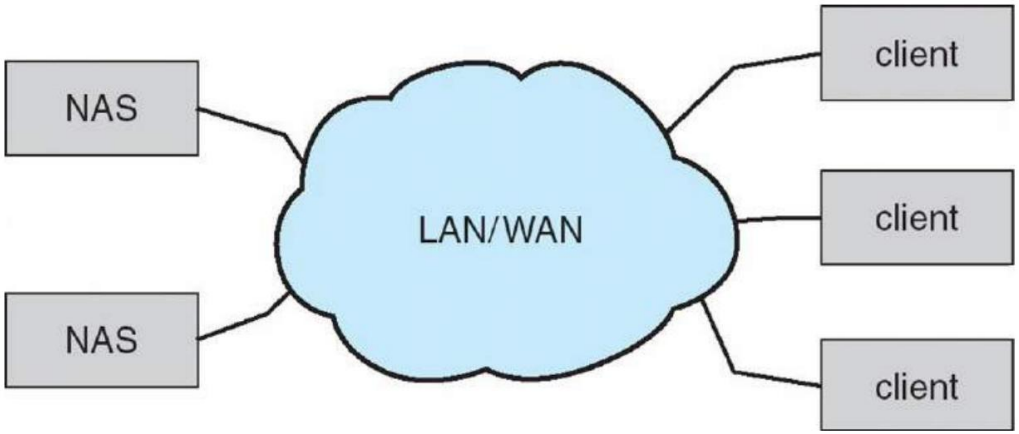
    Se pueden conectar varios hosts y dispositivos de almacenamiento al tejido FC

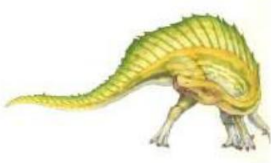


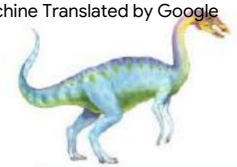


Almacenamiento conectado a la red

- El almacenamiento conectado a la red (NAS) es almacenamiento disponible a través de una red en lugar de a través de una conexión local (como un bus).
  - Conexión remota a sistemas de archivos
- NFS y CIFS son protocolos comunes
- Implementado a través de llamadas a procedimientos remotos (RPC) entre el host y el almacenamiento a través de TCP o UDP en una red IP.
- El protocolo iSCSI utiliza la red IP para transportar el protocolo SCSI.
  - Conexión remota a dispositivos (bloques)

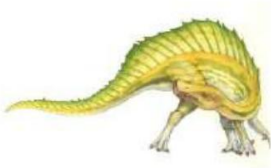


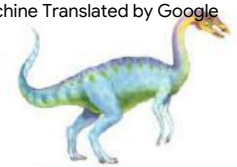




Almacenamiento en la nube

- Similar al NAS, proporciona acceso al almacenamiento a través de una red
  - A diferencia de NAS, se accede a través de Internet o una WAN a un dispositivo remoto.
    centro de datos
- NAS se presenta como un sistema de archivos más, mientras que el almacenamiento en la nube es
  Basado en API, con programas que utilizan las API para proporcionar acceso
  - Los ejemplos incluyen Dropbox, Amazon S3, Microsoft OneDrive,
    iCloud de Apple
  - Utilice API debido a escenarios de latencia y falla (NAS
    los protocolos no funcionarían bien)



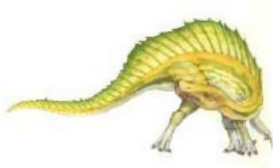


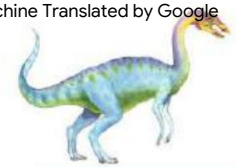
Matriz de almacenamiento

▪ Puede simplemente conectar discos o matrices de discos ▪ Evita el inconveniente del NAS de utilizar el ancho de banda de la red ▪ La matriz de almacenamiento tiene controlador(es) y proporciona funciones a los hosts conectados

- Puertos para conectar hosts al arreglo • Memoria, software de control (a veces NVRAM, etc.)
- Entre unos pocos y miles de discos
- RAID, repuestos en caliente, intercambio en caliente (se analiza más adelante) • Almacenamiento compartido -> más eficiencia
- Funciones que se encuentran en algunos sistemas de archivos

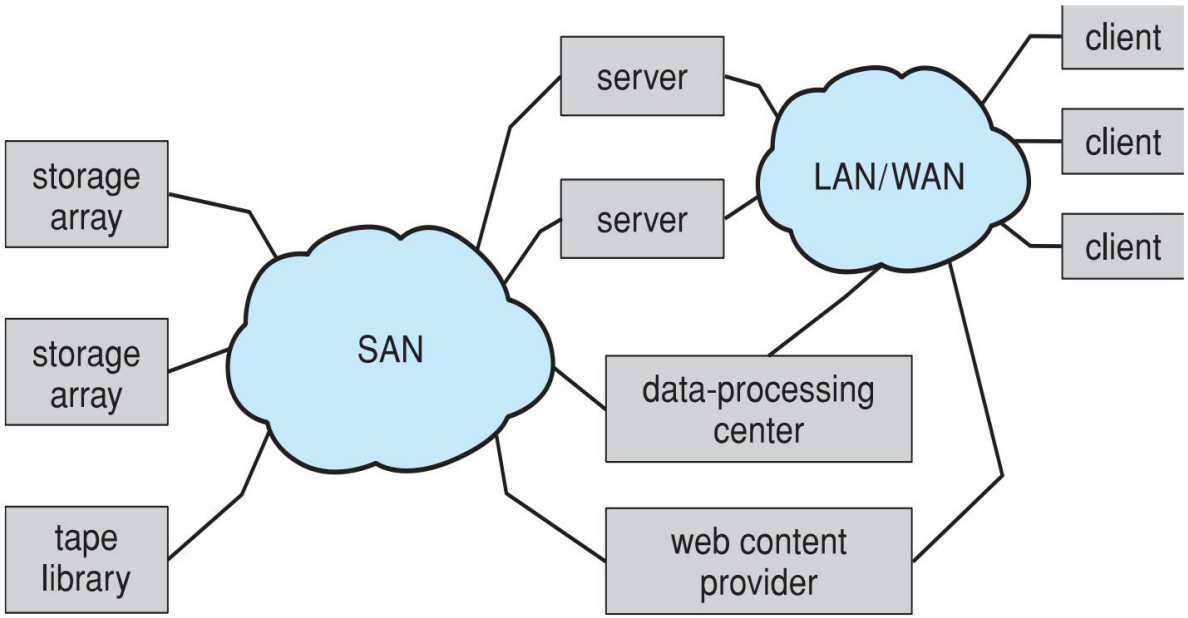
  Instantáneas, clones, aprovisionamiento ligero, replicación, deduplicación, etc.

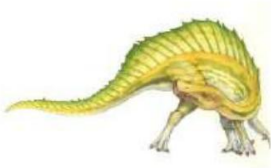


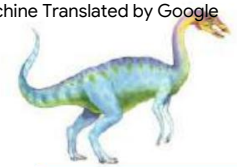


Red de área de almacenamiento

- Común en grandes entornos de almacenamiento
- Múltiples hosts conectados a múltiples matrices de almacenamiento: flexible



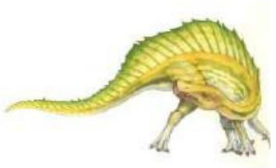


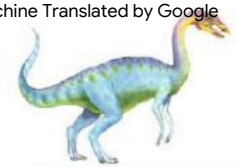


Red de área de almacenamiento (cont.)

- SAN es uno o más arreglos de almacenamiento
  - Conectado a uno o más Conmutadores de canal de fibra o Red InfiniBand (IB)
- Los hosts también se conectan a los conmutadores.
- Almacenamiento disponible a través de LUN Enmascaramiento de matrices específicas a servidores específicos
- Fácil de agregar o quitar almacenamiento, agregar un nuevo host y asignarle almacenamiento
- ¿ Por qué tener almacenamiento separado? redes y redes de comunicaciones?
  - Considere iSCSI y FCOE

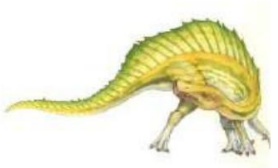
Una matriz de almacenamiento

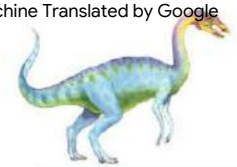




Estructura RAID

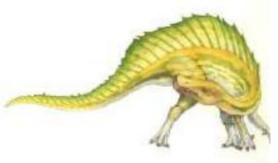
- RAID : conjunto redundante de discos económicos
  - múltiples unidades de disco proporcionan confiabilidad mediante redundancia
- Aumenta el tiempo medio hasta el fallo
- Tiempo medio de reparación : tiempo de exposición cuando podría ocurrir otra falla.
  causar pérdida de datos
- Tiempo medio hasta la pérdida de datos según los factores anteriores
- Si los discos reflejados fallan de forma independiente, considere el disco con un tiempo promedio de falla de 1300 000 y un tiempo promedio de reparación de 10 horas.
  - El tiempo medio hasta la pérdida de datos es 100, 0002 / (2 10) = 500 106 horas, ¡O 57.000 años!
- Combinado frecuentemente con NVRAM para mejorar el rendimiento de escritura
- Varias mejoras en las técnicas de uso de discos implican el uso de múltiples discos trabajando cooperativamente

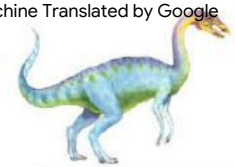




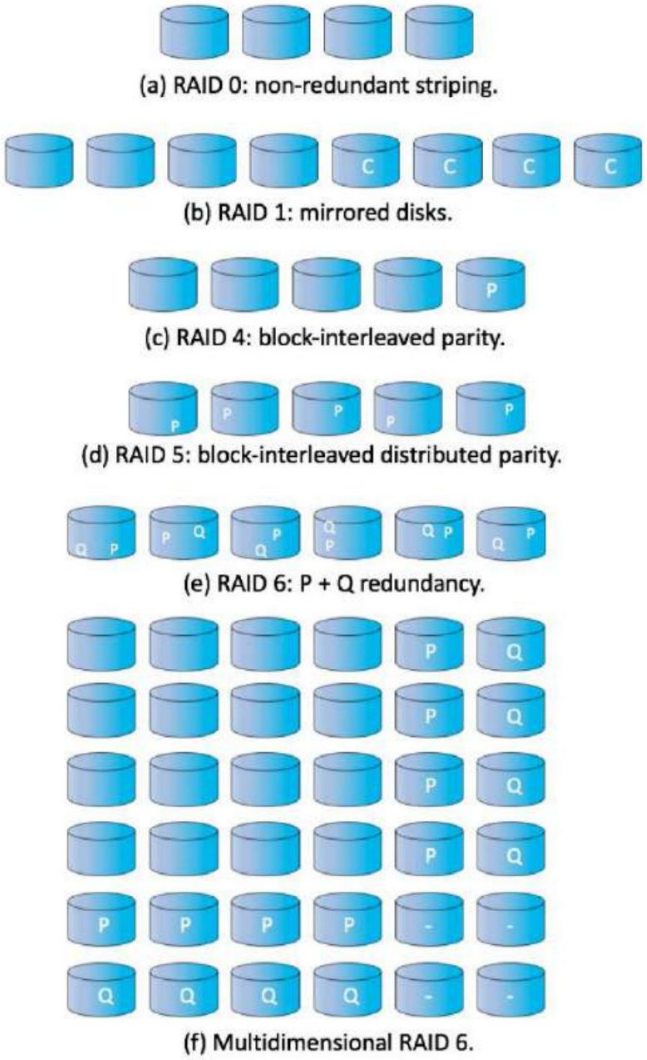
RAID (cont.)

- La división de discos utiliza un grupo de discos como una unidad de almacenamiento.
- RAID está organizado en seis niveles diferentes
- Los esquemas RAID mejoran el rendimiento y mejoran la confiabilidad de el sistema de almacenamiento almacenando datos redundantes
  - Duplicación o sombreado (RAID 1) mantiene duplicados de cada disco
  - Los espejos rayados (RAID 1+0) o las rayas espejadas (RAID 0+1) proporcionan alto rendimiento y alta confiabilidad.
  - La paridad entrelazada de bloques (RAID 4, 5, 6) utiliza mucha menos redundancia
- RAID dentro de una matriz de almacenamiento aún puede fallar si la matriz falla, por lo que La replicación automática de los datos entre matrices es común.
- Con frecuencia, queda una pequeña cantidad de discos de repuesto dinámicos no asignado, reemplazando automáticamente un disco defectuoso y reconstruyendo los datos en él





Niveles de RAID



(a) RAID 0: non-redundant striping.

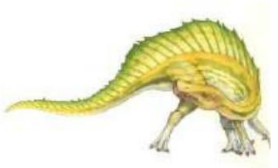
(b) RAID 1: mirrored disks.

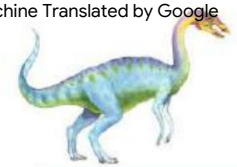
(c) RAID 4: block-interleaved parity.

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

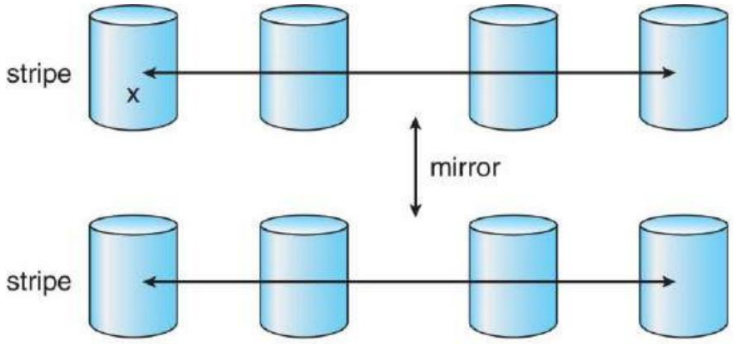
(e) RAID 6: P + Q redundancy.

(f) Multidimensional RAID 6.

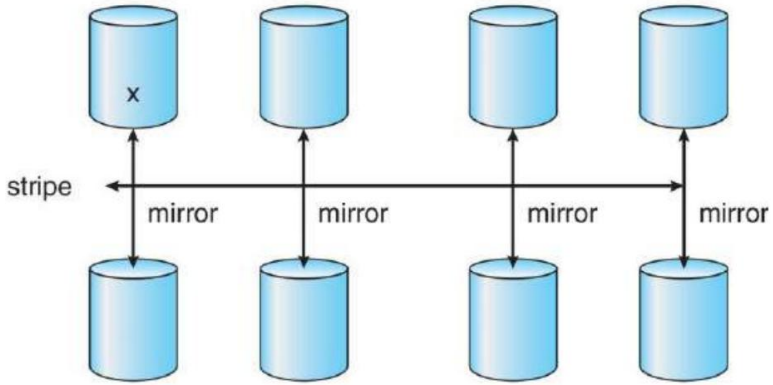




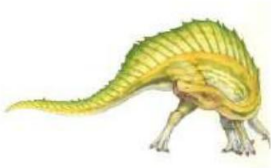
RAID (0+1) y (1+0)

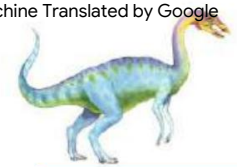


a) RAID 0 + 1 with a single disk failure.



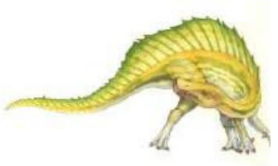
b) RAID 1 + 0 with a single disk failure.

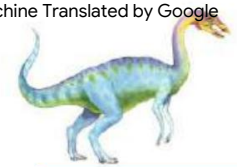




Otras características

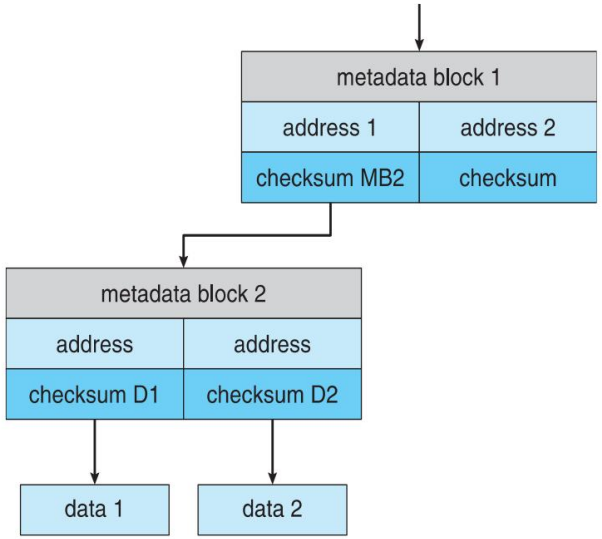
- Independientemente de dónde se implemente RAID, se pueden implementar otras funciones útiles.
  agregado
- La instantánea es una vista del sistema de archivos antes de que se realice un conjunto de cambios.
  (es decir, en un momento determinado)
  - Más en el capítulo 12
- La replicación es la duplicación automática de escrituras entre sitios separados.
  - Para redundancia y recuperación ante desastres
  - Puede ser sincrónico o asincrónico
- El disco de repuesto dinámico no se utiliza; la producción RAID lo utiliza automáticamente
  si un disco falla para reemplazar el disco fallido y reconstruir el conjunto RAID si es posible.
  - Disminuye el tiempo medio de reparación



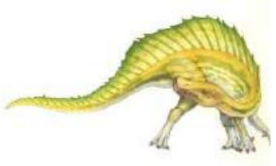


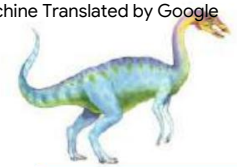
Extensiones

- RAID por sí solo no previene ni detecta la corrupción de datos u otros errores, solo fallas en el disco.
- Solaris ZFS agrega sumas de verificación de todos los datos y metadatos
- Sumas de verificación mantenidas con puntero al objeto, para detectar si el objeto es el correcto y si cambió
- Puede detectar y corregir datos y metadatos. corrupción
- ZFS también elimina volúmenes, particiones
  - Discos asignados en grupos
  - Los sistemas de archivos con un grupo comparten ese grupo, usan y liberan espacio como llamadas de asignación/ liberación de memoria malloc() y free()

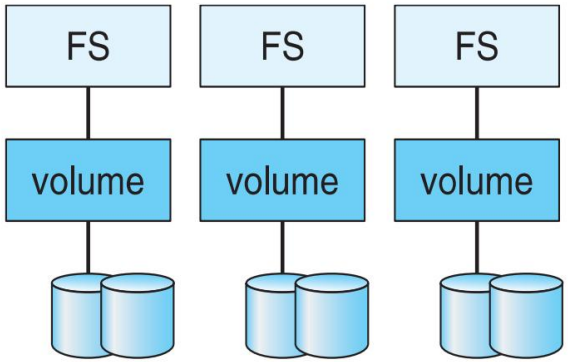


ZFS suma todos los metadatos y datos

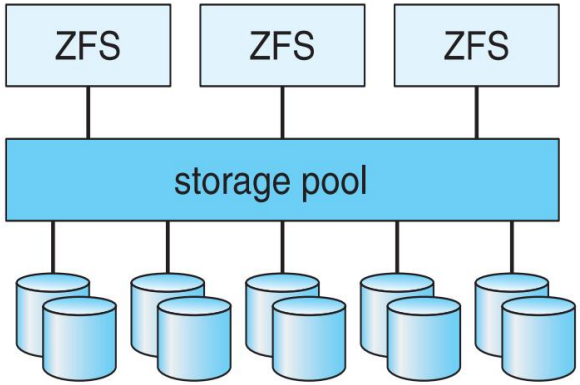




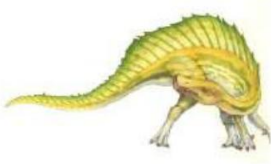
Almacenamiento tradicional y compartido

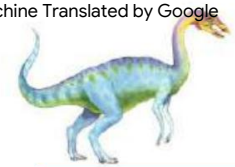


(a) Traditional volumes and file systems.



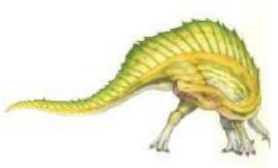
(b) ZFS and pooled storage.

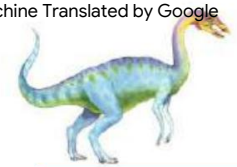




Almacenamiento de objetos

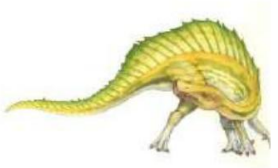
- Computación de propósito general, sistemas de archivos insuficientes para muy Gran escala
- Otro enfoque: comenzar con un grupo de almacenamiento y colocar objetos en eso
  - Objeto sólo un contenedor de datos
  - No hay forma de navegar por el grupo para buscar objetos (sin estructuras de directorios, pocos servicios).
  - Orientado a la computadora, no al usuario
- Secuencia típica
  - Crear un objeto dentro del grupo, recibir una ID de objeto
  - Acceder al objeto a través de esa ID
  - Eliminar objeto a través de esa ID





Almacenamiento de objetos (cont.)

- Software de gestión de almacenamiento de objetos como el sistema de archivos Hadoop (HDFS) y Ceph determinan dónde almacenar objetos y gestiona la protección
  - Normalmente, almacenando N copias, en N sistemas, en el clúster de almacenamiento de objetos.
  - Escalable horizontalmente
  - Contenido direccionable, no estructurado



Fin del Capítulo 11