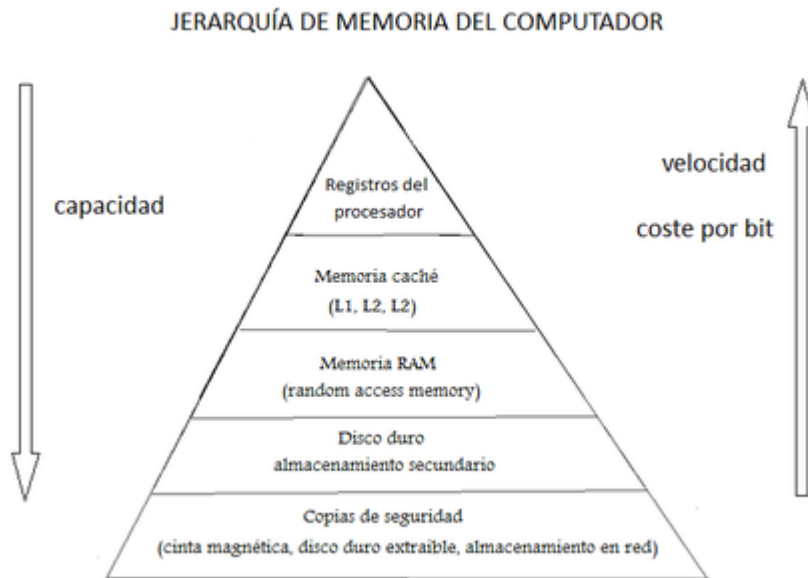


Unidad 5: Memoria

Tipos de memorias, clasificación. Parámetros característicos, tamaño, tiempo de acceso, costo, otros. Memoria principal, formas de organización. Memoria secundaria, organización y formato de datos. Organización jerárquica de la memoria. Dispositivos de almacenamiento externo, disco, cinta, disco óptico, otros. Múltiples unidades de discos (RAID).

Jerarquía de memorias:



Mientras mas se escale en la pirámide, la memoria es mas pequeña, mas costosa por bit y mas rápida.

Para solventar la diferencia de velocidades entre una memoria y otra, se aplican distintos niveles de cache que funcionan de intermediarios, almacenando copias de porciones del programa.

Esto funciona gracias al principio de localidad de memoria.

Memoria, generalidades

Algunos conceptos relacionados con memorias internas:

- Palabra: es la unidad de organización de la memoria. El tamaño de la palabra suele coincidir con el numero de bits utilizados para representar números y con la longitud de las instrucciones.
- Unidades direccionables: en muchos sistemas la unidad direccionable es la palabra.
- Unidades de transferencia: para la memoria principal, es el numero de bits que se leen o escriben en memoria a la vez. Para la memoria externa, los datos se transfieren normalmente en unidades mas grandes que la palabra denominadas "bloques".

Otro concepto distintivo de las memorias es el método de acceso:

- Acceso secuencial: la memoria se organiza en unidades de datos llamadas registros. El acceso debe realizarse con una secuencia lineal específica, es decir que se lee una línea, y luego la línea que esta inmediatamente contigua.
- Acceso directo: se lleva a cabo mediante un acceso directo a una vecindad dada, seguido de una búsqueda secuencial.
- Acceso aleatorio (random): el tiempo de acceso es constante y cualquier celda de memoria puede referenciarse y ser direccionada aleatoriamente. La memoria principal y algunos sistemas de cache son de acceso aleatorio.
- Asociativa: es una memoria de acceso aleatoria que permite hacer una comparación de ciertas posiciones de bits dentro de una palabra buscando que coincidan con unos valores dados y hacer esto para todas las palabras simultáneamente. Por lo tanto, una palabra es recuperada basándose en una porción de su contenido en lugar de su dirección.

Los dos puntos fundamentales de una memoria son su capacidad y sus prestaciones. Estas ultimas se miden por los siguientes parámetros

Tiempo de acceso: en memorias de acceso random es el tiempo que tarde en realizarse una operación de escritura o de lectura. En otros tipos de memorias es el tiempo que se tarda en situar el mecanismo de lectura/escritura en la posición deseada.

Tiempo de ciclo de memoria: concepto asociado a las memorias random. Consiste en el tiempo de acceso mas algún tiempo extra que se requiere antes de realizar otra operación. Este tiempo puede ser necesario para que finalicen las transiciones en las líneas de señal o para regenerar los datos en el caso de lecturas destructivas.

Velocidad de transferencia: es la velocidad a la que se pueden transferir datos hacia o desde una unidad de memoria. Para las memorias random coincide con el inverso del tiempo de ciclo.

Tipos de memorias:

Memorias semiconductoras de acceso aleatorio (memorias RAM)

La característica principal es que es posible tanto leer datos como escribir rápidamente datos nuevos. Tanto la lectura como la escritura se ejecutan mediante señales eléctricas.

La otra características distintiva es que es volátil. Esto implica que debe estar continuamente alimentada. Si se interrumpe la alimentación se pierden los datos, de modo que solo puede usarse como almacenamiento temporal.

Las tecnologías RAM se dividen en estaticas y dinamicas.

Las dinamicas están hechas con celdas que almacenan los datos como cargas en condensadores. La presencia o ausencia de carga se interpreta como el 1 o el 0 binarios. Las RAM dinamicas requieren refrescos periódicos para mantener memorizados los datos. Los valores binarios se almacenan utilizando configuraciones de puertas que forman biestables. Estas memorias son mas densas(celdas mas pequeñas=mas celdas por unidad de superficie) y

mas económicas que las estaticas, aunque requieren circuitería para el refresco. Suelen usarse para memorias de mayor tamaño.

Una RAM estatica mantendrá sus datos mientras tenga alimentación y son mas rapidas que las dinamicas.

Ambos tipos de RAM son volátiles.

En contraste con las memorias RAM, existen las memorias ROM. Una memoria de este tipo contiene un patron permanente de datos que no puede alterarse, por lo tanto no pueden escribirse datos nuevos. Se utilizan en la microprogramación, subrutinas de biblioteca para funciones de uso frecuente, programas del sistema y tablas de funciones.

Organización de las memorias:

El elemento básico de una memoria semiconductora es la celda de memoria. Todas las celdas de memoria comparten ciertas propiedades:

- Presentan dos estados estables o semi estables que pueden emplearse para representar el 1 o el 0 binario.
- Puede escribirse en ellas al menos una vez para fijar su contenido.
- Pueden leerse para detectar su estado.

Básicamente una celda de memoria debe tener tres terminales para transportar señales eléctricas. La primera, la terminal de selección, selecciona la celda para la operación a realizar. La terminal de control, indica el tipo de operación (si se trata de una lectura o escritura). La tercera terminal, se utiliza como salida del estado de la celda en la lectura o proporciona la señal a fijar en la celda en el caso de una escritura.

Lógica del chip

La memoria semiconductora viene encapsulada en chips. Cada chip contiene una matriz de celdas de memoria.

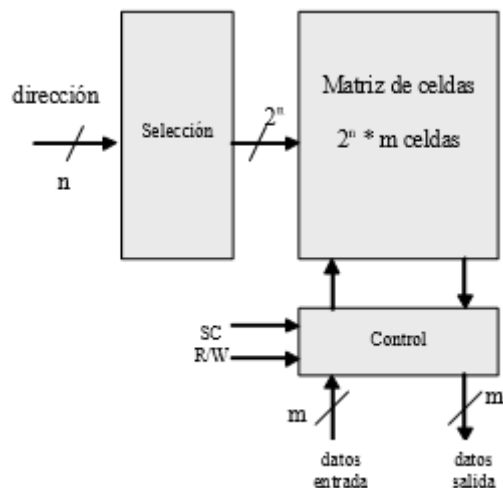
Si la memoria contiene 1 bit por palabra, se necesitaran claramente al menos un numero de chips igual al numero de bits por palabra.

Organización 2d

Es la forma mas sencilla de organización de memoria.

Las celdas forman una matriz de 2^n filas y M columnas, siendo 2^n el numero de palabras del chip y m la cantidad de bits de cada palabra. Cada fila es seleccionada por la decodificación de una configuración diferente de los N bits de dirección. Esta organización tiene el inconveniente de que el selector (decodificador) crece exponencialmente con el tamaño de la memoria.

Igual ocurre al numero de entradas de las compuertas que generan las salidas.



De esta forma, 2^n es el tamaño de bits de direcciones que el decodificador puede re direccionar. M es la cantidad de bits que tiene cada palabra, por tanto el tamaño total de la memoria en bits sería $2^n * m$.

Memoria 2d/2

En lugar de una sola selección de 2^n salidas, en esta organización se utilizan dos decodificadores de $2^{n/2}$, operando en coincidencia. Las líneas de direcciones se reparten entre los decodificadores. Para una configuración dada de las líneas de dirección se selecciona un único bit de la matriz. Por ello también se la denomina organización por bits.

En esta organización se necesitan múltiples matrices, tantas como bits deba tener la palabra. De este modo, para armar una palabra, se selecciona un bit de cada matriz en la misma exacta posición de cada matriz, actuando de forma paralela.

Memoria cache

El objetivo de la memoria cache es lograr que la velocidad de la memoria sea lo mas rápida posible. La cache contiene una copia de partes de la memoria principal. Cuando el procesador intenta leer una palabra de memoria se hace una comprobación para determinar si la palabra esta en la cache. Es es asi, se entrega dicha palabra al procesador. Sino, un bloque de memoria principal consistente en un cierto numero de palabras se trasfiere a la cache y después la palabra es entregada al procesador.

Esto es posible debido al principio de localidad de memoria. Este dice que al ejecutar una instrucción, la siguiente instrucción a ejecutar del proceso es posiblemente, la instrucción contigua en memoria.

Discos magneticos

Organización y formato de los datos

Un disco magnetico es un plato circular construido con metal o plástico cubierto de un material magnetizable. Los datos se graban en el y se recuperan del disco a travez de una bobina llamada cabeza. Durante una operación de lectura o escritura la cabeza permanece

quieta mientras el plato rota bajo ella. El mecanismo de escritura se basa en el campo magnetico producido por el flujo eléctrico que atraiesa la bobina.

La cabeza es un dispositivo realmente pequeño, capaz de leer o escribir una zona del plato que rota bajo ella. Esto da lugar a que los datos se organicen en un conjunto de anillos concéntricos en el plado llamados pistas. Cada pista tiene el mismo ancho que la cabeza.

Los datos se transfieren desde y hacia el disco en bloques. Normalmente el bloque es menor que la capacidad de una pista. De acuerdo con esto los datos se almacenan en regiones del tamaño de un bloque conocidas como sectores. Estos pueden ser se longitud fija o variable.

En la mayoría de los discos la cubierta magnetizable se aplica a ambas caras del plato denominándose discos de doble superficie. Algunas unidades de disco poseen arios platos apilados verticalmente y separados por cierta distancia. Estos discos disponen de multiples brazos y se le conoce como cilindro.

En estos discos las prestaciones se evalúan por los siquientes criterios:

Tiempo de búsqueda: el tiempo que tarda en desplazar el brazo del disco hasta la pista requerida.

Retardo rotacional: es el tiempo que tarda en girar el disco.

Tiempo de transferencia: el tiempo que tarda en escribir o leer un dato en el disco.

Discos RAID

1. Es un conjunto e unidades físicas de disco vistas por el sistema operativo como una única unidad lógica.
2. Los datos se distribuyen a través de las unidades físicas del conjunto de unidades.
3. La capacidad de los discos redundantes se usa para almacenar información de paridad que garantice la recuperación de los datos en caso de fallo de disco.

Los detalles de la segunda y tercera característica cambian según los distintos nieles de RAID. El nivel 0 no soporta la tercera característica.

Esta organización de discos reemplaza luna unidad de disco de gran capacidad por unidades múltiples de menor capacidad y distribuye los datos de forma que se puedan habilitar accesos simultáneos a los datos de varias unidades, mejorando por tanto, las prestaciones de E/S y permitiendo más fácilmente aumentos e la capacidad. Esta estrategia hace hincapié en la necesidad de redundancia. El uso de varios dispositivos, además de permitir que arias cabezas operen simultáneamente consiguiendo mayores velocidades de E/S y de transferencia, incrementa la probabilidad de fallos. Para comenzar esta disminución de seguridad, RAID utiliza la información de paridad almacenada que permite la recuperación de datos perdidos debido a un fallo de disco

| Categoría | Nivel | Descripción | Grado de E/S solicitado Lectura/escritura | Grado de transferencia de datos Lectura/escritura | Aplicación típica |
|-----------|-------|-------------|----------------------------------------------|------------------------------------------------------|-------------------|
| | | | | | |

| | | | | | |
|----------------------|---|----------------------------------------------------|-------------------------|---------------------------|---------------------------------------------------------------------|
| Estructura en tiras | 0 | No redundante | Tiras largas: excelente | Pequeñas tiras: excelente | Aplicaciones que requieren altas prestaciones con datos no críticos |
| Estructura en espejo | 1 | Espejo | Bueno/regular | Regular/regular | Controladores de sistemas; ficheros críticos |
| Acceso paralelo | 2 | Redundante con código hamming | Pobre | Excelente | |
| | 3 | Bit de paridad intercalado | Pobre | Excelente | Aplicaciones con muchas E/S tales como imágenes cad |
| Acceso independiente | 4 | Boque de paridad intercalado | Excelente/regular | Excelente/pobre | |
| | 5 | Paridad distribuida en bloques intercalados | Excelente/regular | Excelente/pobre | Grado de petición alto, lectura intensiva, consulta de datos |
| | 6 | Paridad distribuida dual en bloques intercambiados | Excelente/regular | Excelente/pobre | Aplicaciones que requieren alta disponibilidad |