Tema 1. Modelo genérico de un ordenador. Arquitectura de Von Newman. Estructura física de la CPU.

Índice de contenido

u

Introducción....2

La arquitectura de Von Newman....3

La memoria....4

La memoria principal....4

La memoria cache....5

La memoria secundaria o periférica....6

Periféricos de E/S....6

Los buses....7

Estructura física de la CPU....9

La unidad de control -UC-....9

El reloj del procesador....9

Los registros....10

La unidad aritmético-lógica -ALU-....11

Introducción

Un ordenador es una máquina electrónica de proceso de datos de uso general que trabaja a gran velocidad y con una gran fiabilidad. Las operaciones que efectúa se le indican con las instrucciones que componen los programas.

La parte física del ordenador se denomina "hardware", la parte lógica (datos e instrucciones) se denomina "software".

La arquitectura actual de los ordenadores se basa en los principios de la arquitectura de John Von Neuman, que son:

- Una sola memoria física para los datos y los programas.
- Los contenidos de la misma se direccionan indicando su posición, sin considerar el tipo de dato contenido en la misma.
- La ejecución de las instrucciones es secuencial, salvo que una instrucción ordene romper la secuencia (es decir, realizar un bote dentro del programa).

El primer ordenador en utilizar esta arquitectura fué el EDVAC en 1947, aunque la arquitectura ha ido evolucionando.

Al inicio de la informática se propusieron otro tipo de arquitecturas, como la arquitectura Harvard que proponía memorias físicas separadas para datos e instrucciones. Esta arquitectura se utiliza actualmente en el diseño de las memorias caché de primer nivel y en maquinas especializadas en el proceso digital de señales.

La arquitectura de un computador está orientada a conseguir que este funcione con eficacia al menor coste. Las funciones básicas de un ordenador son:

- **Procesar datos**: El ordenador efectúa operaciones aritméticas y lógicas sobre los datos almacenados en memoria principal.
- **Almacenar datos**: Guarda los datos sobre los cuales el ordenador está trabajando.
- **Transferir datos**: Los datos son transferidos entre el ordenador y el exterior, y entre sus componentes. La comunicación con el exterior se realiza utilizando los dispositivos de E/S del ordenador y la interna utilizando los buses.
- **Control**: Las tres funciones anteriores son controladas por el mismo ordenador siguiendo las instrucciones que componen el programa que ejecuta.

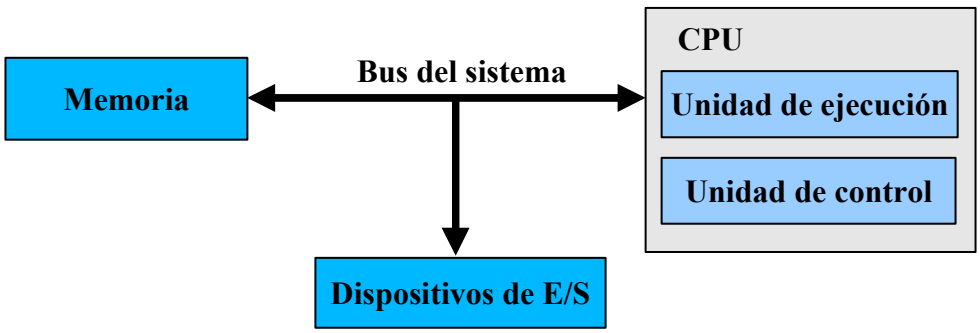
En la arquitectura de Von Newman aparece un tipo de componente específico para cada una de estas funciones.

La arquitectura de Von Newman

Los componentes de un ordenador con arquitectura de Von Newman son:

- **La memoria**: Es donde se almacenan los datos y las instrucciones que componen los programas. Es vista por el procesador como un conjunto de compartimentos numerados donde puede leer y escribir información.
- **El procesador:** Es el encargado de ejecutar las operaciones que indican las instrucciones de los programas y controlar el resto del ordenador para que se cumplan estas instrucciones. Está compuesto de dos partes:
  - La *Unidad de Control*. Se encarga de descodificar las instrucciones y emitir las señales de control apropiadas para que se ejecuten.
  - La *Unidad de ejecución*: Es la que realmente ejecuta las operaciones que indican las instrucciones, conforme a las señales que recibe de la unidad de control.

  La arquitectura original de Von Newman considera estas dos unidades como elementos diferentes.
- **Los dispositivos de E/S:** Proporcionan al ordenador la comunicación con el exterior y la capacidad de almacenamiento permanente de información. Ejemplo: El teclado, un disco duro, la impresora...
- **El bus del sistema:** Es el elemento que interconecta los componentes del ordenador y permite la comunicación entre ellos.



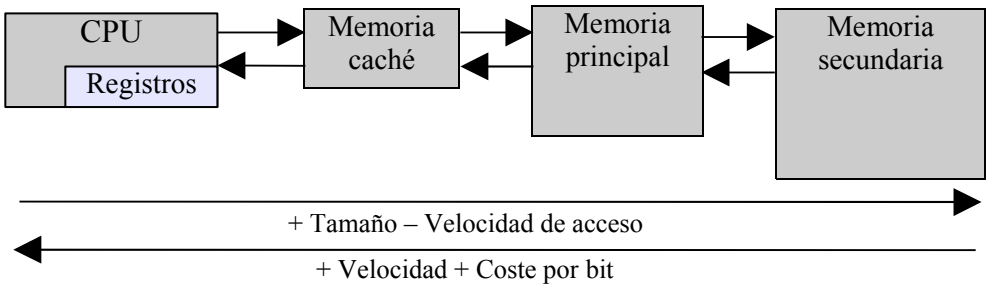
La memoria

Almacena los datos y los programas que utiliza el procesador.

Existen diferentes tipos de memorias que se diferencian entre ellas en su capacidad y en la velocidad de acceso. Siendo las memorias más rápidas, las de menos capacidad.

La memoria se organiza de forma jerárquica. Se reparte entre los dispositivos de forma que se consiga la mayor velocidad posible, al menor coste por bit almacenado.

Esto se consigue gracias al principio de localidad de referencias, según el cual, un programa al ejecutarse accede sólo a una pequeña parte de la memoria durante un periodo relativamente largo. Esta parte accedida, se guarda en la memoria más rápida. Se consigue que el sistema tenga una velocidad de acceso cercana al dispositivo más rápido a un precio por bit del dispositivo más barato.

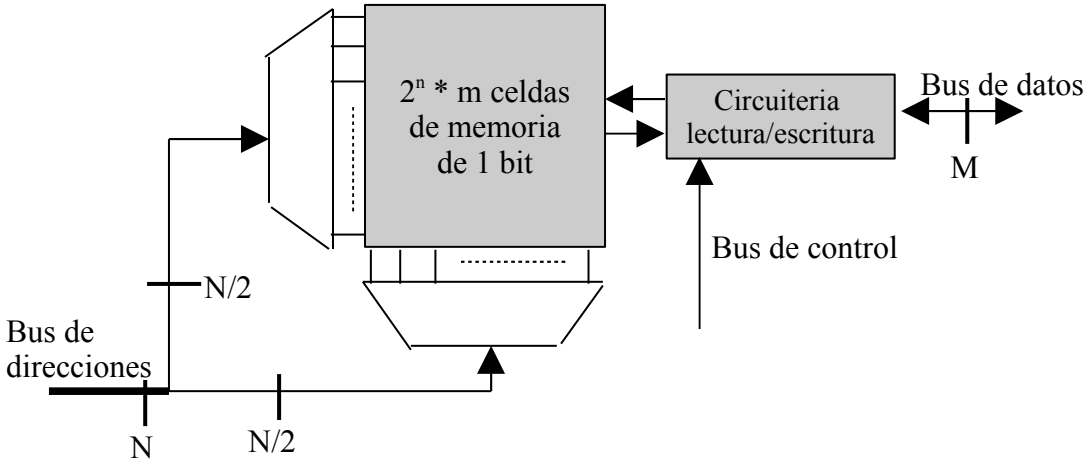


La memoria principal

Contiene los datos y el código de los programas que está utilizando a corto plazo la CPU. Es vista por la CPU como un conjunto lineal de compartimentos numerados de igual tamaño (tamaño de palabra de la memoria) a los que puede acceder aleatoriamente indicando el número (dirección) del compartimento. Como el acceso es aleatorio, también son conocidas como memorias RAM -Random Access Memory-.

Estas memorias, actualmente, son volátiles, es decir, pierden la información almacenada cuando se quedan sin alimentación eléctrica.

La organización de la memoria suele ser de dos dimensiones, para simplificar la circuitería de descodificación de direcciones. Cada bit se almacena en una celda que puede estar implementada por transistores (memorias estáticas) o condensadores (memorias dinámicas).



Esta memoria maneja palabras de m bits, y puede manejar un máximo de 2^n palabras. Su capacidad máxima es de 2^n*m bits.

La memoria cache

Esta memoria se sitúa entre el microprocesador y la memoria principal. Se utiliza para almacenar una copia parcial del contenido de la memoria principal más utilizado.

Esta memoria permite agilizar el trabajo del microprocesador pues trabaja casi a su misma velocidad. Es una memoria que es invisible a los programas que ejecuta el ordenador.

Cuando la CPU accede a una palabra de memoria, el sistema de acceso a memoria la busca en la caché. Si no la encuentra en la memoria caché, trae el bloque, que contiene la palabra, de la memoria principal a la memoria caché.

Así el tiempo medio de transferencia de una palabra es:

$T_{medio} = T_{transferencia\ desde\ caché} + (1\text{-Probabilidad de acierto})T_{transferencia\ de\ memoria\ principal\ a\ caché}$

La probabilidad de acierto es muy alta, por lo que el tiempo medio de acceso a memoria se acerca al tiempo de acceso a caché.

En un ordenador actual pueden aparecer varios **niveles de memoria caché**:

1. *Nivel 1*. Es la memoria caché de menor tamaño y de velocidad similar a la CPU. Se encuentra integrada dentro de la misma CPU y tiene un diseño siguiendo al arquitectura Harvard, datos e instrucciones en memorias separadas, para permitir el acceso simultáneo a ambos (Necesidad de los diseños superescalares).

- *Nivel 2*. Memoria caché más lenta y de mayor capacidad que la de primer nivel. Almacena de forma conjunta datos e instrucciones. Es externa a la CPU y en los diseños actuales tiene un bus de comunicaciones propio e independiente.
- *Nivel 3*. Aparece en los ordenadores multiprocesador. Es una memoria caché compartida por varias CPUs.

Hay tres tipos de cachés según su **política de ubicación** -dónde se coloca el bloque de memoria leído-:

- *Correspondencia directa*. Cada bloque de memoria principal tiene una sola posición en caché donde puede ubicarse. Son cachés de diseño muy simple, y de menor prestaciones que el resto.
- *Completamente asociativa*. Se puede almacenar cualquier bloque de memoria principal en cualquier posición de la caché. Es la que obtiene un mayor rendimiento, pero son caras por tener un diseño complejo.
- *Asociativa por conjuntos*. La memoria caché se divide en varios conjuntos de N bloques. Cada bloque de memoria tiene asociado un conjunto de bloques de la memoria caché y podrá ubicarse en cualquier posición dentro de él. Es la política más utilizada al tener la mejor relación rendimiento/coste.

La memoria secundaria o periférica

Es la memoria que almacena la información a largo plazo, que no está necesariamente en uso. Sus características generales son:

- Memoria no volátil. Aunque el ordenador se apague sigue almacenando la información.
- Gran capacidad de almacenamiento.
- Velocidad de acceso menor que la memoria principal
- Coste por bit almacenado menor que la memoria principal.
- La mayoría de los dispositivos realizan el acceso de forma secuencial a los datos.
- Diferentes formas de almacenar la información: magnético, óptico y eléctrico.

Los dispositivos actuales de almacenamiento son:

- Los **discos duros:** Es el principal medio de almacenamiento no volátil. Suele ser un periférico interno, no extraíble. La información se almacena como campos magnéticos en la superficie de unos discos metálicos. El acceso a la información se realiza de forma secuencial. Permite almacenar cientos de Gigabits.
- Los **discos ópticos** CD-ROM, DVD, CD-RW...: La información se almacena en la superficie de un disco de forma óptica. Existen discos que sólo permiten la lectura, de una sola escritura o que permiten varias escrituras. Son dispositivos extraíbles utilizados para almacenar y transportar de forma física la información entre ordenadores, sobretodo para el almacenamiento multimedia.
- **Memorias flash**. Memoria EEPROM -Electrical-Erasable Programmable ROM- que se ha popularizado recientemente. El acceso a la información se realiza de forma aleatoria, a diferencia del resto de memorias secundarias. Son memorias ligeras y de bajo consumo.

Periféricos de E/S

Son los elementos de los ordenadores utilizados para la comunicación de este con el exterior. Existe una gran variedad de dispositivos diferentes de características muy diferentes.

Los dispositivos se pueden clasificar utilizando diferentes criterios:

- Según el tipo de comunicación: (De salida: un monitor, una impresora..., de entrada: un ratón, un teclado..., de entrada/salida: un modem, una tarjeta de red...)
- El ancho de banda utilizado: (Gran ancho de banda: El monitor, la tarjeta de red..., poco ancho de banda: El teclado, el ratón...)

La clasificación de un periférico como periférico de almacenamiento o de E/S no está perfectamente definida. Existen periféricos, como una grabadora de CD, que tienen ambas funciones.

Uno de los requisitos del diseño de la E/S en los computadores es conseguir manejar de una forma más o menos homogénea la gran variedad de dispositivos de E/S.

Los periféricos no se conectan directamente al bus del ordenador, sino que se conectan a través de unos dispositivos puente, denominados ***puertos de E/S***. Por ejemplo: los puertos PCI o USB. El puerto de E/S tiene las siguientes funciones:

1. Emitir las señales de control hacia el periférico adecuadas a las instrucciones recibidas desde la CPU.
2. Regular a través de búferes, las diferentes velocidades del periférico y de la CPU. Los periféricos suelen ser más lentos que la CPU.
3. Realizar las conversiones que sean necesarias, tanto eléctricas como de codificación de la información.

Los buses

Es el medio de comunicación interno del ordenador e interconecta todos los componentes del mismo. Está formado por un conjunto de conductores eléctricos, por donde circulan las señales que corresponden a la información que trata el ordenador. Estos buses internos transportan la información de forma paralela.

Hay tres clases de líneas en un bus:

- *Las líneas de datos*: Proporciona el camino para transmitir información (datos e instrucciones) entre los componentes del ordenador. Suele constar de 32 o 64 líneas distintas (Anchura del bus).
- *Las líneas de direcciones*: Indica la fuente o destino del dato situado en el bus de datos. Los ordenadores actuales tienen mapeadas en el mismo conjunto de direcciones, las direcciones de la memoria y de los dispositivos E/S.
- *Las líneas de control*: Se utiliza para que la CPU controle al resto de componentes y para sincronizar el acceso y el uso de los buses de datos y direcciones. A través de este bus se envían señales como: Lectura Memoria, Escritura E/S, Petición de Interrupción...

La arquitectura tradicional de los buses del ordenador era tener un bus único que conectaba la CPU con la RAM y el resto de periféricos. Siendo la CPU la encargada de controlar el acceso de todos los componentes al bus del sistema. Este diseño limitaba la velocidad del bus a la del componente más lento y hacía que el diseño de los buses fuera complejo al tener que conectar dispositivos con comportamientos muy heterogéneos.

En la arquitectura original de *Von Newman*, sólo hay un bus en el ordenador, formado por los tres tipos de líneas comentadas anteriormente. Pero actualmente, un ordenador se compone de varios tipos de buses que se interconectan y pueden trabajar de forma simultánea. Actualmente, un ordenador se compone de varios tipos de buses:

- El ***frontside bus***: Conecta el procesador con la memoria caché de nivel 2.
- El ***backside bus***: Es el conjunto de líneas que conectan directamente la CPU con la memoria principal. De esta forma se evita el cuello de botella que suponen la lentitud del resto de componentes. Está optimizado para transferencias del tamaño de un bloque de caché.
  El *frontside bus* y el *backside bus* pueden funcionar de forma simultánea.
- El ***bus del sistema***: Interconecta los dispositivos de alta velocidad. Es más largo y más lento que *backside bus*, al cual se interconecta a través de un puente.
- Un ***bus de expansión***: Interconecta los dispositivos más lentos del sistema. Se conecta al bus de sistema a través de puente-.

Los dispositivos que se conectan a un bus pueden ser maestros, si pueden tomar la iniciativa de tomar el control del bus e iniciar una transmisión, o dispositivos esclavos, cuando tienen que esperar a recibir una solicitud. Hay dispositivos que actúan como maestros y esclavos, ejemplo la CPU, y otros sólo como esclavos, ejemplo la memoria.

Según los ciclos de tiempo, los buses pueden ser:

1. ***Síncronos***: Una de sus líneas transmite una señal periódica, frecuencia del bus, y todas las operaciones en el bus están sincronizadas a esta señal periódica.
2. ***Asíncronos***: No tienen reloj. Las operaciones con el bus no están sincronizadas, las operaciones de los diferentes dispositivos no tienen porque durar lo mismo. Los dispositivos negocian antes de iniciar la comunicación la velocidad máxima que puede alcanzar. De esta forma, la comunicación entre dispositivos rápidos no está condicionada a que haya dispositivos lentos. El inconveniente de estos buses es su complejidad y alto coste.
3. ***Semisincrono***, el bus tiene una señal de reloj, y las operaciones se sincronizan con los ciclos del reloj. Cuando un dispositivo lento utiliza el bus y no puede completar la comunicación en los ciclos establecidos, activa una señal de espera para obtener un ciclo más.
   Este sistema permite tener un bus que no va a la velocidad del dispositivo más lento, sin tener que realizar las complejas operaciones de negociación de la velocidad de comunicación.

El bus, al ser un recurso compartido, necesita implementar una política de arbitraje para cuando varios dispositivos reclamen su uso de forma simultánea.

Para negociar el turno de utilización del bus aparecen los mecanismos de arbitraje de bus. El arbitraje del bus puede ser:

- **Centralizado:** Un dispositivo árbitro determina que dispositivo tiene acceso al bus. Existen diferentes diseños:
  - Con **una sola línea de solicitud** de bus**.** Daisy chain: Es un sistema de arbitraje con prioridades fijas, Polling (encuesta): con prioridades dinámicas.
  - Con **líneas independientes** para cada dispositivo. Más rápido y más caro que el anterior.
- **Descentralizado**: En este caso no existe un árbitro. Por diseño del sistema de buses, cada dispositivo tiene su prioridad. Evita tener un árbitro, pero su diseño es complejo.

Estructura física de la CPU

La CPU -Unidad central de proceso- es el elemento del ordenador encargado de ejecutar las instrucciones que componen los programas. Para ello, realiza una serie de operaciones -leer de la instrucción, interpretarla, buscar los operandos, ejecutar la operación que indica la instrucción y guardar el resultado - que componen el *ciclo de instrucción.*

Actualmente se fabrican como un chip, denominado microprocesador, compuesto de circuitos digitales de alta densidad. Anteriormente, la CPU se componía de una o varias tarjetas de circuitos digitales.

Una CPU puede ejecutar un conjunto de instrucciones determinado, denominado *juego de instrucciones* de la CPU. Las instrucciones tienen un *formato de instrucción* determinado, unos bits indican que operación se debe realizar y otros indican los operandos utilizados en la operación. El juego de instrucciones de una CPU debe ser completo, es decir, debe permitir implementar cualquier funcionalidad.

La unidad de control -UC-

La unidad de control de la CPU tiene dos funciones:

- El *secuenciamiento de instrucciones*: Determina la siguiente instrucción a ejecutar. Detecta los saltos dentro del código del programa, tanto condicionales como incondicionales, y actualiza el registro PC para que se efectúen.
- La *interpretación de las instrucciones*: Decodifica las instrucciones y genera las señales de control necesarias para que sean ejecutadas.

Las CPU actuales **contienen** más de una unidad de ejecución y cada una de ellas tiene su propio controlador. La UC funciona como el supervisor de todas las unidades de ejecución. Procesadores superescalares.

Existen dos formas de implementar las UC:

- Las *UC cableadas*. La UC está formada por un circuito digital que implementa una máquina de estados finitos. Los bits que forman la instrucción son la entrada de la máquina de estados finitos, esta va pasando de un estado a estado emitiendo las señales de control necesarias.
  Estas UC son las utilizadas en los procesadores RISC. Son rápidas y baratas de fabricar, pero poco flexibles.
- Las *UC μprogramadas*. La UC tiene una pequeña memoria ROM con un μprograma -formado de μinstrucciones- que se ejecuta para decodificar la instrucción y emitir las señales de control.
  Es un sistema flexible, pues permite cambiar el juego de instrucciones de la CPU cambiando el μprograma, y permite tener un repertorio de instrucciones complejas, por ello son utilizadas en las CPUs CISC. Son más lentas y caras que las UC cableadas.

El reloj del procesador

La CPU funciona de una forma sincronizada, según los pulsos que le transmite un reloj.

El reloj se compone de un oscilador de cuarzo capaz de generar pulsos eléctricos a un ritmo constante llamados ciclos. Que se miden en Hertz (ciclos por segundo).

La duración de un ciclo viene determinada por la operación elemental más lenta. Por ello, el diseño de las CPU busca que todas las operaciones elementales tarden lo mismo. (Técnicas de segmentación de las operaciones)

Los registros

Es la memoria interna de la CPU, formada por un conjunto de registros. Es la memoria más rápida del ordenador y la de menor capacidad. Es utilizada para almacenar los datos con los que está operando la CPU y su información de control y estado. Esta memoria se encuentra en la cima de la jerarquía de memoria de un ordenador.

Existen 2 tipos de registros dentro de la CPU:

1. *Registros de propósito general -GPR-*. Son utilizados por los programas para almacenar temporalmente información, ya sean datos o direcciones de memoria. Utilizando estos registros se mejora la velocidad de ejecución del programa, pues se evita acceder a la memoria principal, que más lenta que la CPU.

   Muchas arquitecturas de CPU incluyen dos tipos de registros de propósito general, registros para números enteros y registros de más capacidad para números en punto flotante.

   Existen CPU – CPUs memoria-memoria- que no tienen registros directamente accesibles por los programas. Las CPU actuales disponen de muchos registros accesibles. Ej: La mayoría de RICS tienen 32 registros GPR.

2. *Registros de propósito específico*. Son utilizados por la UC para controlar el funcionamiento de la CPU. Son accesibles sólo a los programas en modo privilegiado. Los registros más comunes de este tipo son:

   - El *contador del programa -PC-*. Es el registro interno que almacena la dirección de la próxima instrucción a leer o de la última instrucción leída, según el diseño de la CPU. De esta manera la UC puede saber cuál es la siguiente instrucción que debe ejecutar.

     En la mayoría de CPUs, el PC va incrementándose en una unidad de forma automática, salvo que una instrucción de salto cambie el flujo del programa. En este caso, el PC se actualiza con esta nueva dirección.

     El incremento del PC puede ser de una, 2, o más posiciones, según el tamaño de las instrucciones usadas.

   - El *Registro de Instrucción -IR-*. Almacena la instrucción que está ejecutando en ese momento la CPU.

   - El *Registro de dirección de memoria -MAR-*. Contiene la dirección de la posición de memoria a la que quiere acceder la CPU. Su contenido se transmite por el bus de direcciones al efectuar la operación de lectura o escritura.

   - El *Registro intermedio de memoria -MBR-*. Es registro que almacena el dato transferido hacia o desde la memoria principal. Funciona de buffer y permite independizar el funcionamiento de la CPU de la memoria principal.

   - El *Registro de estado del procesador -PSW-*. Es un conjunto de bits donde cada uno de ellos sirve de 'flag' para indicar el estado de la última operación realizada. Cada arquitectura de procesador tiene sus flags, los habituales son: la Z indica si la última operación ha sido cero, N si ha sido negativa... Estos valores son los evaluados en los saltos condicionales.

- El *Registro acumulador de la ALU*, destino del resultado de última operación de la ALU. En CPU sencillas, como microcontroladores, el acumulador puede utilizarse como operando en las operaciones. En las CPU avanzadas, sólo almacena el resultado de forma temporal, hasta que es trasladado a otro registro.

La unidad aritmético-lógica -ALU-

Se denomina Unidad Aritmético-Lógica a la unidad incluida en la CPU encargada de realizar operaciones aritméticas y lógicas sobre operandos que provienen de la memoria principal y que pueden estar almacenados de forma temporal en algunos registros de la CPU.

Físicamente, la ALU se compone de una serie de circuitos electrónicos que implementan las operaciones aritmético-lógicas.

Junto con sus registros auxiliares de entrada de los operandos y de salida del resultado forma la unidad de ejecución del procesador.

Las operaciones básicas que realiza una ALU son:

- Operaciones aritméticas: suma, resta y multiplicación.
- Operaciones lógicas: And, or, not y exor.
- Desplazamiento de bits, con o sin mantenimiento del bit de signo.

Funcionamiento:

La ALU efectúa sus operaciones, una vez que los operandos ya se encuentran en registros de la CPU, sean registros de propósito general o el MBR. La UC indica a la ALU que operación efectuar activando una señal de control, el resultado de la operación se almacena en un registro acumulador de la ALU.

Tras cada operación, la ALU actualiza en contenido del registro de *estado del procesador*.

Suele haber dos tipos de UAL: especializadas en operaciones con números enteros -generalmente en formato complemento a 2- y otra con números en coma flotante.

Hace algunos años, existía el denominado coprocesador matemático, una UAL especializada en cálculos con números reales que estaba en un microchip diferente al de la CPU.