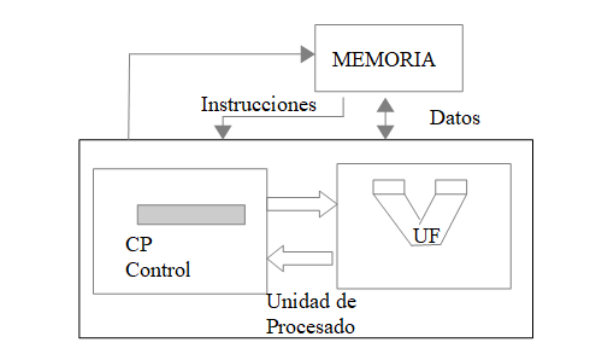
LIBRO 1

Máquina Von Neumann

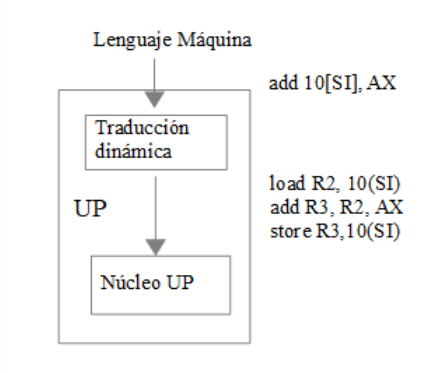
Una máquina von Neumann dispone de posiciones de almacenamiento direccionables y sus valores pueden modificarse mediante un conjunto de instrucciones. La secuencia de control se implementa almacenando las instrucciones en memoria y utilizando un puntero (CP), el cual indica la instrucción que debe interpretarse.



UP: Unidad de Procesado, incluye el control y las unidades funcionales (**UF**) que se utilizan para manipular los valores almacenados en memoria.

Familia: conjunto de máquinas que interpreta un mismo lenguaje máquina (**LM**)

Las instrucciones que interpreta el núcleo de una UP leen los datos fuente del banco de registros y almacenan el resultado en un registro



PRESTACIONES

Latencia: Es el tiempo entre la solicitud de un dato a memoria y la disponibilidad del dato en la UP

Ancho de banda: Es el número de bytes que se transmiten por unidad de tiempo

PROPIEDAD DE LOCALIDAD Y JERARQUÍA DE MEMORIA

PROPIEDAD DE LOCALIDAD

Localidad temporal: Existe una alta probabilidad de que cuando se accede a una posición de memoria esta sea accedida en un futuro cercano

Localidad espacial: Existe un alta probabilidad de que cuando se accede a una posición de memoria sean accedidas, en un futuro cercano, posiciones de memoria cuya dirección es cercana

La idea es explotar esta localidad para reducir la latencia media de acceso a memoria con una jerarquía de memoria determinada

JERARQUÍA DE MEMORIA

A medida que nos alejamos de las UF se incrementa la capacidad, el tiempo de acceso y se reduce el ancho de banda.

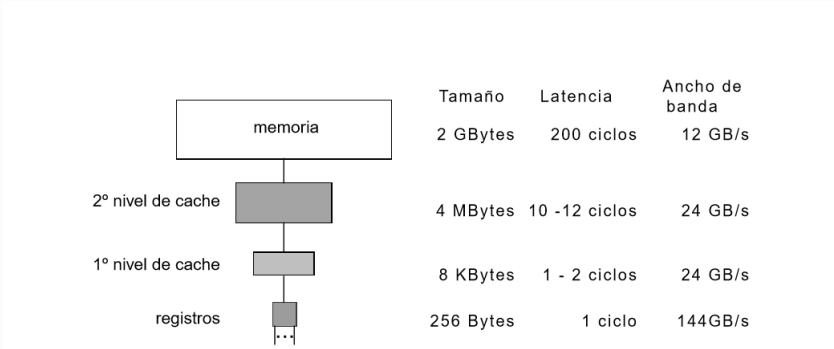


Figura 1.6 *Jerarquía de memoria. Los ciclos se corresponden con ciclos de UP. Se supone un procesador cuya frecuencia de funcionamiento es 1.5 Ghz y que en cada ciclo puede efectuar 8 lecturas y 4 escrituras al banco de registros. El primer nivel de cache tiene dos caminos de acceso de 8 bytes. El segundo nivel de cache tiene un camino de acceso y puede transmitir 32 bytes cada 2 ciclos. Para acceder a memoria se dispone de un camino de acceso que permite transmitir 64 bytes cada 8 ciclos.*

Primer nivel BR, se gestiona el contenido mediante las instrucciones LM, este explota la localidad temporal de las refs a datos. Los siguientes son MC y la gestión de contenidos es usualmente transparente al LM. Algunos procesadores tienen inst de prebúsqueda para traer info desde memoria a alguna cache antes de que se necesite. Los programas no se encuentran en su totalidad en memoria, pero sí en disco y es el SO quien gestiona sus recursos. Las MC explotan la localidad temportal y espacial a las refs de datos.

CMA = cacierto + fR * Pf (Ciclos Medio de UP)

MÉTRICAS DE RENDIMIENTO

RENDIMIENTO DE UN PROCESADOR

Hay varias técnicas que afectan casi solo a un factor. Como que el compilador efectúe optimizaciones que reduzcan el número de inst. Mejoras en la tecnología con el fin de reducir el retardo de propagación de las señales y reducir el tc sin incrementar el CPI. También una organización jerarquía de memoria que explote la localidad de refs a memoria, reduce el CPI. El paralelismo para reducir el CPI, estos procesadores suelen tener HW que utilizan la segmentación adicionalmente

T =	N	x	CPI	x	t_c
	compilador		lenguaje máquina		organización
	lenguaje máquina		organización		tecnología

Figura 1.7 *Dependencias de los factores de la expresión del tiempo de ejecución con el compilador, el lenguaje máquina, la organización y la tecnología.*

T esta formada por los ciclos que utiliza la UP y los ciclos que UP espera para completar los accesos a Memoria (CM)

CM = N * fl * Pf o N * Rl * fR * Pf (Fallos/Inst o Accesos/Inst * Fallos/Accesos)

En la Figura 1.10 se relacionan dependencias de los factores de la expresión anterior con el compilador, lenguaje máquina, organización, jerarquía de memoria y tecnología.

T =	N	x	($CPI\|_{UP}$	+	R_l	x	f_R x P_f)	x	t_c
	compilador		lenguaje máquina		lenguaje máquina		jerarquía de memoria		organización
	lenguaje máquina		organización		compilador				tecnología

Figura 1.10 *Dependencias de los factores de la expresión del tiempo de ejecución con el compilador, el lenguaje máquina, la organización, la jerarquía de memoria y la tecnología.*

LEY DE AMDAHL

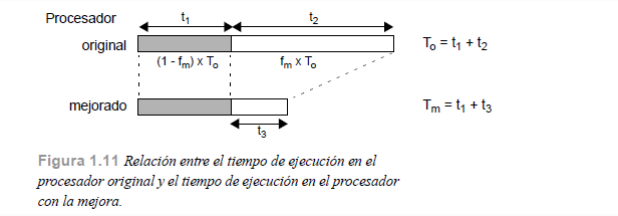


Figura 1.11 *Relación entre el tiempo de ejecución en el procesador original y el tiempo de ejecución en el procesador con la mejora.*

$$Ganancia = \frac{T_0}{T_m} = \frac{T_0}{T_1 + T_3}$$

Como fm = t2/To, tenemos que t1 = (1 - fm)x T0. Entonces, el tiempo con la mejora se puede expresar como Tm = (1- fm) x To + t3.

Sustituyendo en la expresión que calcula la ganancia obtenemos. Como gm = t2/t3 y t2 = fm x To, obtenemos t3 = fm x (To/gm).Efectuando sustituciones en la expresión que calcula la ganancia obtenemos:

$$\text{Ganancia} = \frac{T_o}{t_1 + t_3} = \frac{T_o}{(1 - f_m) \times T_o + \frac{f_m \times T_o}{g_m}} = \frac{1}{(1 - f_m) + \frac{f_m}{g_m}}$$

CONCURRENCIA

Técnicas de paralelismo y segmentación para incrementar el rendimiento cuando se interpretan instrucciones en la UP y cuando se accede al sistema de memoria.

INTERPRETACIÓN DE INSTRUCCIONES

El procesador es binario compatible quiere decir que la mejora es transparente a nivel de código máquina.

Segmentación

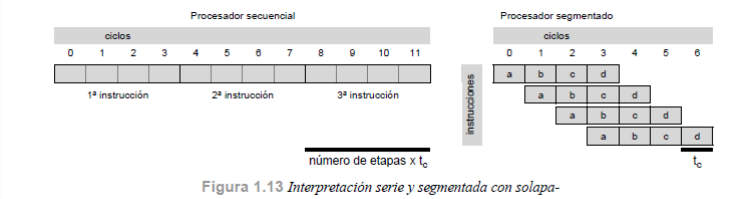


Figura 1.13 *Interpretación serie y segmentada con solapamiento de varias instrucciones.*

Es semejante a una cadena de montaje, cada paso o etapa de la segmentación realiza una parte de la interpretación de una instrucción. En cada ciclo se inicia la interpretación de una nueva instrucción. Procesadores segmentados escalares

En un caso ideal:

$$Ganancia = \frac{N.etapas * t_c}{t_c} = N.etapas$$

Numerador tiempo de eje. de una instrucción en un procesador serie y el denominador de una instrucción segmentada. En este caso vemo que finaliza una instrucción cada ciclo (tc)

Paralelismo

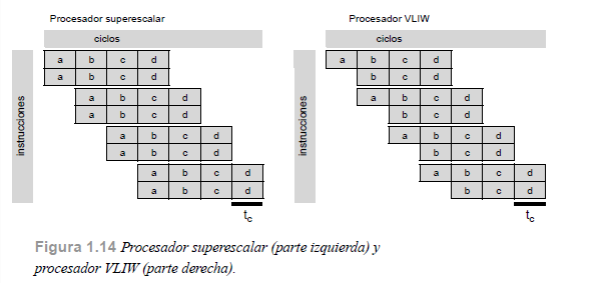


Figura 1.14 *Procesador superescalar (parte izquierda) y procesador VLIW (parte derecha).*

Varias instrucciones están en la misma etapa de interpretación. Procesadores superescalares

Cuando se utiliza paralelismo y segmentación, como en la parte izquierda, la ganancia ideal sería:

$$Ganancia = \frac{N.etapas * t_c}{t_c / N.Inst\ por\ ciclo} = N.Inst\ por\ ciclo * N.etapas$$

Procesador VLIW: Las inst. especifican la ejecución de varias operaciones. Las operaciones agrupadas en una instrucción son independientes entre sí y los tipos de operaciones que se pueden agrupar en una instrucción están restringidos por las UF disponibles.

Mermas en la ganancia

Para los casos ideales se ha supuesto que la segmentación o paralelismo no incrementa el tiempo de interpretación de una instrucción. También se ha supuesto que los recursos del procesador no están limitados y que las instrucciones que se están interpretando concurrentemente son independientes entre sí.

Los ciclos en que se reduce el grado de concurrencia en la interpretación
de instrucciones se denominan ciclos perdidos.

Multihilo

Los procesadores segmentados y superescalares pueden perder bastantes ciclos por situaciones de riesgo debido a que hay que mantener la semántica del lenguaje máquina que se interpreta. En lugar de que estos ciclos se desaprovechen, un procesador multihilo los utiliza para interpretar instrucciones de otros hilos.

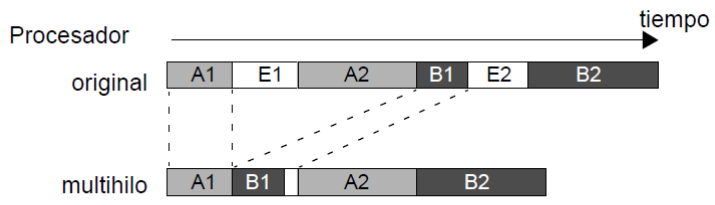


Figura 1.15 *Procesador que interpreta las instrucciones de forma serie y procesador multihilo.*

En E1 se produce un fallo de cache, y mientras se espera al dato, se efectúa B1, para no perder tiempo.

Tipos de multihilo:

- Grueso: En un intervalo de tiempo determinado sólo se están interpretando instrucciones de un hilo.
- Fino: Existen varios hilos activos, pero en un ciclo determinado sólo se inicia la ejecución de instrucciones de un único hilo
- Multihilo simultáneo: Existen varios hilos activos y en cada ciclo se pueden estar ejecutando instrucciones de varios hilos.

La técnica multihilo requiere que la jerarquía de memoria soporte accesos concurrentes.

ACCESO A LA JERARQUÍA DE MEMORIA

Cache no bloqueante

Las cache no bloqueantes permiten servir solicitudes de acceso que son aciertos mientras se está sirviendo un fallo de cache. Este tipo de cache con la inclusión de instrucciones de prebúsqueda en el lenguaje máquina permite soportar la latencia de fallo de cache. Cuando se supera el límite de fallos permitidos, en algunos procesadores, la instrucción de prebúsqueda se convierte en una instrucción nop. Esto es, la instrucción finaliza la interpretación después de detectar el fallo.

Load no bloqueante

La UP sigue interpretando instrucciones después de un fallo de cache, mientras no se interprete una instrucción que utilice como dato fuente el dato que se está esperando cargar en un registro.

Ancho de banda

Memorias multipuerto: se pueden efectuar varios accesos a la misma posición de memoria en paralelo.

Cuando en un nivel de la jerarquía se permite más de un fallo de cache pendiente de servicio, el servicio de los fallos es serie.

Una posibilidad para soportar accesos concurrentes a la información almacenada en un nivel de la jerarquía de memoria es utilizar varios bancos de memoria con un único puerto de acceso por banco y acceso independiente a cada banco (multibanco). permite servir de forma solapada accesos concurrentes si acceden a bancos distintos. Para determinar el banco en el cual se almacena un dato se utiliza entrelazado por dirección. Usualmente se utilizan los bits menos significativos de la dirección para determinar el banco.

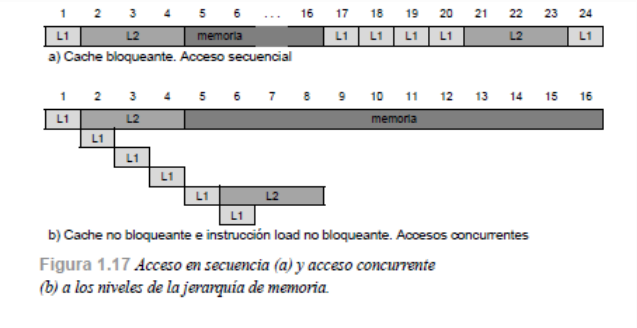


Figura 1.17 *Acceso en secuencia (a) y acceso concurrente (b) a los niveles de la jerarquía de memoria.*

Tendencia en la jerarquía de memoria

En algunos procesadores el acceso a los dos primeros niveles de la jerarquía de memoria se efectúa en paralelo. Reduce la latencia efectiva al segundo nivel de cache en caso de fallo en el primer nivel de cache y acierto en el segundo nivel de cache.

TECNOLOGÍA DE FABRICACIÓN

La tecnología de fabricación utilizada actualmente es CMOS y el elemento básico es el transistor.

POTENCIA Y ENERGÍA

Energía (J)

Potencia (W = J/s)

La potencia consumida en circuitos CMOS tiene tres componentes: conmutación, corriente de fugas y corriente de cortocircuito.

Conmutación -> P = E * f = C * V^2 * f

En la Figura 1.22 se relacionan dependencias de los factores de la expresión que calcula la potencia consumida con la organización (microarquitectura), diseño de circuitos y proceso tecnológico.

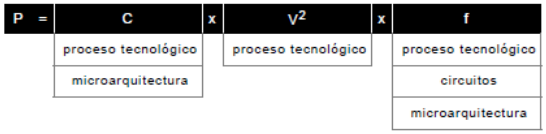


Figura 1.22 *Relación de dependencia de los factores implicados en el cálculo de la potencia.*

TEORÍA DEL ESCALADO

En una transición entre generaciones tecnológicas la teoría de escalado ideal indica que las dimensiones de los dispositivos (transistores) y cables (interconexiones) se escalan por un factor de ~0.7 (30% de reducción).

TENDENCIAS EN LA TECNOLOGÍA DE FABRICACIÓN

FRECUENCIA

La superación de la previsión de la teoría del escalado (1.43) se debe a: a) la reducción del número de puertas empleadas en cada ciclo de reloj con un posible incremento de la segmentación y b) la utilización de técnicas de diseño de circuitos mejoradas que reducen el retardo medio de puerta por encima del 30% por generación tecnológica previsto por la teoría de escalado.

POTENCIA

La potencia consumida por un chip depende de la tecnología de fabricación y del circuito que contiene y como se implementa. Esto es, entre otras cosas depende del tamaño del chip, del estilo de diseño de los circuitos lógicos, de la microarquitectura y de la frecuencia de operación.

TAMAÑO DEL DADO

Los diseñadores, además de aprovechar el incremento de densidad de los transistores, también incrementan el tamaño del dado.

MÉTRICAS QUE RELACIONAN RENDIMIENTO Y POTENCIA

La potencia consumida tiene dos costes. Uno de ellos es la disipación térmica y el otro el suministro de la potencia, mediante una tensión de alimentación y una intensidad de corriente.

Decir que un procesador consume menos que otro puede dar lugar a interpretaciones erróneas. En otras palabras, aun tardando más tiempo en realizar un cálculo, la energía consumida puede ser la misma. Por tanto, la batería se descarga en la misma cantidad de energía en ambos casos. En estas condiciones es mejor una figura de mérito que relacione la energía y las operaciones efectuadas.

MÉTRICA DE EFICIENCIA

Una de ellas es MIPS / Potencia = MIPS / W = Instruccion / J

PDP (power-delay product) Potencia / MIPS = J / instruccion

EDP (energy-delay product) Potencia / MIPS^2 = PDP * t

CONSUMO REDUCIDO

El objetivo es maximizar el tiempo de vida de una batería ya que la carga acumulada es fija.

E = T * P = P / R

E -> Energia

T -> Tiempo de ejecución

P -> Potencia

R -> Rendimiento

Entonces, para mejorar el tiempo de vida de una batería, el beneficio de rendimiento debe ser mayor que la potencia adicional consumida.