**CPU vs MEMORIA**

Históricamente, la CPU ha sido más rápida que la memoria. El aumento en la cantidad y complejidad de los circuitos integrados que componen los chips de CPU y memoria se ha orientado principalmente a: mejorar la performance de la CPU para hacerla más veloz (por ejemplo, mediante el uso de pipeline) y aumentar la capacidad de almacenamiento de la memoria (más memoria y más grandes decodificadores).

Esta diferencia de velocidades implica que después de que la CPU emite una solicitud de lectura a la memoria (a través de los buses de direcciones y control), pasan varios ciclos de reloj antes de recibir la información requerida en el bus de datos. En todos los ciclos de instrucción, la CPU accede a la memoria al menos una vez para buscar la instrucción y varias veces para buscar operandos. Por lo tanto, la velocidad a la que la CPU ejecuta instrucciones está limitada parcialmente por el tiempo del ciclo de memoria.

El problema del desbalance entre la velocidad de la CPU y la memoria no es solo tecnológico, sino también económico. Aunque se pueden construir memorias muy rápidas, casi tan rápidas como la CPU, para lograr la máxima velocidad se requiere que estén dentro del chip de la CPU, evitando así el uso del bus. El acceso a la memoria por el bus del sistema es inherentemente lento en comparación con la velocidad interna de la CPU.

**MEMORIA CACHÉ**

Una forma de compensar el desbalance entre CPU y memoria es usar una jerarquía de memoria. Esta jerarquía combina memorias muy rápidas y pequeñas con memorias grandes y más lentas para obtener una velocidad conjunta (de ambas memorias) casi tan rápida en promedio. La memoria muy rápida y pequeña que integra esta jerarquía es la memoria caché.

La memoria caché es una memoria muy rápida y pequeña intercalada entre la CPU y la memoria principal. Es casi tan rápida como la CPU y de tamaño miles de veces más pequeña que la memoria principal. Cuando la CPU busca una información, si está en la caché la trae desde ahí, y si no, la busca en la memoria principal.

Físicamente, la memoria caché está intercalada entre la CPU y la memoria principal, y puede estar integrada en el mismo chip de la CPU o fuera de él. Debido a su velocidad, la memoria caché es relativamente mucho más pequeña que la memoria principal, conteniendo solo una fracción de la información de esta última.

El uso de la memoria caché se basa en el comportamiento determinístico de los programas, es decir, en su comportamiento predecible. Este comportamiento determinístico se fundamenta en dos principios empíricos:

* **Principio de localidad espacial**: Cuando se accede a una palabra de memoria, es altamente probable que el próximo acceso sea en la vecindad de la palabra anterior.
* **Principio de localidad temporal**: Cuando se accede a una posición de memoria, es altamente probable que en un corto lapso de tiempo dicha posición sea accedida nuevamente.

El principio de localidad espacial se sustenta en la forma en que la CPU accede a las instrucciones y los datos. Estadísticamente, un elevado porcentaje de instrucciones se ejecuta accediendo a instrucciones consecutivas en memoria (acceso secuencial). Además, los programadores suelen asignar variables relacionadas próximas entre sí y las estructuras de datos como matrices o pilas se almacenan en áreas de memoria cercanas, facilitando el acceso.

El principio de localidad temporal se manifiesta en cómo la CPU accede a código, ciclos o bucles y subrutinas, así como a datos y estructuras de datos. Un ejemplo típico es la ejecución de un bucle, donde las instrucciones se acceden secuencialmente y repetitivamente durante todo el tiempo que dura el bucle. Otro ejemplo es el acceso repetitivo a elementos de un arreglo en un bucle, mostrando tanto la localidad temporal como espacial.

Basándose en estos principios, la idea de intercalar una memoria caché muy rápida entre la CPU y la memoria principal puede ser efectiva. La memoria caché almacena la posición de memoria a acceder y sus vecinas, lo que permite un acceso mucho más rápido que si se hiciera a la memoria principal, dependiendo de la tasa de aciertos.

Para facilitar las transferencias entre la memoria principal y la caché, ambas memorias se organizan en bloques de palabras de igual tamaño. Las transferencias entre la memoria principal y la caché se realizan por bloque de datos, siendo la mínima transferencia un bloque. Entre la CPU y la caché, la transferencia es por palabra, ya que así opera la CPU.

El proceso de asignación de bloques de la memoria principal en la caché se llama función de mapeo. Existen tres técnicas básicas de mapeo: asociativa, directa y asociativa por grupos.

Cuando la CPU busca un dato en la caché, puede encontrarlo o no. Si el dato se encuentra, ocurre un acierto (hit), y la CPU obtiene el dato a alta velocidad. Si el dato no se encuentra, ocurre un fallo (miss), y la CPU tiene que obtenerlo de la memoria principal, a una velocidad menor. La eficiencia de la caché se mide a través de la frecuencia de aciertos, es decir, el número de veces que la CPU encuentra el dato en la caché.

En algunos casos, la caché está organizada en varios niveles (L1, L2, L3, etc.). Cuanto más alto es el nivel, más pequeña es la memoria. Dado que típicamente el porcentaje de aciertos de una memoria caché es del 90% (la tasa de fallos es del 10%), organizar la memoria caché en múltiples niveles busca mejorar la eficiencia del 90% de aciertos.

**MEMORIA SECUNDARIA**

La memoria externa comprende dispositivos de almacenamiento que no están conectados físicamente al bus del sistema. Entre estos dispositivos se encuentran los discos magnéticos, discos ópticos, CD-ROM, CD-R, CD-RW y DVD.

**DISCOS MAGNÉTICOS**

Los discos magnéticos son dispositivos electromecánicos compuestos por uno o más platos rígidos, que suelen ser de aluminio o vidrio (este último tiene un menor coeficiente de dilatación que el aluminio). Estos platos están recubiertos con una capa de material magnético, como óxido de hierro, y su superficie debe ser extremadamente uniforme y libre de defectos.

En el disco se generan pequeñas áreas que pueden ser magnetizadas en dos direcciones por un transductor, conocido como cabeza lectora/grabadora. El disco gira, creando un movimiento relativo entre la superficie magnetizable y el transductor durante la lectura o escritura.

La lectura y escritura se realizan mediante una cabeza transductora que permanece estática mientras el plato gira. Los puntos magnetizados en dos direcciones representan los estados binarios de 1 y 0, almacenándose en círculos concéntricos en el disco.

El movimiento relativo del disco respecto a una cabeza lectora/grabadora fija determina un anillo o pista. La cabeza se mueve radialmente en pasos discretos, y cada posición de la cabeza corresponde a una pista o "track". La cantidad de pasos discretos define el número de pistas disponibles en el disco.

Las pistas son concéntricas y están separadas por espacios vacíos denominados "gaps". Reducir los gaps aumenta la capacidad del disco, ya que aumenta la densidad de pistas. Se asume que el número de bits por pista es igual para todas las pistas.

Las pistas se dividen en sectores, siendo el sector la mínima unidad de transferencia del disco. Cada pista contiene un número entero de sectores.

**TIPOS DE DISCOS MAGNÉTICOS**

Los discos magnéticos pueden variar según el tipo de cabeza (fija o móvil), si son removibles o fijos, si tienen uno o ambos lados utilizables, y si tienen uno o múltiples platos. Los tipos de cabezales incluyen de contacto (como en disquetes), de distancia de separación fija, y de separación aerodinámica (conocido como tipo Winchester).

**CARACTERÍSTICAS TÍPICAS**

Los discos suelen tener múltiples platos, con una cabeza lectora/grabadora por cara o superficie. Todas las cabezas se mueven solidariamente, y las pistas alineadas en cada plato forman una estructura lógica llamada "cilindro". Los datos se almacenan por cilindros para reducir el movimiento de las cabezas y aumentar la velocidad de respuesta.

**FORMATO DE GRABACIÓN EN DISCOS**

El formato estándar ST 506 para discos magnéticos tiene sectores con cinco campos: Gap1 (separación entre sectores y el inicio del sector), Id (identificación del sector), Gap2 (separación entre el campo Id y el campo de datos), Data (campo de datos), y Gap3 (separación entre sectores y el fin del sector). Los gaps son espacios sin información útil.

El campo Id consta de SyncByte (patrón de bits para sincronizar la operación), Track (número de pista), Head (número de cabeza lectora/grabadora), Sector (número de sector dentro de la pista), y CRC (bits para detección y corrección de errores).

El campo Data incluye encabezado (para sincronizar la lectura e identificar el sector), datos (de una longitud en bytes normalmente expresada como potencia de 2), y código para errores (CRC).

La capacidad total de un disco se calcula multiplicando el número de sectores, pistas y superficies, y dividiendo por el tamaño del sector. Debido a las restricciones del formato, se desperdicia espacio en las pistas más externas. En formatos de grabación posteriores, se particionan las áreas en zonas para aumentar la capacidad, aunque esto requiere circuitos electrónicos más complejos debido a la variación en la densidad de grabación.

El tiempo de acceso a un sector en un disco consta de dos componentes: el tiempo de búsqueda (seek), que es el tiempo que tarda la cabeza en llegar al cilindro o pista buscado, y el tiempo de latencia (por rotación), que es el tiempo de espera hasta que el sector pase por debajo de la cabeza. El tiempo total de acceso incluye el tiempo de acceso y el tiempo de transferencia de datos.

**FORMATOS DE GRABACIÓN**

La velocidad de rotación angular del disco puede ser fija o variable. En el formato CAV (velocidad constante), la velocidad lineal de los bits varía con el radio del disco, causando que los bits en las pistas más cercanas al centro giren más lentamente que los de las pistas más alejadas. En el formato VAV (velocidad angular variable), se puede mantener una velocidad lineal constante para todos los bits, de modo que los bits y sectores ocupan el mismo tamaño en todas las pistas.

El formato de grabación de un disco define la cantidad de sectores, el tamaño del sector y las funcionalidades de los campos en cada pista y sector. Los formatos pueden ser por hardware (usando marcas físicas para definir el tamaño de los sectores, un método antiguo) o por software (donde el tamaño del sector está determinado por el sistema operativo).