Unidad aritmético lógica

La ALU es una caja negra donde resuelve operaciones (de entrada tiene datos y control) y devuelve resultados, y estados (flags)

Características fundamentales

- tipos de datos
- repertorio de operaciones
- formatos de instrucción
- registros (espacios de almacenamiento mínimo)
- Modos de direccionamiento (RISC vs CISC) depende del tipo de instrucciones

Tipos de datos: direcciones, números, cadena de caracteres, lógicos
Repertorio de instrucciones (Arquitectura) por ejemplo ADD, JMP, CLR de simuproc.

Esta es la frontera donde el diseñador de la computadora y el programador pueden ver la misma máquina. ¿A que se refiere?. Se refiere a que el programador (de bajo nivel) diseña el programa acorde al repertorio de instrucciones y el diseñador de hardware se encarga de que el mismo sea funcional para realizar programas.

Cada instrucción debe tener 4 elementos básicos

- código de operación: especifica que tipo de operación debe ser realizada (no puede faltar nunca).
- referencia a los operandos fuentes: Donde están los datos requeridos para ejecutar la operación.
- Referencia al resultado: donde lo guardo (pueden ser varios)
- Referencia a la próxima instrucción.

La pila la utilizamos cuando no tenemos operandos explícitos, que es donde se van guardando estos operandos de manera implícita.

El repertorio de instrucciones debe ser **funcionalmente completo, eficiente, posee clases de instrucciones regulares y completas, y ortogonal** (máximo tipos de datos, tipos de direcciones y repertorio de instrucciones). Para que el usuario pueda formular cualquier tarea de procesamiento de alto nivel.
Tipos: operaciones aritméticas, operaciones lógicas, movimientos de datos, operaciones de entrada y salida

Tipos de instrucciones: sentencias, condicionales, repeticiones, subrutinas (procedimientos, funciones)
Modo de direccionamiento basados en registros: Se obtiene acceso más rápido a los datos y en la instrucción campos de dirección más pequeños
Direccionamiento por registro: es similar al directo, pero el campo de dirección especifica la dirección de un registro
Direccionamiento indirecto por registro: similar al indirecto, pero el campo de dirección especifica el número de registros que contiene la dirección efectiva del dato.

Memoria

Características
Palabra: es la unidad «natural» de organización de la memoria. El tamaño de la palabra suele coincidir con el número de bits utilizados para representar números y con la longitud de las instrucciones. Por desgracia hay muchas excepciones.
Unidades direccionables: en algunos sistemas la unidad direccionable es la palabra.
Sin embargo, muchos de ellos permiten direccionar a nivel de bytes. En cualquier caso, la relación entre la longitud A de una dirección y el número N de unidades direccionables, es 2^A = N.
Unidad de transferencia: para la memoria principal es el número de bits que se leen o escriben en memoria a la vez. La unidad de transferencia no tiene por qué coincidir con una palabra o con una unidad direccionable. Para la memoria externa, los datos se transfieren normalmente en unidades más grandes que la palabra, este procedimiento es el “block transfer“.
Métodos de acceso
Acceso secuencial: la memoria se organiza en unidades de datos llamadas registros. El acceso debe realizarse con una secuencia lineal específica. Se hace uso de información almacenada de direccionamiento que permite separar los registros y ayudar en el proceso de recuperación de datos. Se utiliza un mecanismo de lectura/escritura compartida que debe ir trasladándose desde su posición actual a la deseada, pasando y obviando cada registro intermedio. Así pues, el tiempo necesario para acceder a un registro dado es muy variable.
Acceso directo: como en el caso de acceso secuencial, el directo tiene asociado un mecanismo de lectura/escritura. Sin embargo, los bloques individuales o registros tienen una dirección única basada en su dirección física. El acceso se lleva a cabo mediante un acceso directo a una vecindad dada, seguido de una búsqueda secuencial, bien contando, o bien esperando

hasta alcanzar la posición final. De nuevo el tiempo de acceso es variable.
Acceso aleatorio (random): cada posición direccionable de memoria tiene un único mecanismo de acceso cableado físicamente. El tiempo para acceder a una posición dada es constante e independiente de la secuencia de accesos previos. Por tanto, cualquier posición puede seleccionarse «aleatoriamente» y ser direccionada y accedida directamente. La memoria principal y algunos sistemas de caché son de acceso aleatorio.
Asociativa: es una memoria del tipo de acceso aleatorio 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. Una palabra es por tanto recuperada basándose en una porción de su contenido en lugar de su dirección. Como en las memorias de acceso aleatorio convencionales, cada posición tiene su propio mecanismo de direccionamiento, y el tiempo de recuperación de un dato es una constante independiente de la posición o de los patrones de acceso anteriores. Las memorias caché pueden emplear acceso asociativo.

Desde el punto de vista del usuario, las dos características más importantes de una memoria son **su capacidad y sus prestaciones**. Se utilizan tres parámetros de medida de prestaciones, **tiempo de acceso, tiempo de ciclo de memoria y velocidad de transferencia.**

CAPACIDAD , TIEMPO DE ACCESO Y COSTO
• A menor tiempo de acceso, mayor coste por bit.
• A mayor capacidad, menor coste por bit.
• A mayor capacidad, mayor tiempo de acceso.

Debido a que el diseñador quiere maximizar el rendimiento-costo, se emplea un jerarquía de memoria.
a) Disminuye el coste por bit.
b) Aumenta la capacidad.
c) Aumenta el tiempo de acceso.
d) Disminuye la frecuencia de accesos a la memoria por parte del procesador
Memorias más pequeñas, más costosas y más rápidas, se complementan con otras más grandes, más económicas y más lentas.

Contenido de una línea/entrada/renglón de Caché:
Bit de validez: indica si la entrada contiene datos válidos. Cuando arranca el sistema todas las entradas son NO VÁLIDAS.

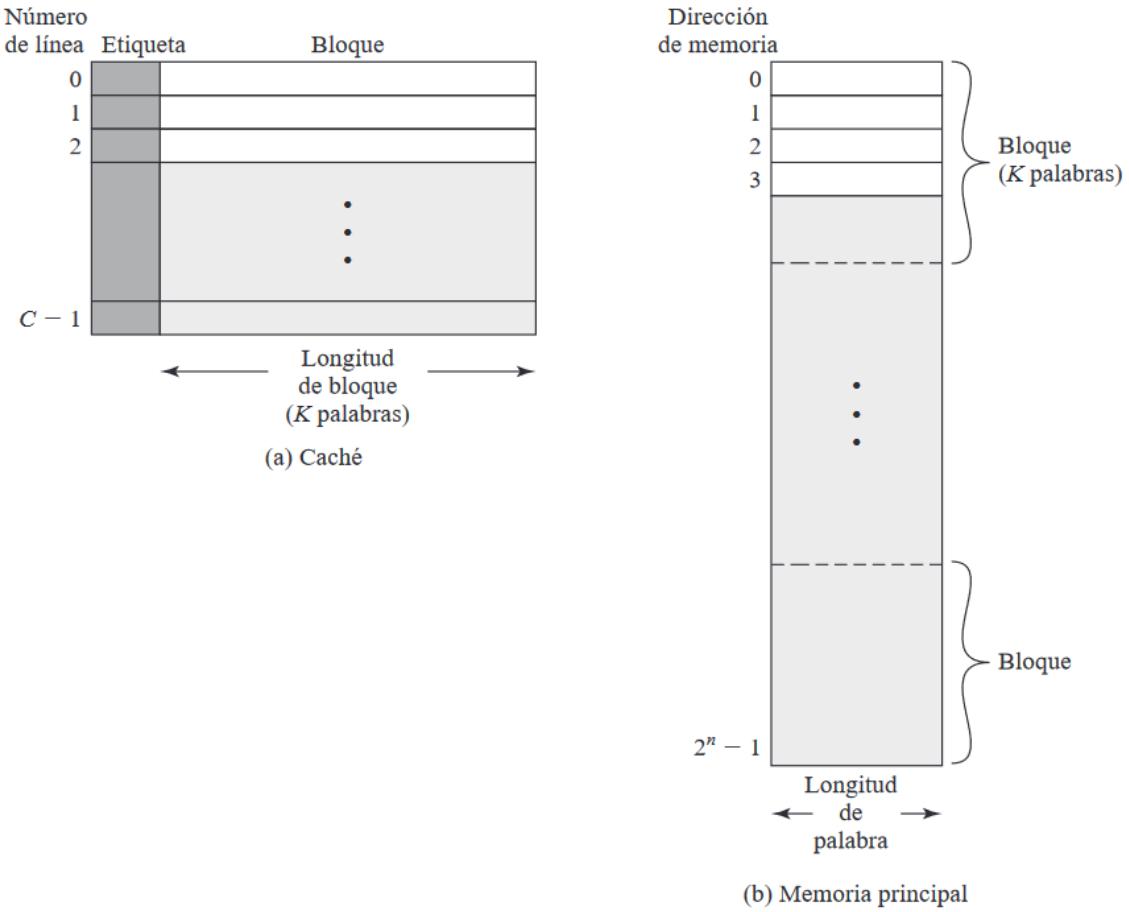
Campo Etiqueta (Tag): identifica al bloque de memoria de donde provienen los datos
Campo de Datos: contiene una copia de los datos de la Memoria Principal.

OBJETIVO PRINCIPAL DE LA MEMORIA CACHÉ: DISMINUIR LATENCIA

Funcionamiento entre memoria caché y memoria principal

La memoria principal consta de hasta 2^n palabras direccionables, teniendo cada palabra una única dirección de n bits.
Esta memoria la consideramos dividida en un número de bloques de longitud fija, de K palabras por bloque. Es decir, hay M 2n/K bloques. La caché consta de C líneas. Cada línea contiene K palabras, más una etiqueta de unos cuantos bits; denominándose tamaño de línea al número de palabras que hay en la línea. El número de líneas es considerablemente menor que el número de bloques de memoria principal (C M). En todo momento, un subconjunto de los bloques de memoria reside en líneas de la caché. Si se lee una palabra de un bloque de memoria, dicho bloque es transferido a una de las líneas de la caché. Ya que hay más bloques que líneas, una línea dada no puede dedicarse única y permanentemente a un bloque. Por consiguiente, cada línea incluye una etiqueta que identifica qué bloque particular almacena. La

etiqueta es usualmente una porción de la dirección de memoria principal.



Elementos de diseño de la caché

Tamaño de caché

Función de correspondencia: forma en la que se decide qué bloque de memoria principal ocupa actualmente una línea determinada de caché. Tipos de algoritmos.

Correspondencia directa. La técnica más sencilla, denominada correspondencia directa, consiste en hacer corresponder cada bloque de memoria principal a solo una línea posible de caché.

ventaja: sencilla y barata **desventaja:** menor tasa de aciertos por si se requieren 2 bloques distintos ubicados en la misma línea de manera recurrente.

Correspondencia asociativa: Cada bloque puede cargarse en cualquier línea de la caché. Interpreta la dirección de memoria como una etiqueta y un campo. **ventaja:** ofrece flexibilidad para cualquier bloque. **desventaja:** circuitos complejos.

Correspondencia asociativa por conjuntos. La correspondencia asociativa por conjuntos es una solución de compromiso que recoge lo positivo de las correspondencias directa y asociativa, sin presentar sus desventajas. En este caso, la caché se divide en v conjuntos, cada uno de k líneas. La lógica de control de la caché interpreta una dirección de memoria como tres campos: etiqueta, conjunto y palabra.

Algoritmos de sustitución

La Correspondencia directa no necesita un algoritmo de sustitución pero la asociativa si.

El más efectivo es el denominado **"utilizado menos recientemente"** se sustituye el bloque que se ha mantenido en la caché por más tiempo sin haber sido referenciado

Otra posibilidad es el primero en entrar primero en salir (**FIFO, First-In-First-Out**): se sustituye aquel bloque del conjunto que ha estado más tiempo en la caché.

Otra posibilidad más es la del **utilizado menos frecuentemente** (LFU, Least-Frequently Used): se sustituye aquel bloque del conjunto que ha experimentado menos referencias.

Y por último está la aleatoria que proporciona unas prestaciones sólo ligeramente inferiores.

Principio de localidad el cual establece que es probable que los datos en la vecindad de una palabra referenciada sean referenciados en un futuro próximo.

Más elementos de diseño

Tamaño de línea

Número de caches

Política de escritura (actualización)

Write Through (Escritura Directa): En la política de actualización "Write Through", cada vez que se escribe un dato en la memoria principal (RAM), se escribe simultáneamente en la caché y en la memoria principal. Esto garantiza que la caché y la memoria principal siempre contengan los mismos datos, manteniendo la coherencia. Aunque esto puede parecer un enfoque simple y seguro, puede generar un rendimiento más lento en comparación con otras políticas, ya que cada escritura implica dos operaciones de escritura (una en la caché y otra en la memoria principal).

Ventajas:

- Mayor coherencia de datos entre la caché y la memoria principal.
- Menos probabilidades de perder datos en caso de fallos.

Desventajas:

- Mayor latencia debido a las operaciones de escritura adicionales.

Write Back (Escritura Diferida o Posterior): En la política de actualización "Write Back", cuando se escribe un dato en la caché, se marca el bloque de caché como modificado pero no se escribe inmediatamente en la memoria principal. La escritura en la memoria principal se realiza solo cuando se reemplaza el bloque de caché o cuando se necesita el espacio para otro bloque. Esto permite reducir la latencia y mejorar el rendimiento, ya que varias escrituras en la misma ubicación de memoria pueden agruparse y escribirse en la memoria principal en una operación única.

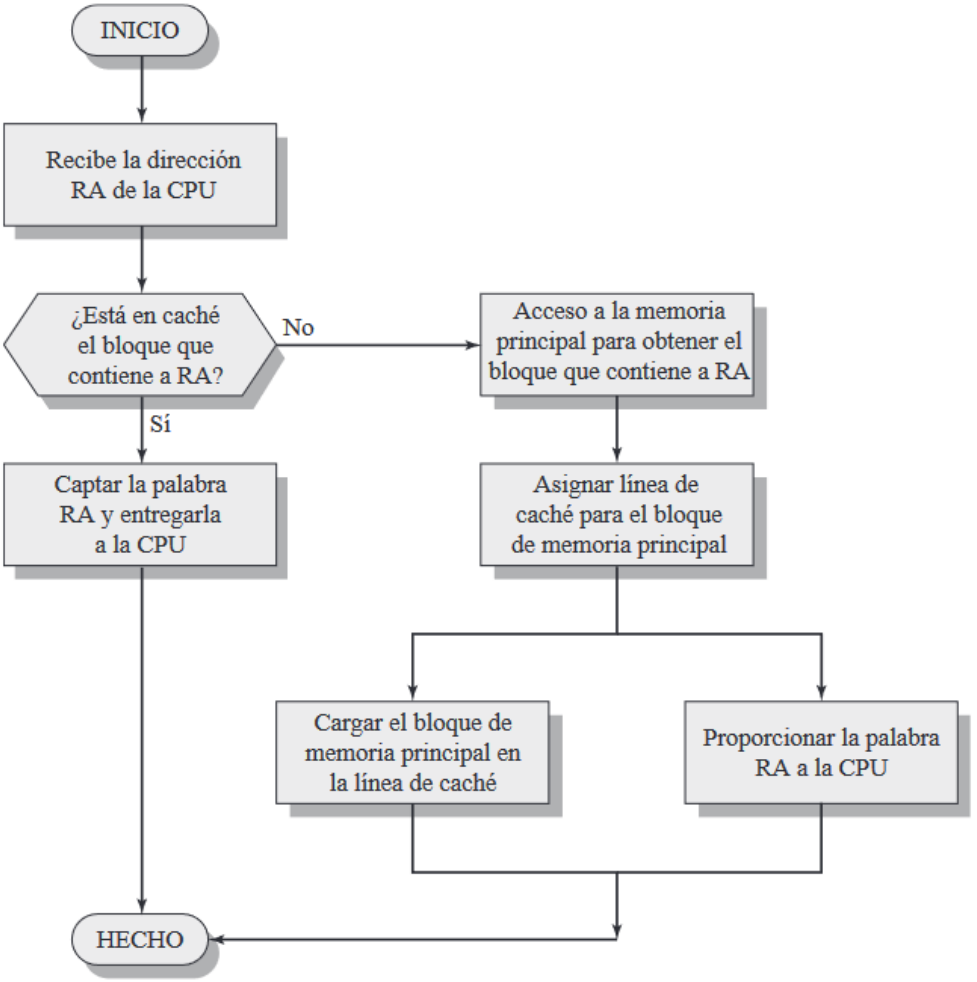
Ventajas:

- Menor latencia, ya que no se requiere una escritura inmediata en la memoria principal.
- Mejor rendimiento debido a la reducción de operaciones de escritura en la memoria principal.

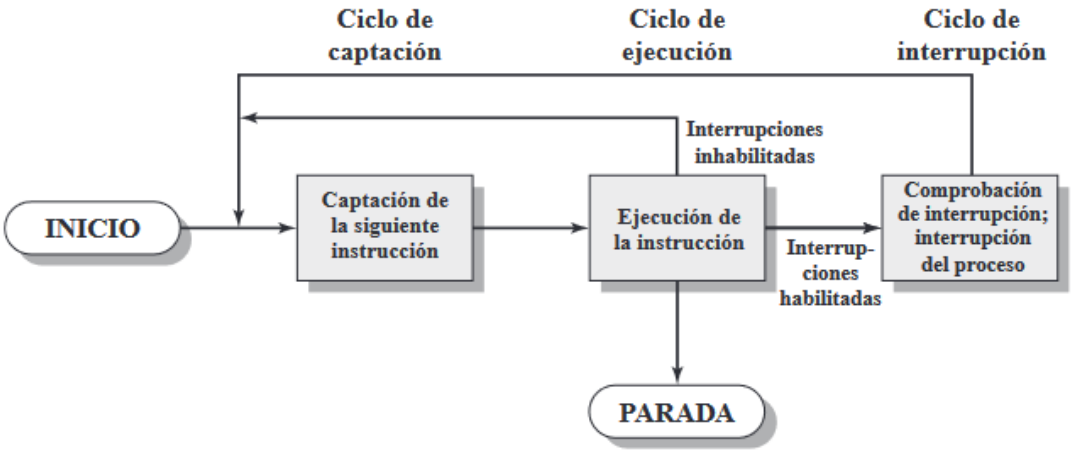
Desventajas:

- Posible pérdida de datos en caso de fallos antes de que se escriban los datos en la memoria principal.
- Puede ser más complejo mantener la coherencia de datos entre la caché y la memoria principal.

En resumen, la elección entre "Write Through" y "Write Back" depende de los objetivos de rendimiento, la complejidad del sistema y la importancia de la coherencia de datos. "Write Through" **es más seguro en términos de coherencia de datos, pero puede ser más lento.** "Write Back" **puede mejorar el rendimiento, pero requiere estrategias adicionales para manejar la coherencia de datos y puede tener riesgos asociados con la pérdida de datos en caso de fallos.**



Flujo de ciclo de instrucción con interrupción



Flujo de ciclo de instrucción básico

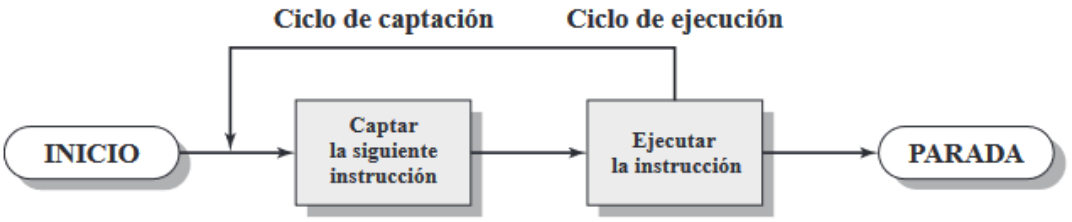


Figura 3.3. Ciclo de instrucción básico.

BUSES

Contiene tres grupos funcionales de líneas de comunicación
- Líneas de datos
- Líneas de direcciones
- Líneas de control

ELEMENTOS DE DISEÑO DEL BUS

Dedicadas: Una línea de bus dedicada está permanentemente asignada a la función específica. Ventajas, comunicación clara y directa entre componentes. La desventaja es que es más costosa porque hay que implementar mas lineas
Multiplexadas: combina varias líneas ahorrando espacio y costos. La desventaja es que se necesitan circuitos más complejos en cada módulo

Métodos de arbitraje del bus
Centralizado: un dispositivo denominado controlador del bus o árbitro asigna el tiempo del bus
Distribuido: cada módulo conectado al bus contiene lógica de control de acceso e interactúa con los otros para establecer quién utiliza el bus

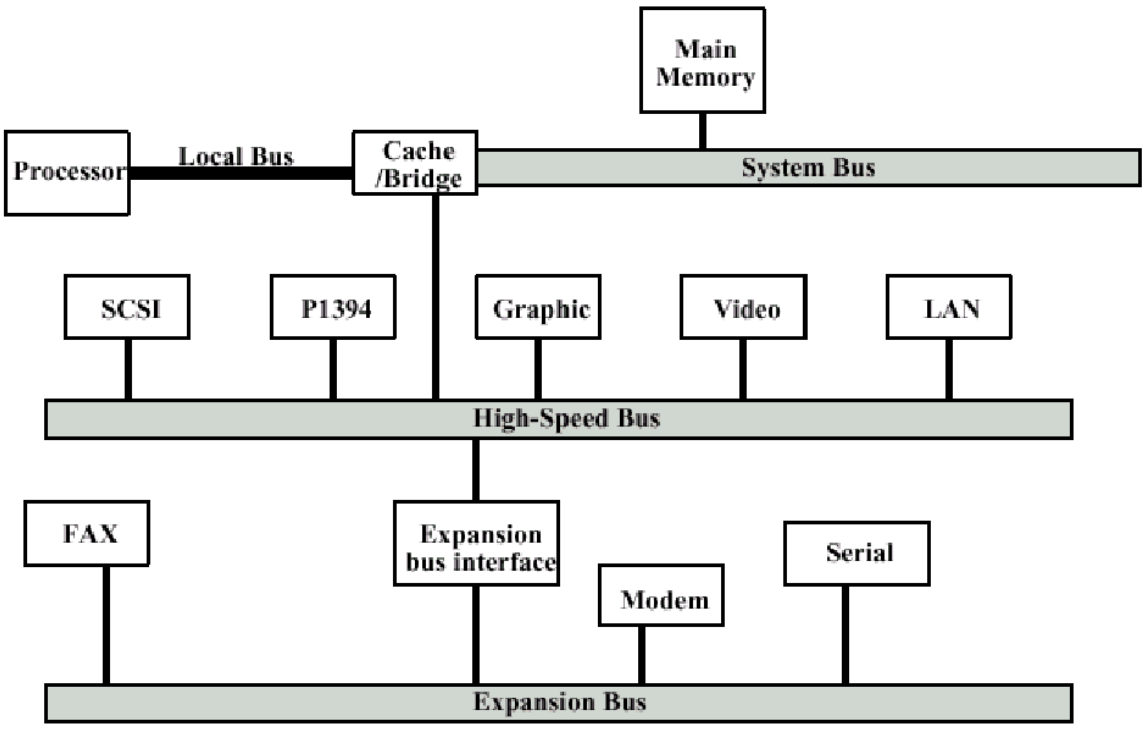
Timing - Forma en la que se coordinan los eventos

Sincrónico: La ocurrencia de eventos en el bus está determinada por un reloj que produce una secuencia de unos y ceros alternados de igual duración
- El reloj es accesible a todos los dispositivos conectados al bus
- La secuencia de un 1 y un 0 constituye un ciclo de bus (o ciclo de reloj)
- Los eventos se inician al comienzo de un ciclo de reloj

Asincrónico: La ocurrencia de un evento sigue a y depende de la ocurrencia de un evento previo

- Permite tomar ventaja de progresos en el rendimiento de los dispositivos y que una mezcla de dispositivos lentos y rápidos –que utilicen nuevas y viejas tecnologías- compartan el bus
- Es más difícil de implementar y probar que timing sincrónico

JERARQUÍA DE BUSES PARA ALTO RENDIMIENTO



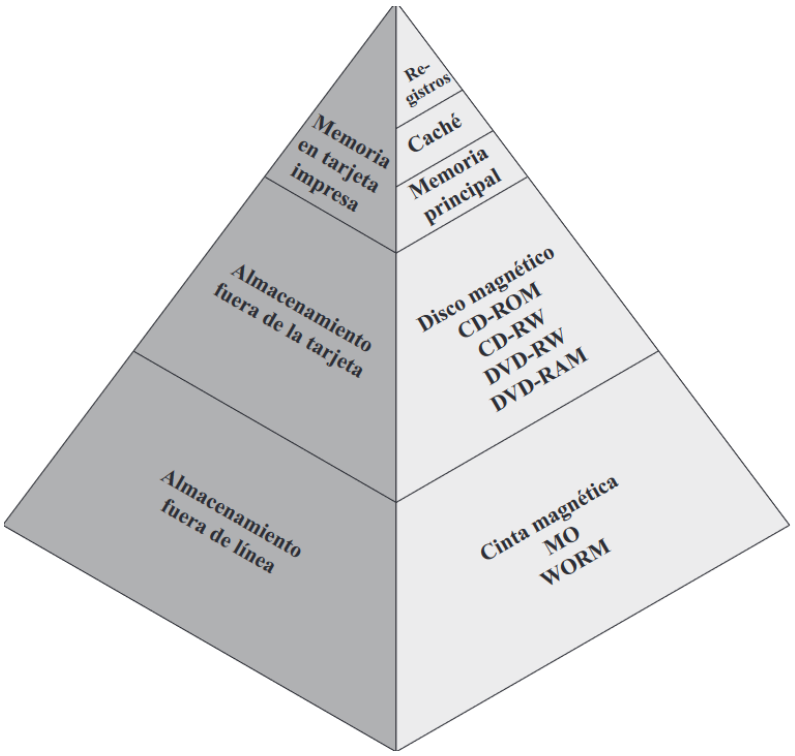
El bus de alto rendimiento se implementa para poder estar a la altura de los componentes de entrada y salida que cada vez son más rápidos

Memoria

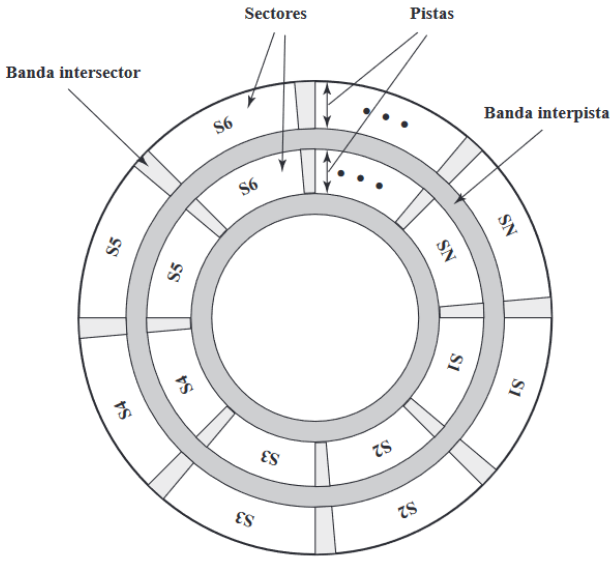
Existe un compromiso entre las tres características clave de coste, capacidad, y tiempo de acceso.

- A menor tiempo de acceso, mayor coste por bit.
- A mayor capacidad, menor coste por bit.
- A mayor capacidad, mayor tiempo de acceso.

Jerarquía de memorias



Disco magnético organización y método de formato



Básicamente, la superficie del disco magnético está compuesta por cientos de pistas las cuales son del tamaño del cabezal y entre estas se encuentra una banda interpista que ayuda a minimizar errores en caso de un desalineamiento del cabezal. Cada pista está dividida por sectores los cuales a su vez se separan con una banda intersector