Evolución histórica de los computadores

En los sólo 50 años de vida de los computadores, los avances en su arquitectura y en la tecnología usada para implementarlos han permitido conseguir una evolución en su rendimiento sin precedentes en ningún otro campo de la ingeniería. Dentro de este progreso la tecnología ha mantenido un ritmo de crecimiento constante, mientras que la contribución de la arquitectura ha sido más variable.

En los primeros años de los computadores (desde el 45 hasta el 70) la mejora provenía tanto de los avances tecnológicos como de innovaciones en el diseño. En una segunda etapa (aproximadamente de los 70 a mediados de los 80) el desarrollo de los computadores se debió principalmente al progreso en la tecnología de semiconductores, que obtuvo mejoras impresionantes en densidad, velocidad y disipación de potencia. Gracias a estos avances el número de transistores y la frecuencia de reloj se incrementaron en un orden de magnitud en la década de los 70 y en otro en la de los 80.

Posteriormente tanto la tecnología como la arquitectura tuvieron una influencia fundamental en dicha evolución, cuyo ritmo se ha acelerado actualmente. En la década de los 90 el número de transistores y la frecuencia de reloj se han multiplicado por 20.

Muchos anuncian que este proceso comenzará a hacerse más lento a medida que nos aproximemos a los límites físicos de la tecnología de semiconductores. Según F. Faggin [Fagg96a], a partir de entonces las innovaciones en la arquitectura de los procesadores serán el motor fundamental de su progreso.

Para estudiar este proceso con mayor detalle usaremos una clasificación de los computadores en generaciones. Estas se dividen habitualmente basándose en la tecnología empleada, aunque los límites entre una y otra son más bien difusos. Cada nueva generación se caracteriza por una mayor velocidad, mayor capacidad de memoria, menor consumo y menor tamaño que la generación anterior. Existen algunas diferencias a la hora de realizar la clasificación en generaciones tecnológicas pero en general podemos decir que la Tabla 1 presenta la clasificación más ampliamente aceptada. En ella se destacan los principales avances tecnológicos y arquitectónicos que tienen lugar en cada una de las etapas.

Generación	Fechas	Característica tecnológica básica	Otros avances tecnológicos	Avances arquitectura
Primera	1946-1957	Válvula de vacío	Memoria ferritas Cintas magnéticas Disco magnético	Modelo Von Neumann
Segunda	1958-1963	Transistor	Incremento capacidad memorias	Memoria virtual Interrupciones Segmentación
Tercera	1964-1971	Circuito integrado SSI-MSI	Disco Winchester	Microprogramación memoria cache
Cuarta	1972-1980	LSI Microprocesadores	Memorias de semiconductores	
Quinta	1981-	VLSI	Incremento capacidad memorias y discos	Computadores RISC Superescalares Supersegmentados

Tabla 1 Generaciones de computadores

ANTECEDENTES: LAS MÁQUINAS DE CALCULAR MECÁNICAS

Los computadores surgen por la necesidad de las personas de realizar cálculos, que llevó a la invención de distintas máquinas para calcular. Ya en el 3.000 a. C. se inventó el ábaco, que puede considerarse el primer antecedente. Pero el primer paso importante en la historia de computadores lo constituyen las primeras máquinas de calcular mecánicas, que se remontan al siglo XVII, construidas por B. Pascal (1642-43) y G. Leibnitz (1674).

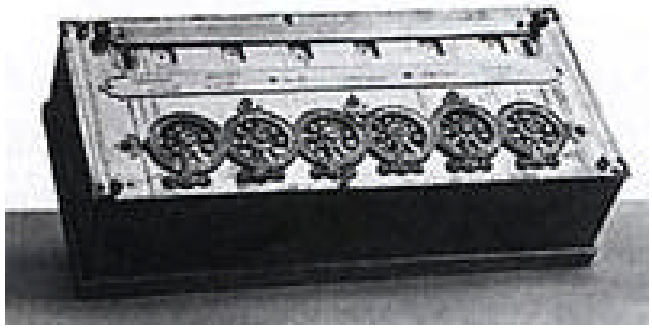

Figura 1: Máquina de Pascal

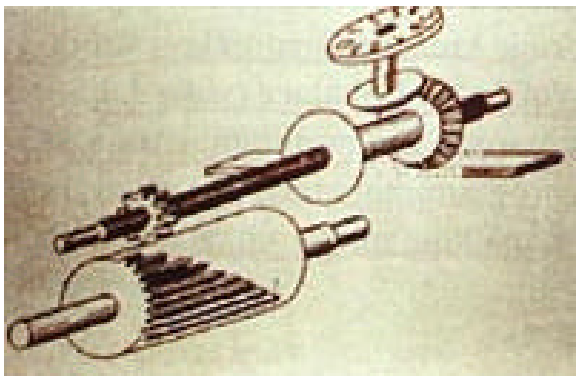

Figura 2: Máquina de Leibnitz

Posteriormente, en el siglo XIX, C. Babbage ideó dos máquinas: diferencial (1822-32) y analítica (1834-35). La primera, que únicamente realizaba un algoritmo, tenía una estructura muy simple. Su principal novedad la constituyó la salida de resultados por medio de perforaciones sobre una placa. La máquina

analítica, por su parte, constituye la primera máquina de propósito general. Era capaz de realizar cualquier operación matemática automáticamente. Tenía una unidad de almacenamiento, una unidad de procesamiento, una unidad de entrada de datos y otra de salida, y en cierto modo, su estructura se mantiene aún en los computadores modernos. No llegó a implementarse por dificultades tecnológicas, pero los computadores electromecánicos Harvard Mark I y Mark II, basados en relés, fueron diseñados por H. Aiken en 1944 y 1947 respectivamente, a partir de estas ideas de Babbage [Coel02].

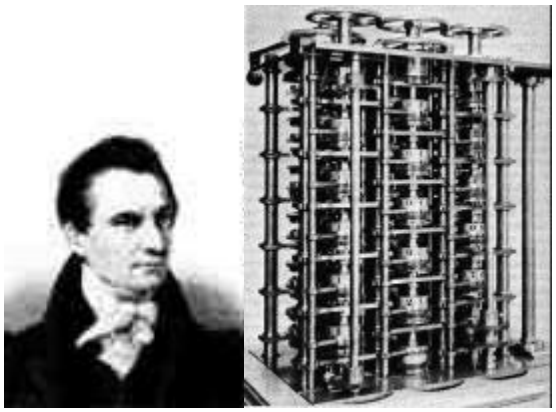

Figura 3: Charles Babbage y su máquina diferencial

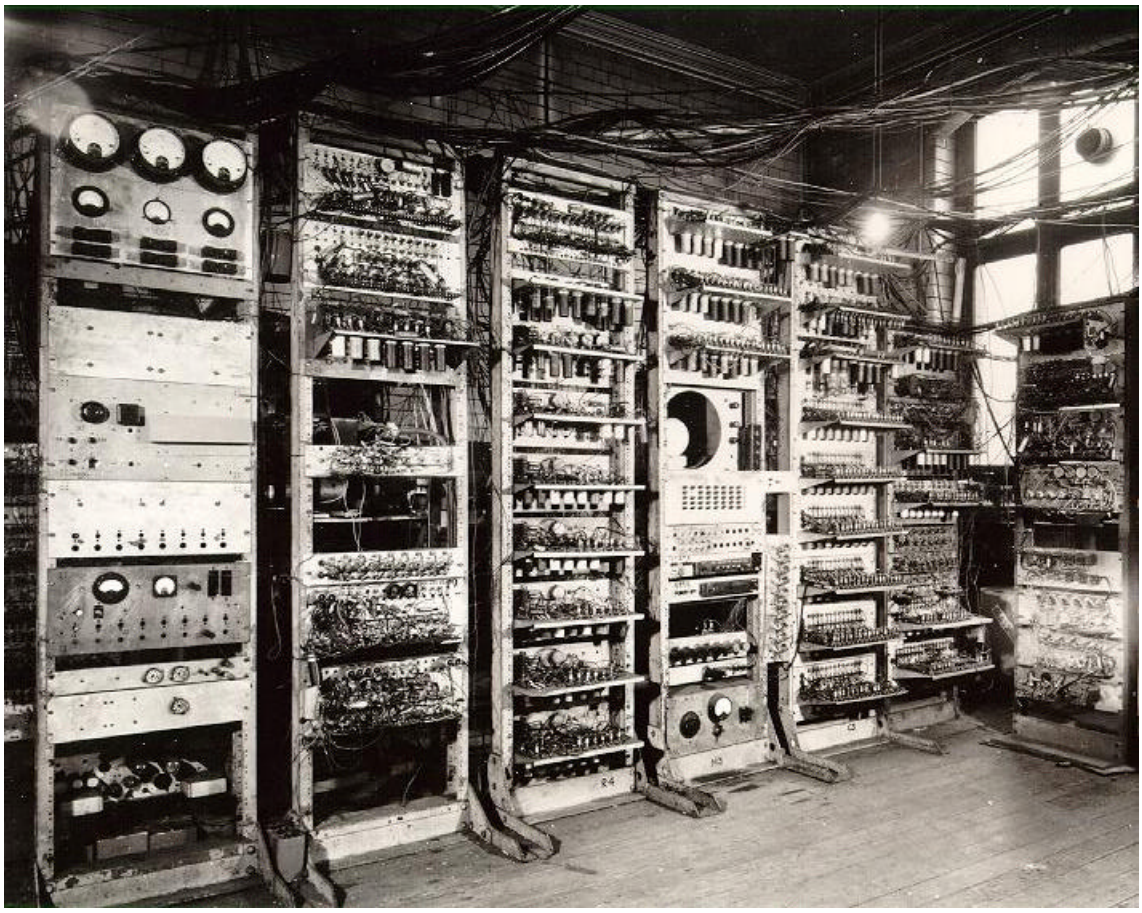

Figura 4: Imagen de la Mark I basada en las ideas de Babbage

Cabe destacar también el papel de Ada Lovelace, (hija de Lord Byron), en el desarrollo histórico de las computadoras modernas que fue casi totalmente ignorado hasta 1979, cuando el Departamento de Defensa de los Estados Unidos decidió utilizar su nombre para el nuevo lenguaje de programación que utilizarían como estándar para desarrollar su propio software interno. Desde entonces, parece que una nueva luz se ha producido sobre la vida de esta matemática tan talentosa, que fue una de las pocas personas que conoció y comprendió detalladamente el trabajo de Charles Babbage, además de haber escrito el primer programa para la inexistente Máquina Analítica.

PRIMERA GENERACIÓN: LAS VÁLVULAS DE VACÍO (1946-1957)

Tecnología básica

En 1904, Fleming patenta la válvula de vacío diodo, con idea de utilizarla para mejorar las comunicaciones de radio. En 1906, Forest añade un tercer electrodo al flujo de corriente de control del diodo de Fleming, para crear la válvula de vacío de tres electrodos.

Los computadores mecánicos tenían grandes dificultades para conseguir aumentar su velocidad de cálculo, debido a la inercia de los elementos móviles. Por ello el uso de válvulas de vacío supuso un gran paso adelante en el desarrollo de los computadores, tanto en velocidad como en fiabilidad, y dio lugar a lo que se conoce como Primera Generación de computadores.

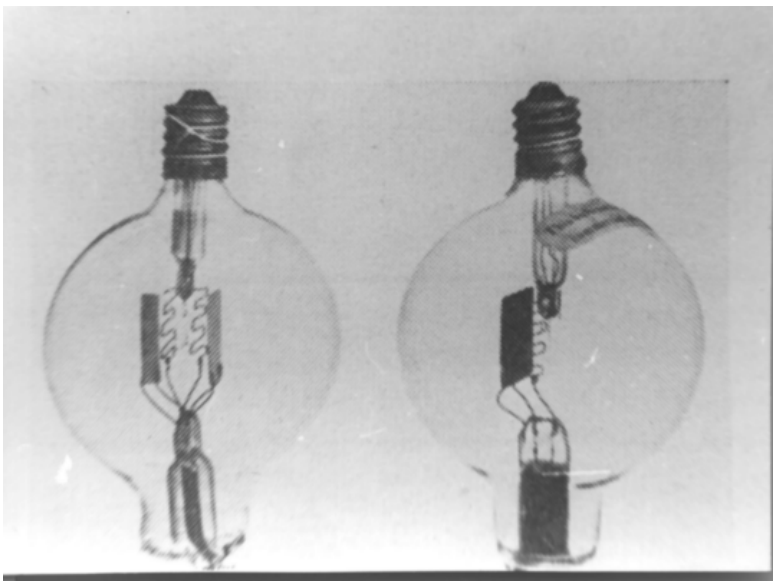

Figura 5: Imagen de las primeras válvulas de vacío

Avances en arquitectura

Habitualmente se considera que los computadores comenzaron con el ENIAC en 1946 y, de acuerdo con esto, la IEEE Computer Society celebró en

1996 los primeros 50 años de los computadores modernos. Sin embargo, J. V. Atanasoff había construido en 1939 un prototipo de computador digital electrónico que usaba aritmética binaria. Por eso desde 1973 se le reconoce como creador del computador moderno [CaBM96].

Si la bomba atómica fue el secreto mejor guardado por los norteamericanos durante la Segunda Guerra Mundial, su equivalente en Inglaterra fue el Colossus, la primera computadora completamente electrónica del mundo que se diseñó explícitamente para poder descifrar los mensajes secretos de los nazis y que A. Turing, T. Flowers y M.H.A. Newman presentaron en Diciembre de 1943 e hicieron operacional en Bletchley Park [Dani96]. . Esto marcó el inicio de la escuela inglesa de cómputo electrónico que le dio al mundo la primera computadora con programa almacenado de la historia, la primera unidad de control microprogramada y muchas otras valiosas contribuciones a la computación moderna.

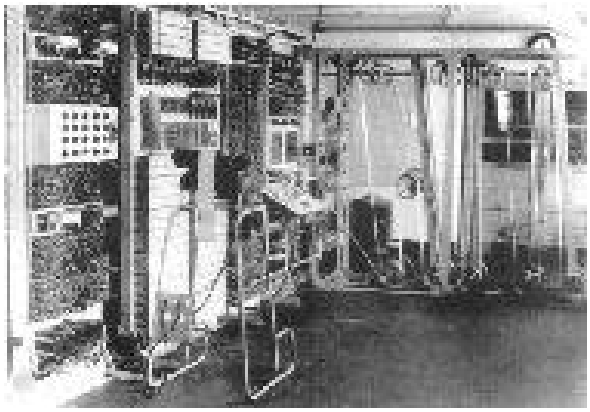

Figura 6: El Colossus, primera computadora totalmente electrónica

Pero es en 1946 cuando se considera que comienza la historia de los computadores. En la Universidad de Pennsylvania, J.P. Eckert y J.W. Mauchly mostraron al mundo el primer computador electrónico de propósito general: el ENIAC (Electronic Numerical Integrator and Calculator). Pesaba 30 toneladas y consumía 150 KW. Estaba construido con 18.000 válvulas de vacío y permitía realizar cálculos a una velocidad 100 veces mayor que una persona. Era programable y la programación se efectuaba mediante cables y conmutadores.

Los datos se introducían mediante tarjetas perforadas. Sus principales inconvenientes eran la tediosa tarea de programación, y la limitada capacidad de almacenamiento.

Para eliminar la necesidad de programar manualmente el computador J. Von Neumann propone un computador de programa almacenado denominado EDVAC (Electronic Discrete Variable Automatic Computer). Su diseño se denomina hoy "modelo Von Neumann", y se sigue manteniendo en la mayoría de computadores actuales, con unidad aritmético-lógica, unidad de control, unidades de entrada/salida, y memoria.

Basado en las ideas del EDVAC, M. Wilkes, de la Universidad de Cambridge, construye en 1949 el EDSAC (Electronic Delay Storage Automatic Calculator), que utiliza la noción de memoria jerárquica y una arquitectura basada en acumulador.

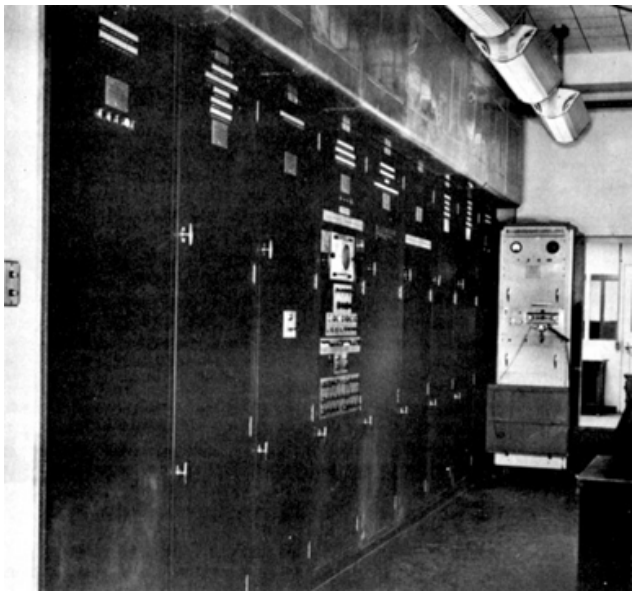

Figura 7: EDVAC

También von Neumann junto con H. Goldstine y A. Burks comenzó a construir un nuevo computador de programa almacenado, denominado IAS (Institute for Advanced Study) cuyo diseño no terminó hasta 1952. El IAS constaba de una memoria principal para almacenar datos e instrucciones, una unidad aritmético-lógica, una unidad de control que interpreta las instrucciones

y provoca su ejecución, y una unidad de entrada/salida dirigida por la unidad de control.

En 1951 Wilkes introduce la idea de la microprogramación para el diseño ordenado de la unidad de control. Esta idea no fue realizable ya que para almacenar los microprogramas se requería una memoria muy rápida, que no estaba disponible en ese momento. Por esta razón, la idea quedó como una mera conjetura académica durante una década. Una vez más, como le sucedió a Babbage, una innovación arquitectónica tuvo que esperar hasta que la tecnología avanzara para permitir su implementación[Wilk51][Wilk53] .

Otras tecnologías

Al mismo tiempo, en el MIT el equipo de J. Forrester trabaja en un computador de propósito especial para tratamiento de señales en tiempo real, el proyecto Whirwind (1949), cuya principal aportación es la utilización de la memoria de ferritas. Esta ha sido la principal tecnología de memoria durante varias décadas [deMi90]. Cada punto de memoria es un toro o anillo de ferrita, que puede presentar dos direcciones de magnetización. Las primeras ferritas fabricadas tenían un diámetro exterior de 3 mm, tenían una capacidad de 2 Kbytes y el tiempo de acceso era de unos 30ms. La conexión de los anillos de ferrita a los transductores se hacía mediante hilos de cobre barnizados que debían hacerse pasar por el interior de las ferritas. Este proceso era de difícil automatización, por lo que debía hacerse a mano.

Por otra parte se extiende el uso de cintas magnéticas para el almacenamiento masivo. También aparece el primer disco magnético en el año 1956, que se utilizó en la máquina RAMAC (Random Access Method of Accounting and Control) de IBM, con una capacidad de 5Mbytes y un tiempo de aproximación de 1s [HePa02].

Figura 8: Usuario utilizando una de las primeras versiones del RAMAC de IBM

Los primeros computadores comerciales de esta generación, que aparecieron en la década de los 50, fueron el UNIVAC I y II (Universal Automatic Computer), fabricados por Eckert y Mauchly y la serie 700 de IBM.

En esta primera generación de computadores, las instrucciones se procesaban en serie: se buscaba la instrucción, se decodificaba y luego se ejecutaba. La velocidad típica de procesamiento que alcanzaron los computadores era aproximadamente 40.000 operaciones por segundo. Eran equipos de gran tamaño, escasa capacidad y difícil mantenimiento, que disipaban mucho calor. Los trabajos se realizaban en monoprogramación y no existía sistema operativo, por lo que los periféricos de entrada/salida dependían directamente del procesador. Se programaba en lenguaje máquina, lo que exigía programadores muy especializados.

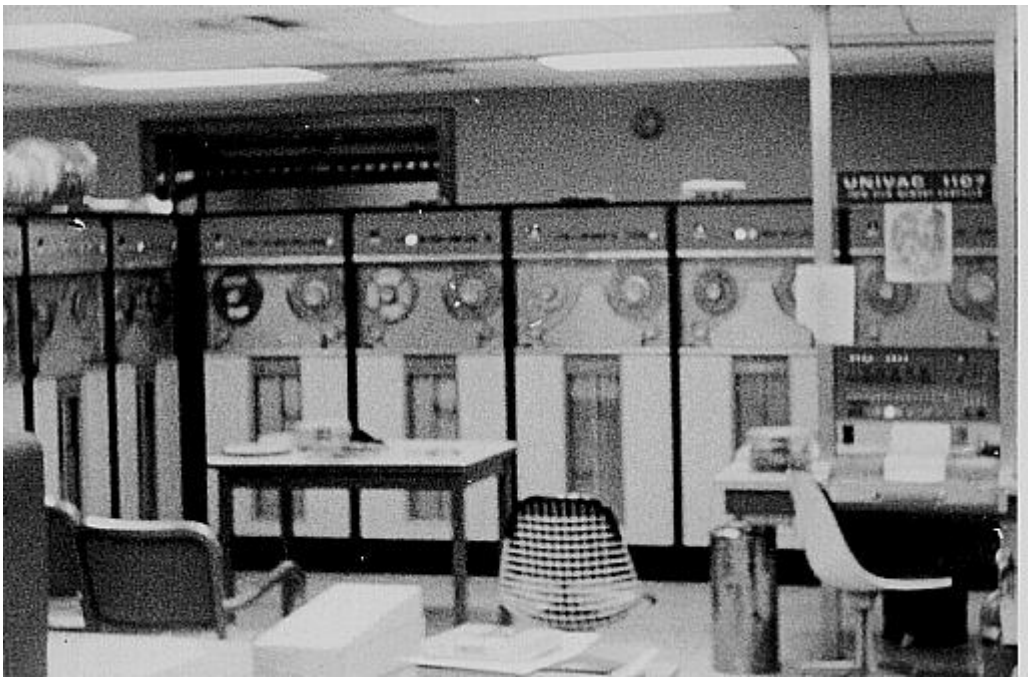

Figura 9: Imagen del UNIVAC II

SEGUNDA GENERACIÓN: LOS TRANSISTORES (1958-1963)

Tecnología

La invención del transistor tuvo lugar en 1948 en los laboratorios Bell por W.B. Shockley, J. Bardeen y W.H. Brattain. Poco a poco la industria de semiconductores fue creciendo y los productos industriales y comerciales sustituían los dispositivos de válvulas de vacío por implementaciones basadas en semiconductores.

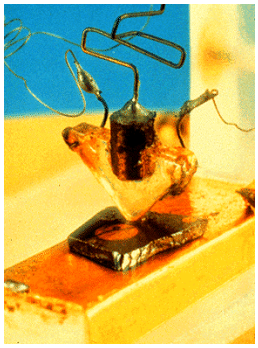

Figura 10: El transistor

La nueva tecnología permite aumentar el rendimiento y la fiabilidad, y reducir de forma drástica el tamaño de los computadores, dando lugar a la Segunda Generación de computadores. La velocidad de ejecución de la CPU se incrementó enormemente, hasta alcanzar 200.000 operaciones por segundo. La disminución de tamaño de los módulos permitió introducir unidades lógicas y aritméticas y unidades de control más complejas.

Por otra parte, el tamaño de la memoria principal de ferritas creció de 2 Kpalabras a 32 Kpalabras, y el tiempo de aproximación cayó de 30 ms a 1,4 ms.

Avances en arquitectura

El incremento de la complejidad de las unidades de control, permitió introducir una de las innovaciones arquitectónicas que posteriormente se ha utilizado en gran escala: la segmentación de operaciones. Con esta técnica, la decodificación de una instrucción se solapa con la búsqueda de la instrucción siguiente y con la ejecución de la anterior.

Figura 11: IBM 7030

En 1961 aparece el IBM 7030 o Stretch, el primer computador que usa segmentación. También tiene memoria entrelazada y predicción de saltos. No tuvo éxito comercial debido, entre otras causas, a que no alcanzó el rendimiento esperado porque el tiempo para recuperarse de un error de predicción era muy largo.

Entre las innovaciones arquitectónicas más importantes de esta generación puede destacarse la utilización de memoria virtual, para facilitar la tarea del programador a la hora de escribir programas demasiado largos para residir completamente en memoria principal. Estos programas debían estar formados por varios segmentos que se cargaban alternativamente desde la memoria secundaria hasta la memoria principal, bajo la supervisión del programa principal. La memoria virtual intentaba aliviar a los programadores de este peso, gestionando automáticamente los dos niveles de la jerarquía de memoria, formada por la memoria principal y la secundaria [HePa02]. La memoria virtual,

y la utilización de interrupciones para la E/S se utilizaron por primera vez en el sistema ATLAS (1962), desarrollado por Ferranti en la Universidad de Manchester, que también usaba segmentación.

El CDC 6600 diseñado por S. Cray de Control Data Corp. en 1964 fue el primer supercomputador comercial de éxito. Tenía arquitectura de carga-almacenamiento y empleaba técnicas de segmentación, además de paralelismo a nivel de unidades funcionales, lo cual le permitía un rendimiento de 9 MFLOPs, superior en un orden de magnitud al del 7094 de IBM.

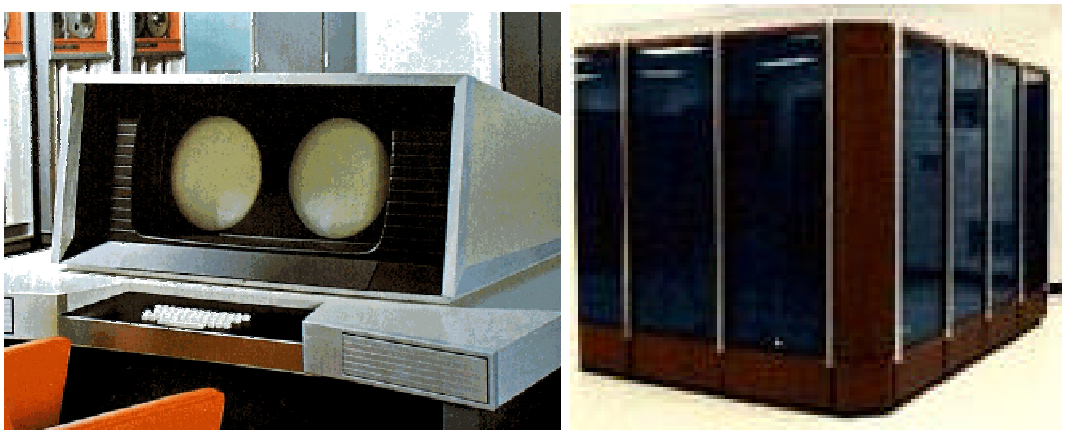

Figura 12:CDC 6600 y CDC 7600 diseñados por Seymour Cray

Además de la serie 7000 de IBM, otra máquina comercial de esta generación fue el PDP-1, lanzado en 1960 por DEC. Una de las innovaciones interesantes del PDP-1 fue el empleo de un terminal de vídeo, así como cierto grado de capacidades gráficas sobre la pantalla de 512x512 pixels.

Con estos computadores comenzó la utilización de lenguajes de programación de alto nivel [Stal97], como FORTRAN -cuyo primer compilador desarrolló IBM en 1957-, COBOL y LISP (ambos en 1959), y apareció el procesamiento por lotes, que será el germen de los sistemas operativos.

TERCERA GENERACIÓN: LOS CIRCUITOS INTEGRADOS (1964-1971)

Tecnología básica

Durante la generación anterior los equipos electrónicos estaban compuestos en su mayoría por componentes discretos -transistores, resistencias, condensadores, etc.- cada uno de los cuales se fabricaba separadamente y se soldaban o cableaban juntos en tarjetas de circuitos. Todo el proceso de fabricación resultaba caro y difícil, especialmente para la industria de computadores, que necesitaba colocar juntos cientos de miles de transistores que había que soldar, lo cual dificultaba enormemente la fabricación de máquinas nuevas y potentes.

Por eso, la invención del circuito integrado a finales de los 50 (J. Kilby de Texas Instruments construye el primero en 1958 y R. Noyce de Fairchild Semiconductor construye otro en 1959) fue la gran clave para el crecimiento de la industria de computadores, y suele tomarse como punto de inicio de la Tercera Generación de computadores.

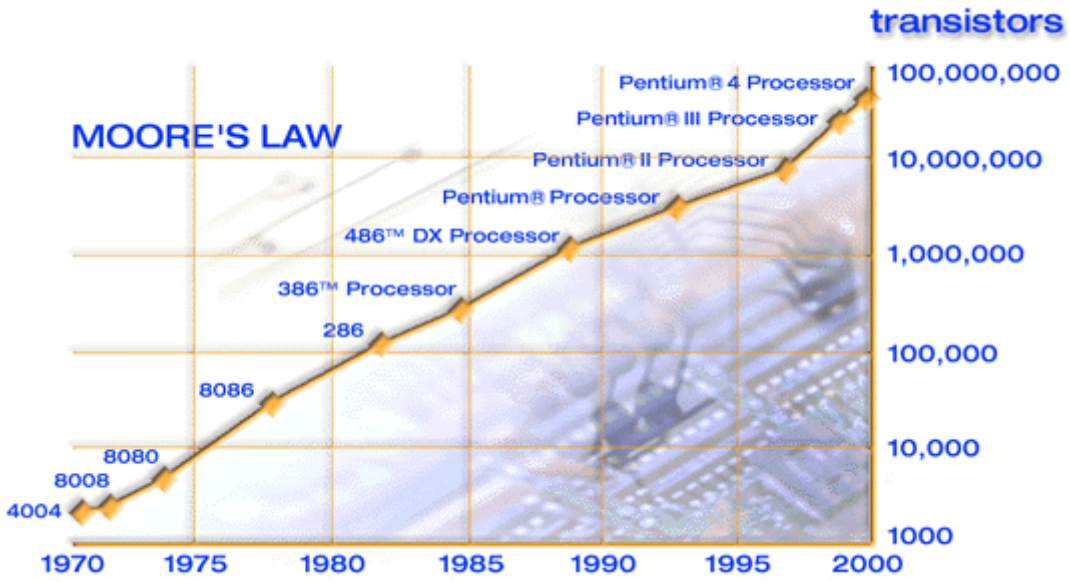

Figura 13: Ley de Moore según Intel

La introducción de circuitos integrados comerciales empezó en 1961 con componentes RTL (resistor-transistor logic), que fueron pronto sustituidos por componentes TTL (transistor-transistor logic). Posteriormente pequeños grupos de dispositivos de tecnologías TTL SSI (Small Scale Integration) fueron reemplazados por dispositivos de tecnologías TTL MSI (Medium Scale Integration) y LSI (Large Scale Integration). Entre 1961 y 1971 los chips se fueron haciendo mayores y los transistores cada vez más pequeños, de modo que el número de transistores en un chip casi se duplicaba anualmente –esta predicción se ha denominado posteriormente "ley de Moore" (ver Figura 13: Ley de Moore según Intel). Así las funciones lógicas que podían realizar los circuitos también se habían complicado considerablemente. De esta forma era posible realizar módulos y unidades de control aún más complejas, sin que el precio de los circuitos se incrementase, y el tamaño de los computadores se redujo considerablemente, a la vez que aumentó su velocidad y disminuyó el consumo.

Los computadores B2500 y B3500 de Burroughs usaron circuitos integrados y fueron construidos en 1968. Los CIs también disminuyeron el coste de los controladores de discos y de la electrónica para controlar los brazos, de forma que se podían incluir dentro de la caja del disco y ésta se podía sellar. Así surgieron los primeros discos que no eran extraibles: en 1965 aparece el disco Winchester. Después, en 1970 aparecen los discos flexibles (floppy).

Avances en arquitectura

En este periodo también tienen lugar importantes innovaciones arquitectónicas. La principal de ellas es la microprogramación, es decir, Describir las operaciones involucradas en la ejecución de una instrucción máquina mediante un conjunto de bits, que representan a las distintas señales de control que es necesario activar. Dicho conjunto de bits o microinstrucción se almacena en un memoria, denominada memoria de control. Este enfoque

había sido propuesto por Wilkes a principios de los años 50, pero la tecnología de memorias disponible no permitió su materialización.

Es realmente en 1964 cuando IBM introduce la microprogramación en la industria de los computadores al lanzar la familia IBM System/360 [AmBB64], en la que todos los modelos, excepto los más avanzados, eran microprogramados. El concepto de familia de computadores, todos con la misma arquitectura pero con distintos precios y prestaciones, contribuyó a que el IBM/360 fuera el mainframe más popular en los 70. Su éxito fue tan grande que los mainframes actuales de IBM todavía son compatibles con él. Entre otras prestaciones, merecen citarse la posibilidad de programar la prioridad de las interrupciones, los mecanismos de protección de memoria y la inclusión de controladores de DMA.

Figura 14: Imagen del IBM 360

Por otra parte, Wilkes propone en 1965 la memoria cache: se trata de añadir un nivel de memoria intermedio entre el procesador y la memoria principal, con una capacidad inferior a la memoria principal pero un tiempo de aproximación mucho menor. La primera implementación comercial la llevó a cabo IBM en su modelo 360/85 en el año 1968, y pronto se hizo común en las grandes máquinas y minicomputadores. Actualmente es uno de los métodos más utilizados para mejorar el rendimiento debido a la creciente diferencia

entre la velocidad del procesador y la de la memoria. El problema de diseño de caches es uno de los compromisos dirigidos por la tecnología, porque los valores óptimos de los tres principales parámetros de diseño de las caches (tamaño, asociatividad y tamaño de bloque) están muy interrelacionados entre sí y dependen de los detalles de cada diseño en particular. Como el tiempo de aproximación a cache está casi siempre en el camino crítico del diseño de un procesador, el tiempo necesario para saber si un dato está en cache suele influir en el tiempo de ciclo del computador y este tiempo suele ser dependiente del tamaño de cache y del grado de asociatividad.

Al nivel de los minicomputadores también se produjo un paso importante, con la presentación en 1965 del PDP-8 de DEC. Cuando la mayoría de los computadores requerían una habitación con aire acondicionado, el PDP-8 podía colocarse encima de una mesa de laboratorio. Los últimos modelos del PDP-8 usan por primera vez estructura de bus.

Figura 15: El NEC-PDP 8 se podía situar en una mesa de laboratorio.

También hubo avances importantes en el campo de los sistemas operativos. IBM crea el OS/360, primer sistema operativo multiprogramado. Además, aparecen el sistema operativo Multics (1965) y después D. Ritchie y K. Thomson crean el Unix (1970) en los laboratorios Bell. Con esta generación de computadores se consiguieron velocidades de procesamiento de 1 millón de instrucciones por segundo (1 MIPS).

CUARTA GENERACIÓN: LOS MICROPROCESADORES (1971-1980)

Tecnología básica

En 1970 tanto la industria de computadores como la de semiconductores habían madurado y prosperado y su unión permitió el desarrollo de la denominada Cuarta Generación de computadores: basados en microprocesador. Esta etapa viene caracterizada nuevamente por un avance tecnológico, como es el desarrollo de la técnica de integración LSI, que permite incluir hasta 100.000 transistores en un único chip. En 1973 se consiguen integrar 10.000 componentes en un chip de 1cm2.

El primer microprocesador, el 4004 de Intel [Fagg96b], surge en 1971 ideado por T. Hoff y construido por F. Faggin. Era un procesador de 4 bits con 2300 transistores en tecnología de 8 micras. Fue fabricado en obleas de 2 pulgadas y empaquetado con 16 pines. Podía direccionar 8 Kbytes de ROM y 640 bytes de RAM. Un año después apareció el 8008, un procesador de 8 bits con 3500 transistores, que podía direccionar 16 Kbytes de memoria y trabajar a 0.5 MHz [Tred96].

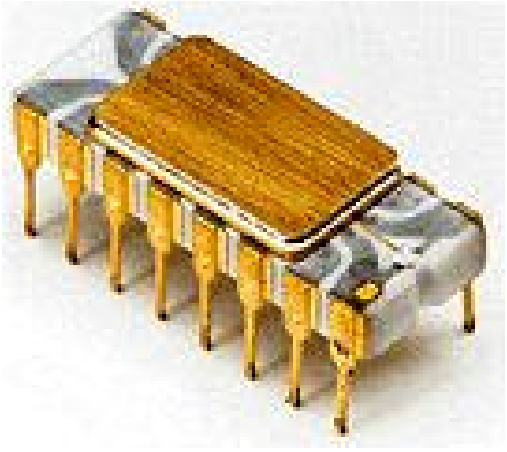

Figura 16: El primer microprocesador, el 4004 de Intel.

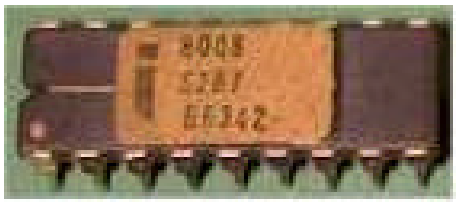

Figura 17: El 8008 contenía ya 3500 transistores en 1972

La primera aplicación del 4004 fue una calculadora de escritorio [Fagg96a]. Sin embargo, dos años después el 4004 se utilizaba en una gran variedad de sistemas empotrados como ascensores, etc. A partir de ese momento cada dos o tres años aparecía una nueva generación de microprocesadores, y los diseñadores los usaban para cualquier producto que pudiera beneficiarse de alguna cantidad de inteligencia, desde juguetes a calculadoras de bolsillo y a computadores personales. Durante los últimos 25 años, a una velocidad impresionante, el microprocesador ha cambiado la estructura de muchas de las industrias existentes e incluso ha empezado a cambiar también el aspecto de la sociedad.

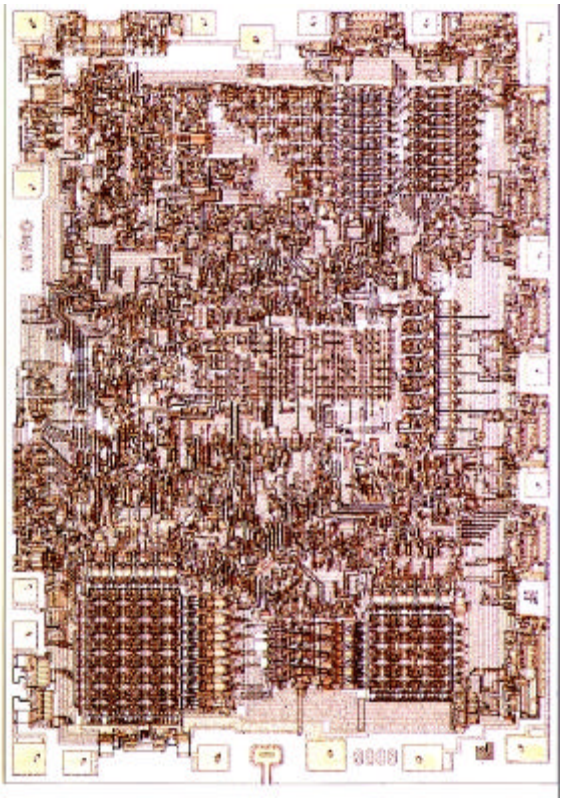

Figura 18: Layout del 8008

Otras tecnologías

Otro de los factores tecnológicos que permiten este abaratamiento de los computadores es la introducción de las memorias de semiconductores. Las memorias de ferritas se caracterizaban principalmente por ser voluminosas, caras y de lectura destructiva. Por eso, un gran avance fue la aplicación de la tecnología de CIs a la construcción de memorias. En el año 1970 Fairchild produjo la primera memoria de semiconductores de una capacidad apreciable. Este primer chip era del mismo tamaño que un único núcleo de ferrita y podía contener 256 bits de memoria. Además presentaba un tiempo de aproximación mucho menor que el de la memoria de ferritas. Sin embargo, su coste por bit era mayor que el del núcleo de ferrita.

En el año 1974, el coste por bit de la memoria de semiconductores cayó por debajo del coste de la memoria de ferritas. Ese año se construyó un chip de DRAM de 4 Kbits. Desde entonces, la capacidad de almacenamiento de las memorias no ha dejado de incrementarse año tras año. Este crecimiento ha ido acompañado por una disminución de su coste y un incremento en la velocidad de aproximación.

Figura 19: MITS Altair 8800

Avances en arquitectura

La arquitectura de los primeros microprocesadores [Fagg96a] fue una adaptación de las ideas usadas con anterioridad en los minicomputadores y los mainframes. Las compañías incorporaban estas ideas en los microprocesadores tan pronto como la rápida evolución de las capacidades dadas por la tecnología y el coste lo permitían. Por eso esta generación se caracteriza principalmente por las mejoras en la tecnología de circuitos integrados, que los microprocesadores aprovechan más que otros computadores debido a su mayor integración, y no tanto por las mejoras arquitectónicas.

Ya en 1974 el 6800 contenía alrededor de 5000 transistores en tecnología NMOS de 6 micras. Operaba a 2 MHz y podía direccionar 64 Kbytes de memoria. También aparecieron el MC6502 y el Intel 8080 entre otros.

La disminución del coste de los CIs conduce a un gran abaratamiento de los computadores, lo cual permite la fabricación de los primeros computadores personales. En 1975 apareció el primer sistema de computador popular basado en microprocesador: el MITS Altair 8800. Estaba basado en el Intel 8080, un microprocesador de 8 bits que trabaja a 2 MHz introducido en 1974. El Apple II se introdujo en 1977, basado en el MC6502, junto con un terminal CRT, un teclado y una disquetera. Fue el primer computador personal con gráficos en color.

En 1978 Intel introdujo el microprocesador de 16 bits 8086, con 29000 transistores, tecnología HMOS de 3 micras, un rango de direcciones de 1 Mbyte y una velocidad de 8MHz. Este diseño fue utilizado por IBM para el computador personal (IBM PC) que se presentó en 1981, para el que elige el PC-DOS de Microsoft como sistema operativo.

Figura 20: El Apple II

En 1979, pensando que la memoria seguiría reduciendo su coste y que los futuros programas se escribirían en lenguajes de alto nivel, Motorola incrementó el ancho de banda con un bus de datos de 16 bits y un bus de direcciones de 32 bits para el MC68000. Este microprocesador podía dar un rendimiento pico de 2 MIPS. Debido a limitaciones de empaquetamiento (tenía 64 pines) los 32 bits se redujeron a 24 en los primeros productos. También se añadieron registros de propósito general de 32 bits, hasta un total de 16. El número total de transistores era de alrededor de 68000, podía trabajar a 5V y a una velocidad de 8 MHz. Apple Computer seleccionó el 68000 para la línea de computadores personales Macintosh.

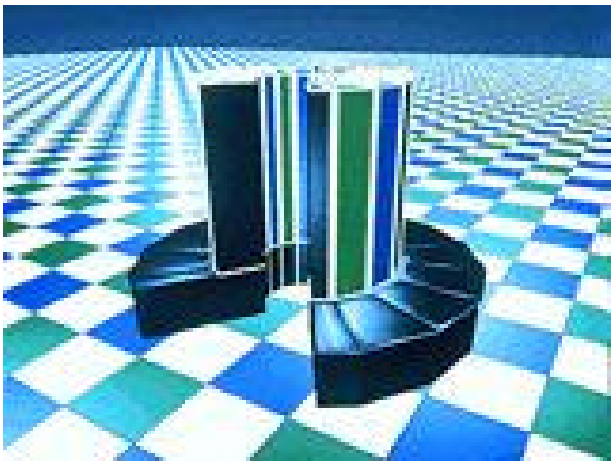

Figura 21: Imagen publicitaria del Cray-1

En 1976 aparece el Cray-1 de Cray Research, el primer supercomputador vectorial. En 1978 DEC presenta el VAX 11/780, un computador de 32 bits que se hace popular para aplicaciones científicas y técnicas. Los diseñadores del VAX buscan simplificar la compilación de lenguajes de alto nivel –en 1972 habían aparecido C, SmallTalk y Prolog, que se sumaron a los ya existentes Fortran, Cobol y Lisp- y para ello crean una arquitectura ortogonal de instrucciones complejas. Además optimizan el tamaño del código para que ocupe menos memoria.

En el sentido opuesto al de los diseñadores del VAX, J. Cocke trabaja en el proyecto 801 de IBM para desarrollar un minicomputador que será origen de las futuras arquitecturas RISC de la siguiente generación.

Figura 22: Imagen del VAX 11/780

QUINTA GENERACIÓN: DISEÑO VLSI 1981-?

Año tras año el precio de los computadores disminuye forma drástica, mientras las prestaciones y la capacidad de estos sistemas siguen creciendo. El incremento de la densidad de integración ha permitido pasar de circuitos con unos pocos miles de transistores a principios de los años 70 a varios millones en la actualidad. Por ello podemos afirmar que la aparición de la tecnología VLSI a principios de los 80 puede considerarse como el origen de la Quinta Generación, que se caracteriza fundamentalmente por la proliferación de sistemas basados en microprocesadores[BuGo97] [Gei90].

Tecnología básica

Los tamaños mínimos de fabricación (minimum feature size, λ) han disminuido desde las 50 micras de los años 60 a las 0.13 micras actuales, mejorándose de este modo tanto la densidad de integración (con un crecimiento anual aproximado del 35%) como la velocidad de los circuitos integrados. En paralelo con esta reducción, las dimensiones máximas del chip también han aumentado, aunque en este caso la evolución es menos predecible, rondando entre el 10% y el 20% anual. El efecto combinado del aumento de la densidad de transistores y del tamaño de los chips ha dado lugar a un aumento en el número de transistores por chip en torno a un 55% anual [HePa02], consiguiéndose integrar en la actualidad del orden de 200 millones de transistores en un único chip. Por citar un ejemplo destacable, el Power4 de IBM integra unos 174 millones de transistores[1][Dief99].

No obstante, a consecuencia de dicha disminución han surgido nuevos desafíos. Al reducirse el tiempo de conmutación de los transistores cobra especial relevancia los retardos de propagación de las señales dentro del propio circuito integrado. Un ejemplo destacable lo tenemos en el diseño del Pentium 4, en el que dos de las 20 etapas de que consta su pipeline se utilizan exclusivamente para propagar las señales dentro del chip [HePa02].

[1] El chip incluye 2 procesadores

Otro aspecto importante a tener en cuenta es el consumo de potencia. En los primeros microprocesadores, el consumo era de tan solo unas decenas de Watios. Actualmente, la potencia máxima disipada por los microprocesadores de gama más alta puede alcanzar entre los 100 y los 150 Watios, siendo probable que en un futuro próximo sean factores relacionados con el consumo los que limiten o bien la cantidad de hardware que pueda ser incluido en el microprocesador o bien la frecuencia de reloj a la puedan trabajar[HePa02].

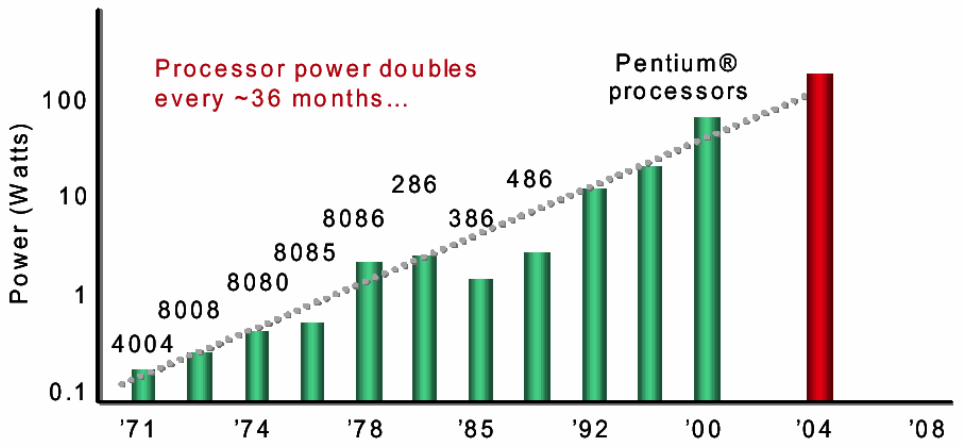

Figura 23: Evolución histórica del consumo de los microprocesadores desarrollados por Intel [MNWS02].

Del incremento en la densidad de integración también se ha beneficiado la tecnología de memorias dinámicas (DRAMs). La densidad (en número de bits por unidad de área) de los módulos DRAM ha aumentado entre un 40% y un 60% anual. Lamentablemente, el tiempo de ciclo se ha ido reduciendo muy lentamente, aproximadamente en un tercio cada diez años [HePa02].

Modelo	Tipo	Ciclo de Reloj Efectivo	Bus de Datos	Ancho de Banda Pico
PC66	SDRAM	66 MHz	64 Bit	0,5 GB/s
PC100	SDRAM	100 MHz	64 Bit	0,8 GB/s
PC133	SDRAM	133 MHz	64 Bit	1,06 GB/s
PC1600	DDR200 DDR200 Dual	100 MHz	64 Bit 2 x 64 Bit	1,6 GB/s 3,2 GB/s
PC2100	DDR266	133 MHz	64 Bit	2,1 GB/s

	DDR266 Dual	133 MHz	2 x 64 Bit	4,2 GB/s
PC2700	DDR333	166 MHz	64 Bit	2,7 GB/s
	DDR333 Dual	166 MHz	2 x 64 Bit	5,4 GB/s
PC3200	DDR400	200 MHz	64 Bit	3,2 GB/s
	DDR400 Dual	200 MHz	2 x 64 Bit	6,4 GB/s
PC4200	DDR533	266 MHz	64 Bit	4,2 GB/s
	DDR533 Dual	266 MHz	2 x 64 Bit	8,4 GB/s
PC800	RDRAM Dual	400 MHz	2 x 16 Bit	3,2 GB/s
			2 x 32 Bit	6,4 GB/s
PC1066	RDRAM Dual	533 MHz	2 x 16 Bit	4,2 GB/s
			2 x 32 Bit	8,4 GB/s
PC1200	RDRAM Dual	600 MHz	2 x 16 Bit	4,8 GB/s
			2 x 32 Bit	9,6 GB/s

Figura 24 Frecuencias de trabajo y los anchos de banda pico de los módulos de memoria comercializados en mayo de 2002[Inst02].

Respecto al ancho de banda proporcionado por cada chip de memoria, durante el mismo periodo de tiempo (diez años) se ha observado un incremento del orden de dos tercios, si bien se han logrado mejoras adicionales mediante el diseño de nuevas interfaces. Los interfaces basados en protocolos asíncronos como page mode, fast page mode o EDO RAM han sido desplazados, compitiendo actualmente por el mercado dos tecnologías con protocolo síncrono (SDRAM): Rambus DRAM (RDRAM) y double-data-rate (DDR) DRAM [Paul02]. En la tabla 1.2.1 se muestran las frecuencias de trabajo y los anchos de banda pico de los módulos de memoria que se comercializaban en mayo de 2002 [Inst02].

Otros avances tecnológicos

Paralelamente al incremento de la densidad de integración, ha aumentado notablemente la capacidad de los sistemas de almacenamiento, y ha disminuido su coste. Hasta 1990 la capacidad de almacenamiento de los discos se incrementaba en un 30% anual. Recientemente, la densidad de integración se ha aumentado en más del 100% anual, aunque como en el caso de las memorias, el tiempo de acceso sólo se reduce por un factor de un tercio cada diez años [HePa02].

Este desequilibrio supone un verdadero problema, sobre todo en los sistemas de memoria virtual. Una de las principales soluciones aportadas a este problema es la técnica RAID (Redundant Array of Independent/Inexpensive Disks) [CLGK94], que surgió por primera vez en 1987. El objetivo de esta técnica es aprovechar la reducción de tamaño y coste de los discos para aumentar la fiabilidad y el rendimiento de los sistemas de almacenamiento masivo. Consiste en utilizar varias unidades de discos que operen independientemente y en paralelo. De esta forma se puede acceder a un bloque de datos en paralelo siempre que los datos de dicho bloque estén adecuadamente distribuidos a lo largo de varios discos, consiguiendo una velocidad de transferencia mucho mayor. Desde el punto de vista del usuario o del sistema operativo este conjunto de discos físicos opera como una única unidad lógica. Además existen discos que almacenan información redundante que permite garantizar la correcta recuperación de los datos en caso de fallo.

También han surgido otras tecnologías de almacenamiento: la óptica y la magneto-óptica [Stal97]. El CD-ROM, introducido por Sony y Philips en 1984, es una memoria de sólo lectura basada en tecnología óptica, que se escribe mediante un rayo láser que realiza hoyos microscópicos sobre una superficie reflectante. Los discos ópticos borrables, de tecnología magneto-óptica, utilizan la energía de un rayo láser junto con un campo magnético para grabarse. Otro

disco óptico, el DVD (Digital Versatil Disk), tiene las mismas dimensiones que un CD-ROM pero puede almacenar hasta 17 Gbytes [Dutt99].

En general estas nuevas tecnologías permiten sistemas de almacenamiento de alta capacidad, seguras e intercambiables, pero su tiempo de acceso es mayor que el de los discos magnéticos, por lo que no suponen una alternativa a aquellos como almacenamiento secundario durante la ejecución de los programas.

Por último, debemos destacar también la mejora que ha experimentado durante estos últimos años la tecnología de red. Tanto la latencia como el ancho de banda han mejorando con el tiempo, si bien durante los últimos años la evolución se ha acelerado notablemente, habiéndose puesto especial énfasis en el ancho de banda. Por citar un ejemplo, hicieron falta unos 10 años para pasar de Ethernet de 10 Mb a 100 Mb, mientras que tan solo cinco años después estuvo disponible la tecnología Ethernet de 1Gb. Esta mejora es debida tanto a la generalización de los dispositivos ópticos como a la mayor densidad de integración de los conmutadores [HePa02]. Aunque hemos dicho que estamos en la quinta generación, dentro de ella podemos distinguir tres etapas en función de los los avances arquitectónicos que han tenido lugar. Veamos con un poco más de detalle cada una de estas etapas y los cambios que las delimitan.

Avances arquitectónicos: primera etapa

La primera etapa de esta generación dura hasta mediados de los 80, cuando el número de transistores en un CI se aproxima a 250.000. En este periodo no hay grandes novedades en la arquitectura de los computadores personales. Para mejorar el rendimiento la secuencia de instrucciones se segmenta a 5 ó más etapas.

Un ejemplo de esta generación es el MC68020, introducido en 1984, el primer microprocesador de 32 bits tanto en los buses de datos y direcciones

como en registros y ALU. Tenía 200.000 transistores, incorporaba por primera vez una cache interna de 256 bytes y una segmentación de hasta 5 etapas [Tred96]. Funcionaba a una frecuencia de 16 MHz.

Por su parte Intel incrementó el ancho de bits de su serie x86 a 32 bits con el Intel 80386 [CrGe87] [Cat90] [Brey95] (1985). Tenía 275.000 transistores y reloj de 33 MHz. Incluía la lógica de gestión de memoria en el chip. No utilizó segmentación.

En esa época los fabricantes de minicomputadores como Hewlett-Packard (HP), Digital Equipment Corp. (DEC), Tektronix y Apollo comienzan a usar microprocesadores para sus CPUs, creando el mercado de las estaciones de trabajo [Tred96], que pronto consiguieron ser mucho más potentes que los mainframes de las generaciones anteriores.

A principios de los años 80, John Hennessy, de la Universidad de Stanford, y David Patterson, de la Universidad de Berkeley, definen la base de los procesadores actuales. Estudios dinámicos sobre ejecución de las cargas de trabajo habituales revelaron que las instrucciones y los modos de direccionamiento complejos se usaban muy poco. Además, las instrucciones eran muy largas y eso aumentaba el tiempo necesario para leerlas de memoria, cada vez más crítico. De estos estudios surgió la idea de diseñar computadores de repertorio de instrucciones reducido o RISC (Reduced Instruction Set Computer), nombre acuñado por D. Patterson. Los repertorios simples facilitaron la labor de construir microprocesadores segmentados ya en 1980-81 [Kogg81][bhcl91]. Los primeros prototipos RISC eran segmentados y el primero en llegar al mercado fue el MIPS R2000 en 1986, seguido del Sparc [Cata91] de Sun, 29000 de AMD, etc... [Feel94] [Gimi87] [Henn96] [Henn99][ibm94] [Pase82] [Kate85].

En el campo de los supercomputadores se presenta en 1986 el Cray XP de cuatro procesadores, que alcanza 713 MFLOPs.

Figura 25: Cray XP

Avances arquitectónicos: segunda etapa

La segunda etapa comienza cuando se pueden integrar un millón de transistores en un CI, a finales de los 80. Los mayores fabricantes de microprocesadores para computadores personales, Intel y Motorola, tenían absoluta necesidad de compatibilidad, por lo que no modificaban significativamente sus arquitecturas. Así para obtener mayor rendimiento con una arquitectura fija se aumentaba la frecuencia de reloj y se usaban los transistores disponibles para añadir memoria cache interna, coprocesadores matemáticos, segmentaciones más profundas y algoritmos de predicción de saltos [Dani96].

En 1989 se lanzan al mercado el i486 y el MC68040 con 1.2 millones de transistores cada uno y rendimiento similar [Yu96]. El MC68040 estaba segmentado en 6 etapas, con memoria cache de instrucciones y datos de 4 Kbytes cada una y coprocesador matemático. El i486 funcionaba a 25 MHz con tecnología de 1 micra y a 50 MHz con la de 0.8 micras, incluía un coprocesador matemático, una cache de 8 Kbytes y soporte para memoria virtual, además de segmentación. El éxito de Intel en el mercado de computadores personales atrae a competidores (AMD, Cyrix, etc.) a crear soluciones compatibles con la familia x86.

En los microprocesadores para estaciones de trabajo aparecen evoluciones de la segmentación como la ejecución supersegmentada y la ejecución

superscalar. En los procesadores supersegmentados se disminuye el tiempo de ciclo a costa de aumentar el número de etapas del pipeline. Los procesadores superescalares son capaces de lanzar más de una instrucción por ciclo con el objeto de explotar el paralelismo a nivel de instrucción (Instruction Level Parallelism, ILP). No obstante, los primeros procesadores capaces de ejecutar múltiples instrucciones por ciclo fueron dos computadores de los años 60, el CDC 6600 [Thor64], del que hemos hablado al describir los avances arquitectónicos de la segunda generación, y el IBM 360/91 [AnST67], donde ya se incluye etiquetado de instrucciones y renombramiento de registros. Ejemplos de esta generación son Sun SuperSPARC, el HP-PA 7100, MIPS R4000, Intel 80960CA y Motorola 88110.

Otros hitos importantes de esta generación son la estación de trabajo Sun 3/260 que en 1986 incorpora por primera vez dos niveles de cache, uno interno y otro externo. En 1991 el Cray Y-MP C90, que tiene 16 procesadores, consigue alcanzar 16 GFLOPs.

Destacar por último, por su impacto en las metodologías de diseño actuales, que a finales de los 80 se funda la organización SPEC (System Performance and Evaluation Company) (http://www.specbench.org/spec/) con el objetivo inicial de proporcionar un método de comparación más realistas para el mercado de servidores y estaciones de trabajo que los hasta la fecha populares MIPS, MFLOPS o los kernels sintéticos tipo Whetstone. La primera propuesta de SPEC hoy se conoce con el nombre de SPEC89 y estaba orientada casi exclusivamente al rendimiento de procesador. A esta versión inicial han seguido tres nuevas entregas, SPEC92, SPEC95 y SPEC CPU2000, así como nuevos benchmarks destinados a medir otros elementos del computador como el sistema gráfico (SPECviewperf y SPECapc) o el sistema de ficheros (SPECSFC). Desde su aparición ha sido un referente importante utilizado tanto por la industria como dentro del mundo académico donde es uno de los bancos de pruebas más utilizados para explorar nuevas ideas o analizar diferentes alternativas de diseño.

Avances arquitectónicos: tercera etapa

A finales de los 90 hemos asistido a la vertiginosa expansión de Internet y del Word Wide Web (WWW), al éxito comercial de los PDAs (personal digital assistants) y a la proliferación de otros productos electrónicos de gran consumo como las consolas de video juegos, las cámaras digitales o los reproductores de MP3. Estos cambios, han dado lugar a tres segmentos claramente diferenciados dentro del mercado de los computadores, cada uno de ellos con diferentes aplicaciones objetivo, diferentes requisitos y diferentes tecnologías: computadores personales, servidores y procesadores empotrados. Vemos un poco más en detalle cada uno de ellos

Computadores Personales y Estaciones de Trabajo (Desktop Computing)

Es el segmento con mayor volumen de negocio. Comprende desde equipos de gama baja, de menos de 1000 dólares, hasta potentes estaciones de trabajo de más de 10.000 dólares. Para la mayoría de los usuarios, las decisiones de compra vienen determinadas por la relación entre el rendimiento (tanto del procesador como del sistema gráfico) y el precio del sistema. Por lo tanto, el factor precio-rendimiento es el principal objetivo de diseño.

Dentro de este segmento, hemos asistido a una intensa batalla por el dominio del mercado de los PCs compatibles, lo que ha incentivado una importante mejora de prestaciones. Actualmente sólo sobreviven la propia Intel y Advanced Micro Devices (AMD), si bien en el ámbito de los equipos portátiles, lucha por hacerse un hueco, aunque con bastantes dificultades económicas, Transmeta Corporation, con su procesador Crusoe del que hablaremos en la sección 0. Es una familia revolucionaria x86-compatible especialmente diseñada para el mercado de ordenadores móviles de mano y de peso ligero. El procesador Crusoe de alto rendimiento consume 60 al 70 por ciento menos de potencia (según el fabricante)y trabaja mucho más refrigerado que los chips que compiten con él, transfiriendo la parte más compleja del trabajo del

procesador -la determinación de instrucciones a ejecutar y cuando - a software en un proceso llamado Code Morphing.

Como consecuencia de esta competencia, el último procesador de la familia x86, el Pentium 4, se ha puesto a la altura de los mejores procesadores RISC incluso en el proceso en punto flotante (Tan solo el IBM Power4, con multiprocessor-on-chip supera al Pentium 4 en SPEC CPU2000fp). No obstante, el éxito de este nuevo componente de la familia x86 no esta en contradicción con las ventajas atribuidas a la filosofía RISC, ya que aunque la arquitectura Pentium mantiene por cuestiones de compatibilidad la ilusión de una arquitectura x86, internamente se utiliza un núcleo RISC. Para ello, en la fase de decodificación se traducen dinámicamente las complejas instrucciones x86 (IA-32) en microoperaciones más sencillas, que se pueden ejecutar fuera de orden por el núcleo RISC de este procesador. Entre las propuestas del ámbito académico recogidas por el Pentium 4 destacan la Trace Cache o el Multithreading Simultáneo (SMT).

El concepto de Trace Cache fue presentado por primera vez en 1996 [RoBS96]. La idea básica es la de capturar el comportamiento dinámico de las secuencias de instrucciones, almacenando trazas de instrucciones en lugar de bloques contiguos (contiguos tras una ordenación estática).

El SMT es otra novedad que lleva bastante tiempo dentro del mundo académico[YaNe95][EELS97]. La idea es permitir que haya instrucciones de diferentes flujos de ejecución conviviendo dentro del procesador. En cada ciclo se realiza la búsqueda de instrucciones para diferentes threads, manteniendo separados recursos como el banco de registros mediante un exhaustivo control en la asignación. En las primeras implementaciones del Pentium 4, el SMT (limitado a 2 threads en este procesador) estaba desactivado. Actualmente comienzan a salir unidades (Pentium Xeon) [Inte9que hacen uso de esta potente posibilidad.

La importancia de las aplicaciones multimedia dentro de este segmento ha motivado la inclusión en todos los microprocesadores de mejoras y extensiones para acelerar específicamente este tipo de aplicaciones. La clave está en que mientras los microprocesadores de propósito general están optimizados para manejar datos de 32 ó 64 bits, en las aplicaciones multimedia es habitual tratar con flujos continuos de datos más cortos (pixels de 8 bits, señal de audio de 16 bits, etc). Para explotar esta característica se han incluido operaciones tipo SIMD, aprovechándose así el ancho de los datapath y las unidades funcionales. Además, es habitual incluir mecanismos automáticos o semi-automáticos para realizar prebúsqueda de datos y operaciones de carga y almacenamiento que evitan (hacen un bypass) los diferentes niveles de cache a fin de paliar los problemas de localidad. Recientemente se han añadido nuevas extensiones para tratamiento de gráficos en 3D, extendiéndose las operaciones SIMD a datos en punto flotante (en simple e incluso en doble precisión). El primero en incorporarlas fue AMD, con la extensión 3DNow! para los procesadores K6-II y posteriores, aunque no tardaron en aparecer las extensiones SSE (Streaming SIMD Extensions) y SSE2 al repertorio x86 de Intel y AltiVec de Motorola en el PowerPC G4.

Servidores

El mercado de servidores esta dominado por multiprocesadores simétricos de memoria compartida (SMPs) y por clusters. En este segmento la relación coste-rendimiento no es tan decisiva, siendo en este caso mucho más relevantes factores como la alta disponibilidad, la escalabilidad o la productividad (throughput). La alta disponibilidad hace referencia a la necesidad de que los sistemas estén operativos en todo momento, lo cual lleva inherente la necesidad de algún tipo de redundancia, ya que en servidores de gran escala son inevitables los fallos. La escalabilidad es también un aspecto importante, ya que las necesidades de cómputo, memoria, disco o entrada-salida de los servidores suele crecer durante el tiempo de vida de estos sistemas. Por último, aunque es importante el tiempo de respuesta que

ofrecen, las métricas de rendimiento más importante a la que deben hacer frente los diseñadores de estos sistemas, como los SPECrate, los TPC o los recientes SPECSFS y SPECWeb, son medidas de productividad.

Los SPECrate se obtienen a partir de los SPEC CPU2000 ejecutando múltiples instancias de dichos benchmarks (habitualmente una por procesador). No obstante, como la productividad de los servidores no sólo depende de la capacidad de computo, SPEC desarrolló los SPECSFS, diseñados para evaluar no sólo el procesador sino también el sistema de entrada-salida, tanto el almacenamiento secundario (discos) como el interfaz de red. La importancia del WWW dentro de este segmento también ha sido recogida por SPEC mediante el SPECWeb, otro benchmark orientado a medir la productividad en el que se simula un entono de múltiples clientes que solicitan páginas (estáticas y dinámicas) y envían información a un servidor Web.

Los benchmarks con más historia dentro de este segmento son los TPC. De hecho, el Transaction Processing Council (TPC) fue creado con anterioridad a la organización SPEC (a mediados de los 80), con el objetivo de crear benchmarks realistas para el procesamiento de transacciones. Como en el caso de SPECSFS y SPECWeb, los TPC evalúan el comportamiento global del sistema, es decir, no solamente el procesador sino también el subsistema de entrada-salida, el sistema operativo y el gestor de base de datos utilizado. Las métricas utilizadas por esta organización son transacciones por minuto (TPM) y TPMs por dolar, aunque también incluyen requisitos para el tiempo de respuesta (sólo se consideran las transacciones que satisfacen dichos requisitos). Para modelar configuraciones realistas, tanto el número de usuarios como el tamaño de la base de datos se escala con la potencia del servidor.

Entre los microprocesadores utilizados en este segmento se encuentra la gama Pentium Xeon de Intel, y otros microprocesadores tipo RISC (también disponibles en el mercado de estaciones de trabajo) como el UltraSPARC III de SUN, el Power 4 de IBM, el HP PA-8700, el Alpha 21264 y el MIPS R14000

utilizado en las máquinas de SGI. No obstante, algunos de estos procesadores tienen un futuro incierto. La tecnología Alpha fue adquirida por Intel, que ha apostado por una nueva arquitectura conocida como **IA-64**, en cuyo desarrollo también participa HP y en la que han demostrado gran interés otros fabricantes como SGI. Esta arquitectura ha suscitado un gran debate dentro del área por hacer uso de **VLIW** (Very Long Instruction Word) del que hablaremos también en la sección 0.

Finalmente destacar que durante los últimos años hemos asistido a la irrupción de los **clusters** como una alternativa económica, especialmente a los multiprocesadores de gran escala.

Procesadores Empotrados

Los procesadores empotrados representan el segmento del mercado con un mayor crecimiento. Están presentes en multitud de dispositivos, desde tarjetas inteligentes y controladores industriales a sofisticados conmutadores de red o consolas de videojuegos. Es por ello el segmento en el que se aprecia una mayor variedad tanto en prestaciones como en coste. El factor de diseño más importante es en este caso el precio. Existen obviamente algunos requisitos relativos con el rendimiento, a menudo relacionados con alcanzar tiempos de respuesta en tiempo real, pero el primer objetivo suele ser alcanzar dichas prestaciones con el mínimo coste posible. De hecho, a diferencia de los computadores personales o los servidores, los benchmarks para este segmento puede considerarse que están aún en su infancia. El intento de estandarización que ha tenido más éxito hasta la fecha son los denominados EEMBC (EDN Embedded Microprocessor Benchmark Consortium). Sin embargo, muchos fabricantes siguen facilitando el resultado de kernels sintéticos ya obsoletos en los otros segmentos como Dhrystone o medidas basadas en MIPS.

Para garantizar tiempo real, los diseñadores de este tipo de sistemas tratan de optimizar el peor caso posible, a diferencia de los microprocesadores de propósito general donde siempre se intenta favorecer las situaciones más

probables a costa incluso de penalizar al resto. Otros factores de diseño importantes en algunas aplicaciones el tamaño de la memoria y el consumo de potencia. El primer factor esta relacionado directamente con el coste del sistema, aunque también se relaciona con el consumo de potencia al ser la memoria uno de los componentes de mayor consumo. Este factor se traslada a menudo en un énfasis por reducir el tamaño de los códigos, existiendo en algunos casos soporte hardware para este propósito. La preocupación por el consumo de potencia esta relacionada en la mayoría de los casos por el uso de baterías. No obstante también guarda relación con el coste, ya que un menor consumo permite por ejemplo utilizar empaquetados plásticos, más económicos que los cerámicos, y evita la necesidad de incorporar ventiladores.

Dentro de este segmento, el diseñador puede optar por tres aproximaciones diferentes:

- Combinación Hardware/Software que incluye algún circuito de propósito específico integrado junto a algún core.
- Procesador empotrado genérico (off-the-self) con el software específico para resolver el problema.
- DSP (Digital signal processor) o un procesador multimedia (media processor) con el software específico para resolver el problema.

En En los 90 la diferencia entre microprocesadores y microcontroladores se ha ensanchado cada vez más. Los primeros buscan aumentar su rendimiento y los segundos se concentran en disminuir el coste y aumentar la integración del sistema, incluyendo en el mismo chip funciones analógicas, todo tipo de memorias y sensores.

Conclusiones

En los primeros años de su existencia, los computadores tenían un coste muy elevado, por lo que su uso y comercialización estaban restringidos para cálculos complejos y por lo tanto los usuarios eran especialistas.

En la segunda generación se produce un gran avance tecnológico con la invención del transistor. Los computadores se usan, como anteriormente, para cálculos científicos complejos. Por eso se aprovecha la mejora tecnológica para aumentar las capacidades de cálculo de los computadores con segmentación y paralelismo y se crean sistemas operativos por lotes y lenguajes de alto nivel para mejorar el aprovechamiento de los mismos. Aparecen los supercomputadores.

En la tercera generación se amplía la gama de los computadores con la creación de los minicomputadores y mainframes. Las mayores capacidades que ofrecen los circuitos integrados permiten o bien aumentar las funciones del computador para resolver problemas aún más complejos, o disminuir su coste, de forma que se abre el mercado a nuevos usuarios, todavía del entorno de la ciencia o empresarial.

Hasta esta época, aproximadamente 1970, los cursos de arquitectura de computadores impartían fundamentalmente aritmética de computadores [Patt98].

A partir de entonces se produce un enorme cambio en el uso de los computadores con la invención de los microprocesadores y la memoria de semiconductores en la cuarta generación. Esto reduce los costes de los computadores, de forma que están disponibles para muchas nuevas aplicaciones. Aparecen los microcontroladores y los computadores personales. Simultáneamente se aprovecha la mayor densidad de circuitos integrados para construir mainframes cada vez más potentes para uso científico y empresarial.

También se crean los supercomputadores vectoriales, para cálculos científicos complejos.

Esto se refleja en que durante la década de los 70 los cursos de arquitectura estudiaban básicamente diseño de la arquitectura del repertorio de instrucciones, especialmente repertorios apropiados para facilitar la tarea de los compiladores (CISC).

Finalmente, en la quinta generación se ha producido la invasión de la sociedad por parte de los sistemas basados en microprocesador. La tecnología proporciona circuitos cada vez más complejos y rápidos, pero la desigual evolución de la velocidad del procesador, la de memoria y la de E/S sugiere cambiar el enfoque seguido para el diseño de la arquitectura (sólo ocuparse del procesador) y considerar el diseño del sistema completo. Se amplía la jerarquía de memoria y se usan buses jerárquicos (local, del sistema, de expansión). Hay mejoras en la gestión de E/S con DMA y procesadores de E/S.

En resumen, los sistemas basados en microprocesador, tanto los computadores personales como las estaciones de trabajo, han aumentado de rendimiento y complejidad de forma vertiginosa. Esto permitió en un primer momento que las estaciones de trabajo sustituyeran a los minicomputadores y las mainframes y posteriormente que los mismos PCs se incorporen a este mercado. De hecho, la mayor parte de los supercomputadores y sistemas multiprocesador actuales se construyen con componentes del mercado de servidores y de PCs de gama alta. Por ello, los cursos de arquitectura actuales imparten diseño de CPU, del sistema de memoria, del de E/S y multiprocesadores. También es necesario ampliar la oferta de materias optativas para complementar la formación y los conocimientos.

Un objetivo fundamental: el incremento del rendimiento

A lo largo de esta exposición se ha comprobado que uno de los objetivos fundamentales de los avances tecnológicos y arquitectónicos ha sido incrementar el rendimiento de los computadores. Esta tendencia se ha visto matizada en los últimos años con la aparición de otros objetivos de diseño que también deben tenerse en cuenta como el consumo (en el creciente mercado de computadores portátiles y los procesadores empotrados), la mayor integración del sistema (en sistemas empotrados) y la fiabilidad [Henn99]. A pesar de esto, conseguir mejorar el rendimiento de los computadores continuará siendo un objetivo clave debido a la competencia y a que nuevas mejoras abren nuevas posibilidades de resolver problemas cada vez más complejos.

Los avances tecnológicos han permitido integrar más y más elementos en un único circuito integrado, es decir, los arquitectos disponen de más espacio y componentes para fabricar el procesador. La pregunta es, ¿Cómo distribuir estos componentes para mejorar el rendimiento?. En esta sección se describirán diferentes ideas que han ido apareciendo en los últimos años y al final del capítulo se mostrarán algunas de las líneas abiertas en la actualidad.

Procesadores Superescalares

Actualmente, el mercado de procesadores de propósito general está dominado por procesadores superescalares con ejecución fuera de orden. Se caracterizan por realizar la búsqueda de múltiples instrucciones por ciclo. Estas instrucciones son decodificadas y entran en una ventana donde se controlan las dependencias entre ellas. Aquellas que están libres de dependencias (es decir, que tienen sus operandos fuente disponibles), se emiten para su ejecución en las unidades funcionales. En este punto hay varias alternativas de implementación, desde las que proponen emitir (issue) las instrucciones desde la ventana a las unidades funcionales directamente, hasta las que emiten (dispatch) las instrucciones a unas colas (o estaciones de reserva) colocadas delante de las unidades funcionales y allí se ejecutan en orden. Una vez ejecutadas las instrucciones, éstas escriben sus resultados en el banco de registros y se retiran del pipeline en orden de programa para poder mantener las interrupciones precisas. Algunos ejemplos de procesadores superescalares son el MIPS R14000, el Alpha 21264, el HP PA-8700 [www.hp.com], el Intel Pentium 4 [www.intel.com] o el AMD Athlon [www.amd.com] [AMD99b]. El diagrama de bloques del MIPS R14000 se puede ver en la Figura 26.

El R14000 es un representante típico de los procesadores RISC modernos que son capaces de la ejecución especulativa de instrucciones. Como en el procesador Alpha de Compaq hay dos unidades de punto flotante independientes para la suma y la multiplicación y, además, dos unidades que realizan la división flotante y operaciones de raíz cuadrada (no se muestran en en la figura).

Su sucesor, el R18000 está planificado para su comercialización durante el año que viene y tendrá una estructura diferente, aunque no se sabe mucho aún sobre él. La frecuencia de reloj aumentará a 600 - 800 MHZ pero la diferencia fundamental será que el R18000 será un procesador dual con 4 unidades de punto flotante por chip, de la misma manera que IBM planifica dos CPUs sobre

un chip en sus procesadores POWER4. Para el R18000 se puedenesperar un pico teórico de > 3 Gflop/s.

Aunque los procesadores Pentium no se aplican en sistemas integrados paralelos actualmente, juegan un papel principal en la comunidad de cluster. Hay dos modos principales de aumentar el rendimiento de un procesador: aumentando la frecuencia de reloj y aumentando el número de instrucciones por ciclo (IPC). Estos dos accesos están generalmente en el conflicto: cuando uno quiere aumentar el IPC, el chip se hace más complicado. Esto tendrá un impacto negativo sobre la frecuencia de reloj porque hay que realizar y organizar más trabajo dentro del mismo ciclo de reloj. Intel ha escogido una velocidad de reloj alta (al principio aproximadamente el 40 % más que la del Pentium III con la misma tecnología de fabricación) mientras el IPC ha disminuido entre un 10 y 20 %. El procesador Athlon tiene muchos rasgos que están también presentes en procesadores RISC modernos, en realidad el AMD Athlon es un clon en lo que concierne a la Arquitectura de Juego de Instrucciones de la familia x86 del Intel. En la Figura 27 se muestra el diagrama de bloques del AMD Athlon.

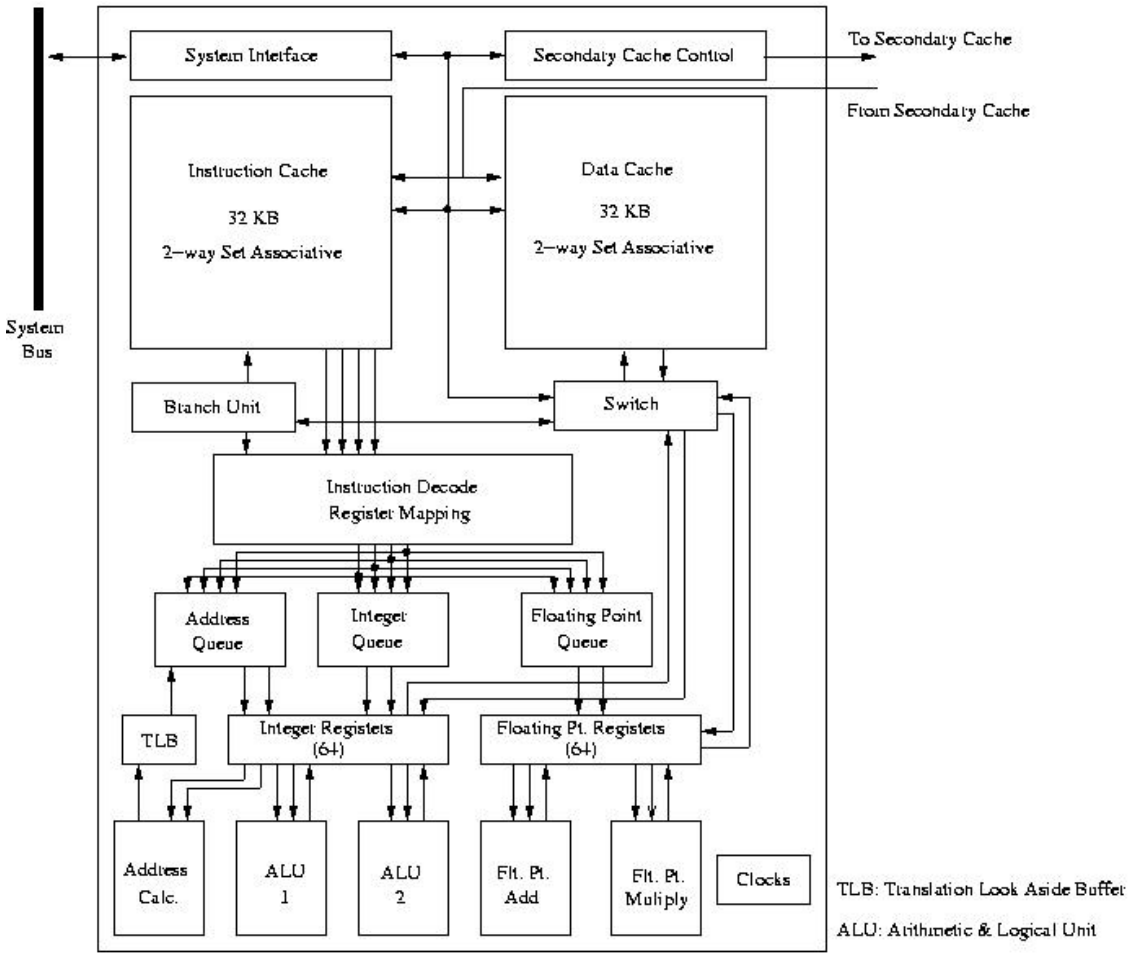

Figura 26: Diagrama de Bloques del MIPS R14000

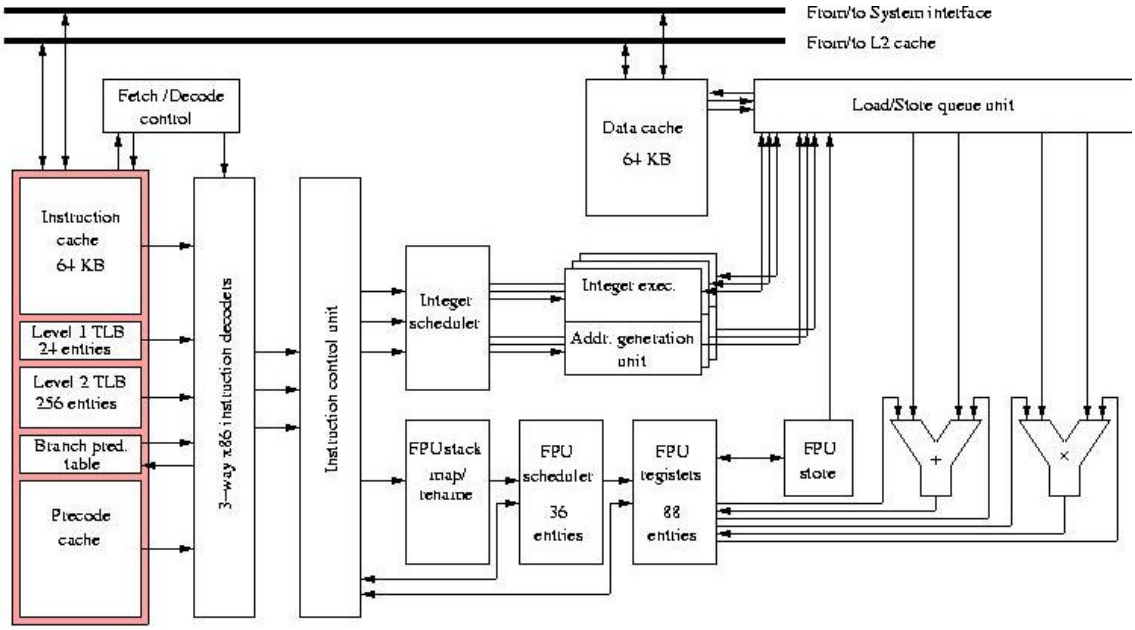

Figura 27: Diagrama de Bloques del AMD Athlon

El principal problema que plantea este tipo de arquitecturas desde el punto de vista del rendimiento, es el bajo aprovechamiento del ILP potencial en la ejecución de programas reales. Algunos estudios muestran que una máquina hipotética con un ancho de 2 instrucciones (es decir, una que es capaz de emitir y completar un máximo de 2 instrucciones en un solo ciclo de máquina) alcanza un promedio de instrucciones por ciclo (IPC) de entre 0,65 y 1,40. En cambio, una máquina hipotética similar con un ancho de 6 instrucciones podría alcanzar como máximo entre 1,2 y 2,4 IPC [Oluk96]. Esta merma del IPC potencial se debe principalmente a los siguientes factores:

- Limitación de recursos. A primera vista este problema parece tener fácil solución gracias al aumento constante de la capacidad de integración de los circuitos integrados, pero aumentar la complejidad del procesador va en detrimento de otros factores como el consumo y, posiblemente, del ciclo de reloj.

- Dependencias. Las dependencias en los programas pueden ser debidas al flujo de datos, al flujo de control o a la reutilización de registros. En el caso de las dependencias de control, la mayor parte de los procesadores incluyen un predictor de saltos, el cual permite continuar la búsqueda de instrucciones por uno de los caminos de una manera especulativa. Más recientemente, el Alpha 21264 incluye una tabla que permite predecir dependencias entre loads y stores, de tal manera que se puede trabajar con especulación de loads y stores.

- El ancho de banda de la memoria. A pesar de las técnicas desarrolladas para reducir la demanda de ancho de banda con memoria (memoria cache, TLB, prebúsqueda, predicción de dependencias entre instrucciones de carga y almacenamiento, etc.), las instrucciones de acceso a memoria siguen siendo uno de los principales cuellos de botella de los computadores actuales.

- Compatibilidad binaria. La imposibilidad de cambiar el juego de instrucciones (ISA) del procesador limita en gran medida la implementación de diferentes modelos de ejecución que permitan obtener ganancias en el rendimiento sin afectar a otros factores como el consumo y el reloj.

- La influencia desigual de determinadas mejoras en el conjunto de aplicaciones a ejecutar. Determinados mecanismos pueden producir un efecto beneficioso en algunas aplicaciones y pernicioso en otras, por lo tanto, sería deseable tener la posibilidad de su control en tiempo de ejecución.

Procesadores VLIW y Arquitectura EPIC

Una alternativa para simplificar el hardware que los procesadores superescalares dedican a la extracción dinámica de paralelismo a nivel de instrucción la constituyen los procesadores VLIW (Very Long Instruction Word). El diseño de estos procesadores es bastante sencillo. Las instrucciones definen varias operaciones que el procesador puede realizar en paralelo. En este caso, es el compilador es el responsable de encontrar las operaciones que pueden lanzarse juntas y crear una instrucción que contenga estas operaciones. Esto se realiza de forma estática y el hardware no necesita chequear las dependencias.

Los primeros procesadores VLIW fueron coprocesadores que se añadían a un procesador central. En la década de los 80 aparecieron más computadores que incluían procesadores VLIW, como el Trace de Multiflow o el Cydra-5 [SiFK97] de Cydrome.

En la actualidad, este tipo de arquitectura es frecuente en el segmento de los procesadores empotrados. Algunos ejemplos son el Star Core, de Motorola y Lucent, la familia C6xxx [HePa02] de Texas Instruments, o el Trimedia de Philips.

La gran ventaja de los procesadores VLIW radica en su simplicidad de diseño, que les permite tener frecuencias de reloj altas con un bajo consumo. Sin embargo presentan dos inconvenientes claros. Por un lado, el compilador tiene todo el peso en la obtención del paralelismo a nivel de instrucción. Como éste no dispone de toda la información que se genera dinámicamente, suele ser bastante conservador a la hora de tomar decisiones. Esto hace que se requieran técnicas sofisticadas de compilación. Otro problema está en la incompatibilidad entre diferentes generaciones. Como la instrucción larga tiene un formato fijo, donde hay una operación para cada unidad funcional, si se añaden más unidades funcionales la instrucción queda obsoleta y, por lo tanto, los programas se tienen que recompilar.

Profundizando en esta línea, Intel y HP han propuesto una nueva arquitectura denominada EPIC (Explicitly Parallel Instruction Computing). El compilador es el responsable de explotar el paralelismo haciéndolo explícito en el código máquina. Las instrucciones se agrupan de 3 en 3 y a cada grupo se le añade información sobre la dependencia entre instrucciones. Como alternativa a la ejecución especulativa se hace uso de la predicción.

Procesadores Multithreaded

Las arquitecturas multithreaded aparecen debido a las limitaciones que imponen las dependencias de datos y la alta latencia de memoria en los procesadores convencionales. Se caracterizan por mantener varios flujos ejecutándose a la vez dentro del procesador. Dichos flujos pueden ser de una misma aplicación, con lo que se reduciría su tiempo de ejecución, o de aplicaciones diferentes, con lo que se aumentaría la productividad (el throughput) del sistema. Dentro de este tipo de arquitecturas pueden distinguirse arquitecturas de grano fino y arquitecturas de multithreading simultáneo [EELS97].

Las primeras tiene como objetivo principal esconder la latencia de memoria. En cada ciclo, el procesador ejecuta instrucciones de threads diferentes, es decir, no se llegan a emitir en el mismo ciclo instrucciones de flujos diferentes. Uno de los primeros procesadores que siguieron esta filosofía fue el MTA de Tera. También se puede encontrar este paradigma en cada uno de los microprocesadores específicos que se encuentran dentro del procesador de red IXP1200 de Intel [www.Intel.com].

Las arquitecturas de multithreading simultáneo (Simultaneous Multithreading, o SMT) permiten que haya instrucciones de diferentes flujos de ejecución conviviendo a la vez dentro del procesador. En cada ciclo se realiza la búsqueda de instrucciones para los diferentes threads, manteniendo separados recursos como el banco de registros. Dependiendo de la implementación, las instrucciones de los diferentes threads comparten la cola de instrucciones y las unidades funcionales. Como ya se ha comentado en la sección anterior, el ejemplo más destacable de este tipo de arquitecturas es el Pentium 4 Xeon.

El principal problema de estas arquitecturas es el soporte que requieren para la generación de los threads, especialmente si lo que se pretende es reducir el tiempo de ejecución de una aplicación extrayendo lo que se conoce

como paralelismo a nivel de thread (Thread Level Parallelism o TLB). A este respecto existen varias alternativas: el compilador puede realizar un análisis del código y determinar qué threads se pueden generar. Esta aproximación puede ser inviable para aplicaciones donde no existe paralelismo explícito de grano fino. Otra posibilidad es que sea el programador quien, a través de directivas (por ejemplo OpenMP), especifique el número de threads de que consta la aplicación. La alternativa más agresiva es dejar que sea el hardware el encargado de extraer, especulativamente, el TLP de las aplicaciones.

Procesadores en cluster

Las aproximaciones más utilizadas para mejorar el rendimiento de los procesadores superescalares han consistido en aprovechar el aumento en la escala de integración para incrementar el tamaño de las caches, el buffer de reordenamiento y las ventanas de instrucciones.

No obstante, como ya hemos comentado en la sección anterior, esta mejora en la escala de integración también lleva asociado un aumento en los retardos de las comunicaciones dentro del chip.

Una manera de aumentar la escalabilidad y reducir los efectos de los retardos es organizar la ventana de instrucciones en diferentes ventanas, más pequeñas e independientes, desde donde las instrucciones se emiten a las unidades funcionales. Esta arquitectura es más escalable, ya que las ventanas son pequeñas y por lo tanto la lógica de selección y emisión es más sencilla.

El principal problema que se plantea en estas arquitecturas radica en la lógica de decodificación y distribución de instrucciones entre las diferentes ventanas. Como la comunicación de datos entre instrucciones de la misma cola, el objetivo es minimizar las comunicaciones entre instrucciones que no pertenecen a la misma cola. Como ejemplo de microprocesador actual donde se ha incorporado esta idea está el alpha 21264. Dicho procesador divide las unidades de aritmética entera en dos clusters, cada uno asociado a un banco

de registros, consiguiéndose reducir de esta forma el tiempo de acceso al banco de registros al disminuirse el número de puertos de lectura y escritura. También hacen uso de esta organización en cluster algunos procesadores empotrados y DSP, como la familia TI TMS320C6x de Texas Instruments [www.ti.com].

Multiprocesadores en un chip

Otra manera de aprovecharse del aumento de la escala de integración sin complicar la arquitectura del procesador es la integración en un único chip de varios procesadores. El objetivo en este caso no es tanto disminuir el tiempo de ejecución de las aplicaciones sino más bien aumentar la productividad (throughput). El beneficio de este tipo de arquitecturas puede ser importante en aplicaciones paralelas, donde el tiempo de comunicación puede verse sensiblemente reducido.

Una de las primeras propuestas de este tipo de arquitecturas fue el Standford Hydra [ONH+96], en el que se propone un sistema formado por cuatro procesadores en un mismo chip compartiendo la cache de segundo nivel. Cada uno de estos procesadores tiene un ancho de emisión de dos instrucciones por ciclo y son más sencillos que los procesadores actuales.

El ejemplo más importante de procesador comercial es el Power4 de IBM, que incluye dos procesadores en un mismo chip, comunicados por una cache de segundo nivel que también reside dentro del chip.

Debido a las características de las aplicaciones que ejecutan, también es una organización común en el ámbito de procesadores de red. Ejemplos comerciales son el IBM Rainer o el Intel IXP1200, que incluye varios procesadores multithreaded dentro del chip.

Supercomputadores

Dentro del campo de los supercomputadores podemos distinguir dos tipos: en primer lugar los procesadores vectoriales, que explotan el paralelismo a nivel de datos, y en segundo lugar los multiprocesadores.

Computadores vectoriales

Además de las técnicas de segmentación, se han desarrollado otras que permiten aumentar el rendimiento de los computadores. Como ya se ha mencionado anteriormente, una de las limitaciones más serias para explotar el paralelismo a nivel de instrucción es la falta de paralelismo inherente en los programas. No es el caso sin embargo de muchas de las aplicaciones utilizadas en ciencia e ingeniería basadas en el procesamiento de vectores, donde existe una gran cantidad de paralelismo a nivel de datos. Por eso algunas de las soluciones que se presentan para explotar este paralelismo están orientadas al tratamiento de grandes estructuras de datos. Se pueden distinguir fundamentalmente dos tipos:

Los computadores matriciales o computadores en array [Hwan93] se caracterizan por disponer de múltiples réplicas de la sección de procesamiento de datos, supervisadas por una única unidad de control. De esta forma pueden operar en paralelo (y síncronamente) sobre distintos elementos de una misma estructura de datos. Las primeras ideas sobre este tipo de arquitectura son del Illiac IV [Barn68], diseñado en 1962, de la que sólo se implementó la cuarta parte y se obtuvo un rendimiento de 50MFLOPs, 20 veces menos de lo esperado. En 1985 se utilizaron las ideas del Illiac IV para construir la Connection Machine (CM), que tenía 65.636 procesadores de 1 bit. Otros ejemplos son MasPar MP-1, CM-2, DAP600, etc. Este tipo de arquitectura está en desuso para máquinas de propósito general, debido a su poca flexibilidad (no es escalable), y a que no puede utilizar las ventajas que proporciona la tecnología de microprocesadores. Sin embargo, este estilo de arquitectura sigue

teniendo cierto interés en diseños de propósito especial, como el procesamiento de imágenes.

Los computadores vectoriales segmentados son capaces de ejecutar instrucciones especiales para manipulación de vectores. Para ello disponen de unidades funcionales segmentadas para procesamiento vectorial, capaces de aceptar un componente del vector por ciclo de reloj. Los primeros computadores vectoriales realizaban las operaciones entre vectores desde memoria, como el CDC STAR-100 de 1972 y el CYBER-205, de 1981. Desde finales de los 80, están orientados a registro, es decir, todas las operaciones sobre vectores (excepto la carga y almacenamiento) se realizan en registros. Los principales ejemplos de este tipo de arquitecturas son los computadores de Cray (Cray-1, Cray-2, Cray X-MP, Cray Y-MP, Cray C-90, Cray J-90, Cray T-90, Cray SV1), los computadores japoneses de Fujitsu (VP100/200, VPP300/700, VPP5000), Hitachi (S810/820, S3000, SR8000 -este último no es estrictamente un computador vectorial sino seudo-vectorial que lee un flujo de datos consecutivos de memoria y los alimenta a la unidad aritmética) y NEC (SX/2, SX/3, SX-4, SX-5) y los minicomputadores Convex (C-1, C-2, C-3, C-4).

Hay que señalar, que en estas arquitecturas vectoriales también juega un papel fundamental la tecnología de compiladores, ya que para poder ejecutar en estas máquinas las aplicaciones científicas extrayendo su paralelismo inherente, es imprescindible contar con compiladores capaces de generar código vectorial a partir de programas escritos en lenguajes de alto nivel secuenciales [Zima91].

Multiprocesadores

Las arquitecturas paralelas basadas en microprocesadores surgen debido al pequeño tamaño, el bajo coste y el alto rendimiento de los microprocesadores. Estas arquitecturas ofrecen importantes ventajas en cuanto a la fabricación, la relación precio/rendimiento y la fiabilidad frente a los computadores más

tradicionales. Se pueden encontrar tres tipos de computadores paralelos [Hwang93]

- Los multiprocesadores de memoria compartida o de acceso a memoria uniforme (UMA)
- Los multicomputadores de memoria distribuida por paso de mensajes.
- Los multiprocesadores de acceso a memoria no uniforme (NUMA)

Los multiprocesadores de memoria compartida se caracterizan por que los microprocesadores comparten un único espacio de direcciones de memoria. Se les llama también de UMA porque todos los accesos a memoria tardan el mismo tiempo. Las tareas de cada procesador pueden compartir datos de la memoria, por eso es necesaria la sincronización y el acceso exclusivo para asegurar la consistencia. Ejemplos típicos son el SG Power Challenge (basado en el procesador R8000/R10000), el DEC 80000 basado en el procesador DEC 21164 y Sun Server 10000.Su principal ventaja es la facilidad de programación. Como inconveniente tienen su falta de escalabilidad.

Esto potenció el estudio de los multicomputadores de memoria distribuida que consistían en múltiples procesadores conectados en red principalmente como malla o como hipercubo. Ejemplos típicos son el IBM SP2 basado en el procesador Power-2 o el Vpp 5000 basado en un procesador vectorial VLSI El principal problema de estas arquitecturas es la dificultad de realizar una programación eficiente. Los multiprocesadores de acceso a memoria no uniforme (NUMA) se sitúan entre los distribuidos y los de acceso a memoria uniforme y se caracterizan por tener espacios de memoria virtual únicos. Es decir, su memoria física esta distribuida pero, desde el punto de vista lógico, es un solo espacio de direcciones. En este caso la memoria es de acceso no uniforme porque el tiempo de acceso depende de la memoria local a la que se acceda. A su vez se pueden clasificar en los sistemas que no tienen coherencia

cache (nccNUMA) como el Cray T3E basado en el procesador Alpha, y los que tienen coherencia cache (ccNUMA).

#	Fab	Computer	R_{max}(GFlops)	País	Año
1	NEC	Earth-Simulator	35860.00	Japan	2002
2	IBM	ASCI White, SP Power3 375 MHz	7226.00	USA	2000
3	HP	AlphaServer SC ES45/1 GHz	4463.00	USA	2001
4	HP	AlphaServer SC ES45/1 GHz	3980.00	France	2001
5	IBM	SP Power3 375 MHz 16 way	3052.00	USA	2001
6	HP	AlphaServer SC ES45/1 GHz	2916.00	USA	2002
7	Intel	ASCI Red	2379.00	USA	1999
8	IBM	pSeries 690 Turbo 1.3GHz	2310.00	USA	2002
9	IBM	ASCI Blue-Pacific SST, IBM SP 604e	2144.00	USA	1999
10	IBM	pSeries 690 Turbo 1.3GHz	2002.00	USA	2002
11	IBM	SP Power3 375 MHz 16 way	1910.00	UK	2002
12	IBM	pSeries 690 Turbo 1.3GHz	1840.00	USA	2002
13	Hitachi	SR8000/MPP	1709.10	Japan	2001
14	Hitachi	SR8000-F1/168	1653.00	Germany	2002
15	SGI	ASCI Blue Mountain	1608.00	USA	1998
16	IBM	SP Power3 375 MHz	1417.00	USA	2000
17	IBM	SP Power3 375 MHz 16 way	1293.00	Germany	2001
18	IBM	SP Power3 375 MHz 16 way	1272.00	USA	2001
19	NEC	SX-5/128M8 3.2ns	1192.00	Japan	2001
20	IBM	SP Power3 375 MHz	1179.00	USA	2000

Tabla 2: Los 20 supercomputadores más potentes en Junio de 2002

Cada año, desde 1993, la organización TOP500 Supercomputing Sites publica en junio y noviembre la lista con los 500 computadores más potentes, utilizando como medida de su rendimiento el benchmark Linpack. En la Tabla 2 aparece la versión de los 20 primeros en Junio de 2002 [TOP99].

	1997	1999	2001	2003	2006	2009	2012
Tamaño característico (micras)	0.25	0.18	0.15	0.13	0.1	0.07	0.05
Voltaje de alimentación (V)	1.8-2.5	1.5-1.8	1.2-1.5	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6
Transistores por chip (M)	11	21	40	76	200	520	1,400
Bits DRAM por chip (M)	167	1,070	1,700	4,290	17,200	68,700	275,000
Tamaño del dado (mm2)	300	340	385	430	520	620	750
Dimensión máxima del chip (mm)	17.3	18.4	19.6	20.7	22.8	24.9	27.4
Frecuencia de reloj local (MHz)	750	1,250	1,500	2,100	3,500	6,000	10,000
Frecuencia de reloj global (MHz)	750	1,200	1,400	1,600	2,000	2,500	3,000
Máxima potencia por chip (W)	70	90	110	130	160	170	175

Tabla 3: Resumen de la predicción de la SIA para procesadores de gama alta.

Tendencias y problemas actuales

Una vez presentados los avances en tecnología y arquitectura a lo largo de la evolución de los computadores y algunas de las últimas propuestas para computadores de alto rendimiento, completamos el análisis del área de conocimiento presentando los retos que deben afrontarse actualmente y las tendencias en el desarrollo de los distintos factores que pueden influir en el futuro de arquitectura y tecnología. Al introducir el concepto de arquitectura de computadores establecimos su dependencia de la tecnología, por un lado, y de las aplicaciones, por el otro. Por eso examinaremos el estado de estos dos factores. En la tecnología distinguiremos entre la tecnología básica usada por los procesadores y memorias, la de semiconductores, y la que usan los sistemas de almacenamiento secundario.

Avances y limitaciones en la tecnología de semiconductores

La tecnología de semiconductores ha realizado constantes progresos desde la invención de los circuitos integrados. En 1965 G.E. Moore predijo que el número de transistores en un circuito integrado se duplicaría anualmente. En 1975, cuando la tecnología de semiconductores alcanzada la adolescencia, Moore revisó su predicción a la baja duplicando el número de transistores cada

18 meses y la tendencia se ha mantenido así durante 20 años, siendo denominada ley de Moore [Gepp98a] (Figura 28).

Todo el mundo discute si se seguirá cumpliendo o no, y durante cuánto tiempo. En 1994 la Asociación de Industrias de Semiconductores (SIA) realizó una predicción sobre la evolución de los distintos aspectos de la tecnología. La predicción tuvo que ser revisada al alza en muchos factores en 1997 y además la introducción de nuevas tecnologías se ha acelerado de 3 años a 2 [Gepp98b]. En la tabla 1.6 aparecen las principales predicciones de la SIA para procesadores de gama alta [FlHR99].

Sin embargo, existe una triple amenaza para la ley de Moore: costes de fabricación, complejidad del diseño y test de circuitos tan complejos en tiempo razonable y límites tecnológicos, tanto de retardo de interconexiones como de consumo y disipación de potencia [FlHR99]. A continuación revisaremos los principales aspectos de cada uno de estos problemas.

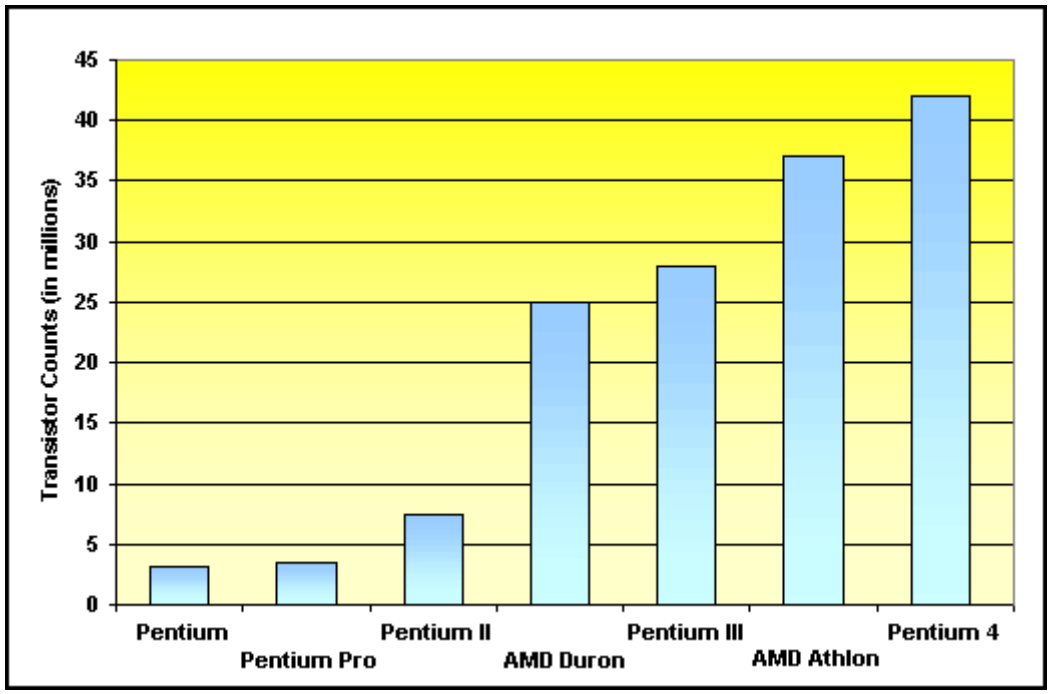

Figura 28: Ley de Moore

Costes de fabricación

La segunda ley de Moore dice que el coste de construir una fábrica de semiconductores se duplica cada 3 o 4 años. Es una opinión generalizada que serán los costes de fabricación y no los retos tecnológicos los que limiten la mejora de los CIs [ChDo98]. Algunos predicen que a partir del 2005 ya no será rentable disminuir el tamaño de los transistores y que dejará de cumplirse la primera ley de Moore.

Hasta ahora ésta se ha mantenido porque las mejoras exponenciales de tecnología y productividad han creado un aumento del mercado de circuitos integrados, aumento que a su vez ha producido grandes ganancias. Estas han permitido invertir mucho dinero en investigación y en fábricas de elevadísimo coste, como por ejemplo 2.000 millones de dólares el de una fábrica para chips de Pentium Pro [Yu96].

El coste creciente sólo se puede mantener si el volumen de ventas es enorme, así que es necesario abrirse a nuevos mercados. Esto puede imponer nuevos requisitos al diseño, ya que para el consumidor medio son importantes la fiabilidad y la facilidad de uso. Según Hennessy [Henn99] estamos en el umbral de una nueva era en la que todo el mundo usará servicios de información y aparatos basados en computadores. Cuando el usuario medio empiece a utilizar estos sistemas esperará que funcionen y sean fáciles de usar. Otra alternativa es que se simplifiquen los procesos y los equipos de fabricación de forma que el coste de las fábricas no siga creciendo.

Límites en la tecnología de semiconductores

Los avances en tecnología de semiconductores aumentan tanto la velocidad como el número de transistores que pueden incluirse en un chip. Cada generación disminuye un 30% cada dimensión -lateral y vertical- de los transistores y su retardo [Bork99] [FlHR99].

Esto debería provocar una mejora del ciclo de reloj de un 43%. Sin embargo ha mejorado un 50% debido a que se disminuye el número de puertas por ciclo

de reloj, aumentando la profundidad de segmentación [Matz97]. Por ejemplo, el UltraSPARC-III de Sun está segmentado en 14 etapas, de aproximadamente 8 retardos de puertas lógicas cada una [HoLa99].

La densidad de transistores debería duplicarse cada generación y los chips de memoria lo han cumplido. En cambio, los chips de lógica no alcanzan la densidad máxima por la complejidad de las microarquitecturas, que como veremos más adelante necesita herramientas CAD mejores.

Actualmente se están utilizando tecnologías de 0.25 micras y de 0.18 micras. Parece que la tendencia es a llegar a las 0.1 micras en el 2006, acercando los sistemas de litografía ópticos actuales a los límites impuestos por la física. Para tamaños por debajo de 0.1 son necesarias nuevas herramientas litográficas, que pueden estar basadas en rayos X, haces de electrones o de iones, de las cuales hay prototipos en desarrollo [Gepp98b].

El factor que más limita la mejora del rendimiento hoy en día son las conexiones. También se necesitan avances en el área de consumo de potencia y aquí el diseño asíncrono u opciones como la del procesador Crusoe pueden ser una vía de solución..

Las conexiones son uno de los grandes límites para conseguir mejorar el rendimiento porque disminuir su tamaño aumenta la resistencia y/o la capacidad. El número de capas de metal para interconexiones se ha incrementado de 2 a 6 y seguirá incrementándose a medida que se necesiten más interconexiones entre los dispositivos.

Según [StCo91], frecuencias de reloj superiores a 1 GHz (ciclo de reloj de 1 ns) pueden ser un límite absoluto para las interconexiones metálicas entre tarjetas de chips o módulos multichip. En general, a velocidades de 400 MHz y superiores, las conexiones deben ser punto a punto.

Será necesario encontrar nuevos materiales con menos resistencia y menos capacidad para poder sobrepasar los límites dados por las interconexiones. Las

conexiones de aluminio tienen alta resistencia y son vulnerables a electromigración, problemas que aumentan al disminuir la anchura de las líneas. El cobre era una alternativa mejor pero había dificultades para su fabricación. En 1997 IBM y Motorola anunciaron procesos de fabricación con 6 niveles de interconexiones de cobre, que mejora rendimiento, consumo, densidad y coste, pero sobre todo fiabilidad [Gepp98b].

Además se combinan conexiones de cobre con aislantes de baja constante dieléctrica para reducir la capacidad de los cables y las interferencias. Un año después del desarrollo de la tecnología de interconexiones de cobre IBM ya la está usando para el PowerPC 750 a 500 MHz [Beck93], un PowerPC empotrado (consumen menos y son más rápidas, así que son muy convenientes para empotrados). En Mayo IBM anunció su nuevo servidor S/390 G6 con interconexiones de cobre [IBM99a]. El chip, que incluye la cache de nivel 2, tiene, gracias a la nueva tecnología, el doble de transistores que el del S/390 G5 y ocupa un 10% menos de área (121 millones de transistores, que ofrecen 16 Mbytes de memoria, en un chip de 16.5 mm). También lo incorporará a las familias de servidores RS/6000 y AS/400 [Gepp99].

A pesar de las mejoras en las conexiones, el retardo de las mismas y el aumento de la frecuencia de reloj implican que las arquitecturas grandes necesitan ser modulares y que la ubicación sea adecuada para evitar la presencia de cables largos. Ya vimos ejemplos de ello, en el UltraSPARC-III (la señal de parada global del pipe tendría mucho retardo, así que se elimina) y, sobre todo, en el Alpha 21264: todas las conexiones externas al chip son canales punto a punto de alta velocidad y el banco de registros se divide en dos para que su tamaño sea menor y se pueda acceder en un ciclo. En este sentido, en la predicción de SIA es muy llamativa la diferencia entre el crecimiento de la frecuencia local del reloj y la global.

El consumo también es un factor que limita el aumento del rendimiento tanto para microprocesadores empotrados, como para los microprocesadores

de alto rendimiento, más rápidos, que necesitan más potencia y necesitan disiparla. Esto fuerza a alcanzar compromisos entre tamaño del dado (que el tamaño del dado no crezca mucho), la alimentación (disminuir el voltaje de alimentación) y la frecuencia, tal como se aprecia en la tabla 1.6. Disminuir la alimentación aumenta la susceptibilidad a errores blandos (debido a que la energía necesaria para cambiar el estado de un biestable es cada vez menor) y precisa de técnicas de refrigeración potentes [Bork99].

También se investiga en el uso de nuevos materiales para aumentar rendimiento y disminuir consumo [Gepp99] como por ejemplo en silicio-germanio.

Por otra parte surge el problema del empaquetamiento [Slat96]. A medida que los diseñadores colocan más funciones en un chip, éstos necesitan más patillas de entrada/salida. Las tecnologías actuales que proporcionan circuitos a precios aceptables, no permiten empaquetar con más de 200 patillas. Las que permiten mayor número elevan considerablemente el coste de los CI. Por esta razón, se necesitan nuevas tecnologías de empaquetamiento que permitan empaquetar grandes cantidades de patillas a costes bajos.

Complejidad de diseño y test

La complejidad del diseño y el tamaño del equipo de diseño se han convertido en una de las barreras más importantes al avance de la tecnología [Yu96]. Un ejemplo que muestra el crecimiento de ambos factores es el diseño de dos microprocesadores MIPS. En 1985 fue terminado el MIPS R2000 tras 15 meses de diseño. Tenía 0.1 millones de transistores. El equipo de diseño lo formaban 20 personas y la verificación constituía un 15 % del coste total. El MIPS R10000 terminado en 1996 tenía 6.8 millones de transistores. El equipo de diseño lo formaban más de 100 personas, que tardaron 3 años. La verificación costó más del 35% del total [Henn99]. De este ejemplo se deduce que el tiempo de diseño se ha duplicado y el tamaño del equipo se ha quintuplicado.

La validación y test de los microprocesadores actuales cada vez ocupan más parte del esfuerzo de diseño. Actualmente consumen del 40 al 50% del coste de diseño de un chip de Intel y el 6% de los transistores en el Pentium Pro. Por un lado los equipos de test son más caros debido al mayor número de pines y frecuencia de reloj. Y por otro, el tiempo de test aumenta constantemente debido a la complejidad de los chips y a los requisitos de calidad [Yu96].

Además, para diseños con tecnologías de tamaños característicos tan pequeños como los actuales es muy importante analizar el comportamiento temporal y el consumo, ya que la ubicación y conexionado pueden afectarles mucho. Esto presenta nuevos retos en el campo del test.

Otras mejoras necesarias tienen que ver con la integración entre los distintos niveles del proceso de diseño. Al planear un nuevo microprocesador de alto rendimiento los diseñadores tienen que tomar muchas decisiones, que incluyen organización superescalar, lanzamiento de instrucciones en desorden, ejecución especulativa, predicción de saltos y jerarquía de caches. La interacción entre las distintas características de la microarquitectura frecuentemente es contra-intuitiva y se formulan preguntas sobre las ventajas potenciales de rendimiento [MoWM99]. Los compromisos de diseño complejos requieren modelado del rendimiento preciso y a tiempo [LiSh97].

Este modelado debe realizarse en varios niveles de abstracción para que sea preciso y rápido [BoCA99]. Al empezar el proceso de diseño se usaban habitualmente modelos de rendimiento: en este nivel de abstracción se quiere definir la mejor microarquitectura que implementa una arquitectura dada, y "mejor" quiere decir la que produce mayor rendimiento en términos de CPI. Estamos siendo testigos de la creciente necesidad de tener en cuenta más y más ligaduras de bajo nivel en el modelado y análisis de alto nivel (en fases tempranas de diseño). Esto es debido al aumento de integración de los circuitos: los diseños actuales usan muchos millones de transistores que operan a frecuencias cercanas al GHz. A esa velocidad los retardos de interconexiones

y cables determinan significativamente el ciclo de reloj, así que la partición y ubicación de los bloques lógicos es un tema que debe tratarse con cuidado en los niveles más altos de diseño para evitar sorpresas posteriores (como vimos en el apartado anterior son necesarias decisiones de diseño de alto nivel para evitar retardos de conexiones demasiado largos).

Por lo tanto se necesita más integración entre las metodologías de modelado y validación en distintos niveles. Para arquitecturas VLIW se debería incluir el compilador dentro del modelo, pero de momento no se ha hecho debido a que es muy complejo.

Por otro lado, los diseñadores de aplicaciones empotradas no están tan interesados en el aumento constante de rendimiento sino en integrar más funciones en el chip. Al evolucionar hacia el "sistema en un chip" los microprocesadores empotrados se especializan, ya que diferentes aplicaciones necesitan diferente memoria, controladores de periféricos e interfaces. Esto aumenta la demanda de microprocesadores que puedan ser un bloque de un ASIC. Varias compañías ofrecen cores de microprocesadores y deben proporcionar también otros bloques complejos y herramientas para diseñar, depurar, verificar y testear los chips.

Para que estos componentes complejos (Intelectual Property) puedan ser utilizados en otros diseños es necesario que sean diseñados pensando en su reuso, tanto si son hard-IP (componentes ya ubicados, conectados y verificados) como si son soft-IP (descripción RT sintetizable).

Además para que sea posible el diseño de "sistemas en un chip" son precisas mejoras en las herramientas para el co-diseño hardware y software, verificación formal y optimizaciones a nivel RT y la integración más estrecha de diseños físicos y lógicos [Mart99]. Hace 10 años el cuello de botella de la tecnología eran las técnicas de fabricación pero hoy es el problema de diseñar chips grandes y muy densos con componentes muy pequeños.

Limitaciones de los sistemas de almacenamiento

Los computadores hoy en día tienen que almacenar todas las formas de información: archivos, documentos, imágenes, sonidos, vídeos, datos científicos y otros tantos nuevos formatos de datos. Se han realizado grandes avances técnicos para capturar, almacenar, analizar y visualizar datos [Gray96]. La sociedad cada vez necesita más y más información, sobre todo con el surgimiento de las aplicaciones multimedia y el acceso a datos a través de navegadores, como veremos más adelante.

En los últimos años se ha conseguido mejorar la capacidad y el coste de los sistemas de almacenamiento enormemente, pero la velocidad no se ha incrementado en la misma proporción. Por ejemplo, la velocidad de acceso a los discos (en la Figura 1.6, en la pág. 28, puede observarse la disminución del coste), se ha incrementado en menos de un factor 2, la de las cintas en 3, mientras que la de la CPU lo ha hecho en varios órdenes de magnitud.

Esto es una limitación grave en el incremento del rendimiento global de los computadores. De acuerdo con la ley de Amdahl, el incremento global en el rendimiento de un sistema depende del incremento del rendimiento de cada una de sus partes y del tiempo que se utilizan éstas. Por tanto, si sólo se mejora el rendimiento de la CPU, no se produce un incremento proporcional en el rendimiento global. Y puede deducirse que ignorar la velocidad de almacenamiento de datos, conduce a mayor pérdida de rendimiento a medida que la CPU se hace más rápida.

Por esta razón, uno de los objetivos tecnológicos actuales es conseguir sistemas de almacenamiento masivo con menores tiempos de acceso. Una posible solución son los SSD (Solid State Disks) implementados con DRAMs y una batería para que sean no volátiles. Pero el problema es el coste, que es al menos 50 veces el coste de los discos magnéticos. Esto seguramente conducirá a sistemas con memorias DRAM masivas en el futuro.

TENDENCIAS EN LAS APLICACIONES

Las tendencias en el uso de los computadores son muchos más difíciles de prever que los desarrollos de la tecnología, porque influyen muchos factores sociales. Por un lado, debemos tener en cuenta la evolución del software del sistema y por otro la de las aplicaciones típicas. Ya hemos mencionado que para extraer más paralelismo de los programas es necesario que los compiladores sean capaces de extraerlo automáticamente. En lo que respecta a la carga de trabajo veremos a continuación distintos aspectos que influyen sobre ella, aunque su evolución es con frecuencia sorprendente.

En primer lugar Internet y WWW tienen una fuerte influencia en la sociedad actual. Actualmente la información está centralizada en Internet, así que muchas personas que no trabajan en campos tradicionalmente relacionados con los computadores tienen que usarlos para acceder a bases de datos, bibliotecas, etc. Esto implica también que se acentúa la tendencia a que un computador no sea básicamente un dispositivo de cálculo sino uno de comunicación y, como veremos más adelante, a que el usuario no sea experto.

Dicha comunicación, aspecto crucial en la sociedad de la información en que vivimos, se realiza fundamentalmente mediante imágenes y sonido. Por tanto, los datos multimedia pasan a reemplazar en muchos casos a los aritméticos tradicionales.

En general, cada vez más capacidad de procesamiento se usa para comunicaciones e interfaz de usuario y tienen mayor importancia las cargas de trabajo multimedia.

Además, la gran importancia de Internet está cambiando la forma en la que se ejecutan las aplicaciones, de modo que el modelo cliente-servidor se utiliza para acceder a un creciente número de servicios. Esto influye en el tipo de computadores que son necesarios para ejecutarlas, y por tanto en la estructura

del mercado de computadores. Como mencionamos al hablar de la evolución de los mismos, actualmente se transforma todo en clientes y servidores.

Un tercer aspecto importante es la integración de los computadores en la sociedad, que lleva a que estos sean omnipresentes. Por un lado, esto implica que la informática cada vez se relaciona más con otros campos (entretenimiento, electrónica de consumo) y, por otro, que cada vez los usuarios son menos expertos. Cambian por tanto el uso que se les da (comunicación, procesamiento de textos, hojas de cálculo, etc) y las prioridades de diseño (facilidad de uso a través de un interfaz agradable, precio, fiabilidad). En el mercado de aplicaciones portátiles, sistemas empotrados y electrónica de consumo otros objetivos son integración y bajo consumo.

A continuación veremos con mayor detalle como pueden afectar estas tendencias en el uso de aplicaciones a la evolución de la arquitectura de computadores.

Influencia de Internet

El desarrollo enorme de las redes globales, utilizadas actualmente como un medio más de comunicación, y la facilidad que proporcionan para obtener información de cualquier lugar del mundo, ha dado lugar a otro tipo de función imprescindible en un segmento importante de computadores: la conexión a Internet. El computador es un medio para poder conectarse a la red y navegar de un sitio a otro en busca de información. Esto, a su vez, ha generado un nuevo tipo de necesidades hardware y software en los computadores, orientadas fundamentalmente a la comunicación por red y visualización de los datos.

Internet y especialmente WWW ofrecen enormes cantidades de información y por ello aumenta el valor de los dispositivos de computación para el consumidor medio. Según Slater, si la red alcanza su pleno potencial, dispositivos de acceso a la red de bajo coste y fácil uso pueden romper la

necesidad de compatibilidad con la arquitectura x86: se ejecuta el navegador en cualquier arquitectura y las aplicaciones descargadas a través de la red pueden ser independientes del procesador si están escritas en Java [Slat96].

Sin embargo, esta tendencia a usar dispositivos de prestaciones reducidas para el acceso a Internet (algunos de los cuáles, como el computador de red, citaremos en el siguiente apartado) se ve compensada por el aumento de la información multimedia y en 3D, que, como veremos más adelante exige gran ancho de banda de comunicaciones.

Además de este uso, Internet también ofrece nuevas posibilidades de computación para realizar proyectos complejos en colaboración. El modelo cliente-servidor se une a la creciente complejidad de los proyectos, como por ejemplo el diseño de microprocesadores y otros circuitos integrados. Esto ha llevado, gracias a la implantación de redes locales de alta velocidad, al éxito de las redes internas ("intranets") de empresa, que permiten el trabajo en equipo. Dado que el computador de red requiere mucho ancho de banda en la red, es probable que sólo sustituya a terminales en intranets.

Evolución de los segmentos de productos

Existen básicamente dos factores que están cambiando la forma de ejecutar aplicaciones en relación con Internet. La primera es la ya mencionada sobre el uso del modelo cliente-servidor en el acceso a los servicios que Internet ofrece. De acuerdo con ello el mercado de microprocesadores se dividirá en clientes y servidores [FlHR99]. Se añade un nuevo segmento al mercado de computadores formados por dispositivos de comunicación de propósito específico y/o reducidas prestaciones, que veremos a continuación.

El "network computer", que se puede traducir como "computador en red", está desarrollando actualmente un concepto arquitectónico diferente del computador o la estación de trabajo. Sus características fundamentales son la falta de sistema de almacenamiento y el disponer exclusivamente del software

necesario para realizar la conexión con Internet, desde los protocolos TCP/IP y FTP, hasta los estándares de la WWW, como HTML, HTTP y Java. El resto, incluyendo cualquier aplicación o datos, puede tomarse de la red.

Según Hennessy, estamos en la era "post-PC" [Henn99], en la que una parte importante son los "aparatos de información" (information appliances) para conectarse a Internet y utilizar servicios. Deben cumplir 2 requisitos: que sean baratos y fáciles de usar.

También se incluyen en este primer segmento los microprocesadores empotrados para otras aplicaciones. Las ventas de microprocesadores empotrados de 32 y 64 bits ya son casi el doble que las de PC de sobremesa y para el 2003 serán más del triple.

Dentro de este segmento serán importantes la disponibilidad de aparatos y servicios, la facilidad de mantenimiento, el coste y el rendimiento de comunicaciones.

Los clientes (esto incluye PCs, computadores de red y procesadores empotrados para electrónica de consumo), deben ser baratos y consumir poco, aunque aumentar el rendimiento es inevitable debido a la competencia. Cuando el coste lo permita incorporarán funciones de memoria y procesamiento de señal.

Los servidores tienen distinto objetivo de diseño: el rendimiento. Es importante usar el espacio de manera eficiente. El tiempo de ciclo limita la complejidad del chip. Es necesario desacoplar regiones del dado y para ello se propone integrar varios procesadores en un dado, o usar redundancia para aumentar la fiabilidad.

El segundo factor afecta fundamentalmente a los servidores: el comercio electrónico se construye sobre una arquitectura de 3 niveles que es escalable: en el primer nivel están los PCs, "clientes finos" (thin clients) y otras "aplicaciones de información" que, mediante el uso de un navegador, permiten

acceso a Internet. El segundo nivel lo constituyen los servidores Unix y NT que ejecutan el software de servidor y HTTP. Este nivel se conecta al tercero, que contiene las bases de datos. Estas están almacenadas en mainframes porque es donde estaban originalmente y costaría mucho convertirlas a otros formatos.

Como vimos anteriormente, hasta 1995 se decía que los mainframes no tenían futuro por la mejora de rendimiento de los PCs (Figura 1.9). Internet puede cambiar esto porque las bases de datos que contienen toda la información residen en mainframes. Lewis [Lewi99] opina que una de las principales diferencias entre los mainframes y los PCs y servidores es que los primeros tienen alta fiabilidad debida al uso de redundancia. Los requisitos de computación que Internet exige en fiabilidad, seguridad y capacidad de procesamiento no pueden ser cubiertos por otros.

Entre 1995 y 1997 las ventas de IBM S/390 se triplicaron. Los costes disminuyen un 32 % anual. La sexta generación S/390 G6 tiene un rendimiento mayor en un 50% al del G5. Se pueden conectar hasta 12 CPUs para conseguir 1,600 MIPS. Tiene 32 Gbytes de memoria, que pueden ser ampliados a cientos de terabytes.

Multimedia y sus implicaciones

Nuestra sociedad está orientada hacia la información, con demanda multimedia en expansión, y por ello depende de dispositivos de comunicación y computación móviles basados en una infraestructura de red [Naka99]. Para ser eficaces, estos dispositivos deben consumir muy poco. Los procesadores empotrados que se fabrican actualmente incluyen varios componentes (memoria, procesadores de señal).

La tendencia general es a crear sistemas que incorporen entrada/salida con vídeo y audio y reconocedores de voz [Bell96]. Por eso actualmente es necesario incorporar varios millones de transistores en un circuito para el interfaz humano y las funciones gráficas.

Si el procesamiento de datos multimedia dinámicos (video, animación, música, etc) se integra en las aplicaciones, la carga media cambiará significativamente, influyendo en el diseño de procesadores. Las características de las aplicaciones son [DiDu97]:

1) Procesamiento en tiempo real.

2) Tipos de datos continuos: los datos de entrada suelen provenir del muestreo en el tiempo de una señal analógica representable, para lo que bastan 8 bits, en lugar de los 32 o 64 usados para representación de enteros. También hay datos en punto flotante.

3) Paralelismo de grano fino en los datos: está inherente en las aplicaciones gráficas y de procesamiento de señal, que consisten en repetir el mismo procesamiento a cada elemento del flujo de entrada. Esto sugiere usar sencillas unidades de ejecución SIMD en lugar del complejo paradigma necesario para extraer paralelismo de las dependencias de datos y control entre instrucciones en superescalares.

4) Paralelismo de grano grueso: las aplicaciones contienen en su mayoría varias "hebras" con tiempo de ejecución crítico (para video conferencias es necesario codificar y decodificar video y audio y otras tareas subordinadas). Estas hebras de ejecución independientes permiten aplicar hardware con paralelismo espacial y temporal: procesadores con muy alta frecuencia, segmentación profunda y hardware multi-hebra, así como multiprocesadores simétricos.

5) Localidad de referencias a instrucciones en bucles pequeños: estos bucles consumen la mayor parte del tiempo de procesamiento, por lo que existe mucha correlación entre la aceleración del bucles y la de la aplicación

6) Gran ancho de banda con memoria: los conjuntos de datos son enormes y no pueden ser manejados por las caches, que empeoran aún más debido a la falta de localidad en los datos.

Para tratar con estos datos hay dos alternativas: procesadores de procesamiento de señal de propósito específico y dotar a los de propósito general de características que aceleren el procesamiento de multimedia. Se puede conseguir bastante mejora con soporte arquitectónico relativamente sencillo. Inicialmente se propusieron extensiones para procesamiento en dos dimensiones (MMX de Intel, VIS para Sun Sparc, MDMX para Silicon Graphics MIPS V, MVI para Digital Alpha, MAX2 para HP-PA) y posteriormente se añadieron para gráficos en 3D [Digi92].

También se pueden utilizar un tipo de procesadores especializados en estas tareas: las arquitecturas DSP (Procesamiento de Señales Digitales) proporcionan múltiples caminos de datos, instrucciones de longitud larga (VLIW) y otras características especiales que los hacen más rápidos, pero a menudo difíciles de programar.

Por otra parte, la anchura de banda necesaria entre la CPU y la memoria es cada vez mayor debido a las mayores funciones de comunicación. Actualmente hay periféricos muy rápidos que requieren grandes intercambios de datos con el procesador, como discos magnéticos (5,000 KB/s), redes de área local (1,000 KB/s) y dispositivos gráficos (30,000 KB/s). Los sistemas de buses actuales tienen que incrementar su rendimiento, para que la transferencia no sea un cuello de botella, que incremente aún más el problema de acceso a memoria se usan buses organizados jerárquicamente: un bus local que conecta la CPU con la memoria cache externa, un bus del sistema que se usa para conectar la cache con la principal y uno o varios buses de expansión o de E/S para conectar distintos periféricos (Figura 1.13).

[Acc02] Accenture, Varios autores. "Expectativas de los estudiantes universitarios ante su inserción laboral". http://careers.accenture.com/location/spain/insercionlaboral.pdf. 2002.

[Acm68] Curriculum committee on computer science, Curriculum 68: Recommendations for the undergraduate program in computer science, Communications of the ACM, 11(3) marzo 68, pp151-197

[Acm79] Curriculum committee on computer science, Curriculum 78: Recommendations for undergraduate program in computer science, Communications of the ACM, 22, (3) marzo79,pp 147-166

[Alex84] N. Alexandridis. Microprocessor System Design Concepts. Computer Science Press, 1984.

[Alv74] Luis Alves Mattos. "Compendio de didáctica general". Ed. Kapelusz. 1974

[AmBB64] G. Amdahl. G. Blaauw, F. Brooks. Architecture of the IBM System/360, IBM Journal Research and Development, V 8, pp 8-21, 1964.

[AMD99b] AMD, http://www.amd.com/products/cpg/athlon/.

[Andr82] M. Andrews. Programming Microprocessor Interfaces for Control and Instrumentation. Prentice-Hall, 1982.

[Ang96] J.M.Angulo. Estructura de computadores. Paraninfo.1996.

[Baer80] J. Baer. Computer System Architecture. Computer Science Press, 1980.

[Baer84] J. Baer. Computer Architecture, Computer, V 17, n.10, pp 77-87, 1984.

[Barn68] G. Barnes. The ILLIAC IV Computer, IEEE Trans on Computers, V 17, n.8, pp 746-757, 1968.

[Beck93] M. Becker. The PowerPC 601 Microprocessor, IEEE Micro, Octubre 1993.

[Bell96] G. Bell, "The system-on-a-chip, microsystems computer industry", IEEE Micro, Diciembre 1996, pp. 52.

[Bhan96] D. Bhandarkar. Alpha Implementations and Architecture. Digital Press, 1995.

[BhCl91] D. Bhandarkar, D. Clark. Performance from Architecture: Comparing a RISC and a CISC with Similar Hardware Organization, Proc. of Fourth Conf. on Architectural Support for Programming Languages and Operating Systems, IEEE/ACM, pp. 310-319, 1991.

[Blas90] M. de Blasi. Computer Architecture. Addison-Wesley, 1990.

[BoCA99] P. Bose, T.M. Conte, T.M Austin, "Challenges in processor modeling and validation", IEEE Micro, Mayo-Junio 1999, pp. 9-14.

[Bork99] S. Borkar, "Design challenges of technology scaling", IEEE Micro, Julio-Agosto 1999, pp. 23-29.

[Brey95] B. Brey. The Intel 32-Bit Microprocessors: 80386, 80486 and Pentium. Prentice Hall, 1995.

[BuGo97] D. Burger, J.R. Goodman, "Billion-transistor architectures", IEEE Computer, Septiembre 1997, pp. 46-48.

[CaBM96] B. Carlson, A. Burgess, C. Miller, "Timeline of computing history", IEEE Computer, Octubre 1996, pp. 25-110.

[Car00] Marcelo Careaga Butter. "Fundamentos para un modelo cibernético de educación". Universidad de Concepción. Dirección de Docencia. 2000.

[Cat90] A.Cattania. 80386 Arquitectura y programación. Grupo editorial Jackson. 1990.

[Cata91] B. Catanzaro (ed.). The SPARC Technical Papers. Springer-Verlag, 1991.

[Cava84] J. Cavanagh. Digital Computer Arithmetic: Design and Implementation. McGraw-Hill, 1984.

[CC2001] IEEE/ACM. The joint task force on computing curricula. Computing curricula 2001. Computer Science. Final Report, 15-dec-2001. www.ieee.org

[ChDo98] P.K. Chatterjee, R.R. Doering, "The future of microelectronics", Proc. of the IEEE, vol. 86, nº 1, Enero 1998, pp. 176-183.

[Che94] P.M.Chen, G.A.Lee, G.A.Gibson,R.H.Katz,D.A.Patterson,"RAID: high-performance, reliable secondary storage.ACM Computing Surveys 26:2 June 1994,pp, 145-188

[CLGK94] P.M. Chen, G.A. Lee, G.A. Gibson, R.H. Katz, D.A. Patterson, "RAID: high-performance, reliable secondary storage", ACM Computing Surveys 26:2, Junio 1994, pp. 145-188.

[Coel02] C.A. Coello. Breve Historia de la Computación y sus Pioneros. Volumen I : Los Orígenes del Hardware, Fondo de Cultura Económica, 2002.

[CrGe87] J. Crawford, P. Gelsinger. Programación del 80386/387. Anaya, 1987.

[Dani96] R.G. Daniels, "A Participant´s Perspective", IEEE Micro, Diciembre 1996, pp. 21-31.

[deMi90] P. de Miguel, "Fundamentos de los computadores", Paraninfo, 1990.

[DeMi91] P. De Miguel. Arquitectura de Computadores. Fundamentos e Introducción al Paralelismo. Paraninfo, 1991.

[DGMP90] P. De Miguel, M.I. García, M. Martínez, F. Pérez. Problemas de Estructura de Computadores. Paraninfo, 1990.

[DiDu97] K. Diefendorff, P.K. Dubey, "How multimedia workloads will change processor design", IEEE Computer, Septiembre 1997, pp. 43-45.

[Dief99] K. Diefendorff, "Power4 focuses on memory bandwith", Microprocessor Report vol. 13, nº 13, Octubre 1999.

[Digi92] E. C. Digital. Alpha Architecture Handbook. Digital Equipment Corporation, 1992.

[Dutt99] A. Dutta-Roy, "Computers", IEEE Spectrum, Enero 1999, pp. 46-51.

[ErLa85] M. D. Ercegovac, T. Lang, Digital Systems and hardware/firmware algorithms. John Wiley & Sons, 1985.

[Fagg96] F. Fagginn. The Microprocessor, IEEE Micro, Diciembre, 1996, pp. 7-9

[Fagg96b] F. Faggin, M. Hoff, S. Mazor, M. Shima, "The History of the 4004", IEEE Micro, Diciembre 1996, pp. 10-20.

[Fel94] J.M.Feldman, C.T.Retter. Computer architecture. A Designer text based on a generic RISC. 1994

[Fer85] Juan José Ferrero. "Teoría de la educación. Fenomenología del hecho educativo". Universidad de Deusto. Bilbao 1985.

[Fer98] Juan José Ferrero. "Teoría de la educación. Lecciones y lecturas". Universidad de Deusto. Bilbao 1998.

[FeSM96] M. Fernández, V. Sánchez, I. Martín. Tecnología de computadores. Síntesis, 1996.

[GeHP93] [Gei90] R.Geiger, P.E.Allen, N.Strader. VLSI Design techniques for analog and Digital circuits. McGraw-Hill. 1990

[Gepp98a] L. Geppert, "The media event: Moore´s Law mania", IEEE Spectrum, Enero 1998, pp. 20-21.

[Gepp98b] L. Geppert, "Solid state", IEEE Spectrum, Enero 1998, pp. 23-28.

[Gepp99] L. Geppert, "Solid state", IEEE Spectrum, Enero 1999, pp. 52-56.

[Gil97] Rafael Gil Colomer (Director). "Filosofía de la educación hoy. Diccionario filosófico Pedagógico. Ed. Dykinson. Madrid 1997.

[GiMi87] Ch. Gimarc, V. Milutinovic. A Survey of RISC Processors and Computers of the Mid-1980s, Computer, V 20, n.9, pp. 59-85, 1987.

[Goor99] A.J. van de Goor. Computer Architecture and Design. Addison Wesley, 1999.

[Gray96] J. Gray, "Evolution of Data Management", IEEE Computer, Octubre 1996, pp. 38-46.

[Gwen96] L. Gwennap, "Digital 21264 Sets New Standard", Microprocesor Report vol 10, nº 14, Octubre 1996.

[Ham96] V.C.Hamacher, Z.G.Vranesic, S.G.Zaky. Computer Organization. McGraw-Hill. 4ª Edición.1996

[HaVZ02] V. Hamacher, Z. Vranesic, S. Zaky. Computer Organization. 4ª edición. McGraw-Hill, 1996.

[Hay87] J.P.Hayes. Diseño de sistemas digitales y microprocesadores. McGraw-Hill. 1987.

[Haye88] J. Hayes. Computer Architecture and Organization. McGraw-Hill, 1988.

[Hein93] J. Heinrich. MIPS R4000 User's Manual. Prentice Hall, 1993.

[Henn96] J.L. Hennesy, "RISC microprocessors", IEEE Micro, Diciembre 1996, p. 27.

[Henn99] J. Hennessy, "The future of systems research", IEEE Computer, Agosto 1999, pp. 27-33.

[HePa02] J.L. Hennesy, D.A. Patterson, "Computer architecture. A quantitative approach", Morgan Kaufmann, 3ª ed., 2002.

[HoLa99] T. Horel, G. Lauterbach, "UltraSPARC-III: Designing Third-Generation 64-bit performance", IEEE Micro, Mayo-Junio 1999, pp. 73-85.

[Holt88] W. Holt (ed.). Beyond RISC!: An Essential Guide to Helwett-Packard Precision Architecture. Helwett-Packard, 1988.

[Huss70] S. Husson. Microprogramming: Principles and Practices. Prentice Hall, 1970.

[Hwa79] K.Hwang. Computer arithmetic: Principles, architecture and design. John Wiley and Sons. 1979

[Hwa90] K.Hwang, F.A.Briggs. Arquitectura de computadores y procesamiento paralelo. McGraw-Hill. 1990.

[Hwan93] K. Hwang. Advanced Computer Architecture: Parallelism, Scalability and Programmability. McGraw Hill 1993.

[IBM94] IBM. The PowerPC Architecture: A Specification for a New Family of RISC Processors. Morgan Kaufmann, 1994.

[IEEE77] A Curriculum in Computer Science and Engineering, IEEE Press, EHO 119-8, 1977.

[IEEE83] The 1983 IEEE Computer Society Model Program in Computer Science and Engineering, IEEE Press, EHO 212-1, 1983.

[Ine02] Instituto Nacional de Estadística. "España en Cifras 2001". http://www.ine.es/espcif/espcifes/espcif01.htm. 2002.

[InSt02] In-Stat/mDR. "Microprocessor report online". 29 Abril 2002

[Inte99] Intel, http://www.intel.com/pentiumiii/xeon/.

[Kate85] M. Katevenis. Reduced Instruction Set Computer Architectures for VLSI. The MIT Press, 1985.

[Kham87] A. J. Khambata. Microprocessors / Microcomputers: Architecture, Software and Systems. 2ª edición. John Wiley and Sons, 1987.

[Kogg81] P. Kogge. The Architecture of Pipelined Computers. McGraw Hill, 1981.

[Kore02] I. Koren. Computer Arithmetic Algorithms. Prentice-Hall. 2002.

[LeEc89] H. Levy, R. Eckhouse. Computer Programming and Architecture: The VAX. 2ª edición. Digital Press, 1989.

[Lewi99] T. Lewis, “Mainframes are dead, long live mainframes”, IEEE Computer, pp. 102-104.

[LiSh97] M.K. Lipasti, J.P. Shen, “Superespeculative microarchitecture for beyond AD 2000”, IEEE Computer, Septiembre 1997, pp. 59-66.

[Liu91] Y. Liu. The MC68000 Microprocessor Family. Prentice-Hall, 1991.

[Liv93] P.E.Livadas, C.Ward. Computer Organization and the MC68000 Prentice hall International Editions. 1993

[Lor83] Konrad Lorenz, "La etología: Entrevista con Alain de Benoist". Ed. Nuevo Arte Thor. Barcelona 1983.

[Lor87] Konrad Lorenz, "Hablaba con las bestias, los peces y los pájaros". Ed. Labor, Barcelona 1987.

[Mart99] B. Martin, “Electronic design automation”, IEEE Spectrum, Enero 1999, pp. 57-61.

[Matz97] D. Matzke, “Will physical scalability sabotage performance gains?”, IEEE Computer, Septiembre 1997, pp. 37-39.

[Memo02] Memoria del curso 2000/2001 de la facultad de Informática. Universidad Complutense de Madrid.

[Milen94] M. Milenkovic. Sistemas Operativos. Conceptos y diseño. McGraw Hill, 1994.

[MoWM99] M. Moudgill, J.-D. Wellman, J.H. Moreno, “Environment for PowerPC microarchitecture exploration”, IEEE Micro, Mayo-Junio 1999, pp. 15-25.

[Naka99] T. Nakamura, “Introducing cool chips”, IEEE Micro, Julio-Agosto 1999, pp. 9-10.

[Nas85] Ricardo Nassif. "Teoría de la Educación. Problemática pedagógica contemporánea". Ed. Cincel. Bogotá 1985.

[NoHo81] R. Noyce, M. Hoff. A History of Microprocessor Development at Intel, IEEE Micro, V 1, n. 1, pp 8-21, 1981.

[Omon94] A. R. Omondi. Computer Arithmetic Systems. Prentice Hall, 1994.

[ONH+96] K.olukonut, B. Nayfeh, I. Hammond, K. Wilson, K-Y. Chang. "The case for a single chip multiprocessor". Proc. 7 th Int. Conf. On Architectural Support for Programming Languages and Operating Systems. (ASPLOS), 1996, pp 1 - 22.

[PaHe98] D. Patterson, J. Hennesy. Computer Organization & Design. The Hardware/Software Interface. Morgan Kaufmann, 1998.

[PaSe82] D. Patterson, C. Sequin. A VLSI RISC, Computer, V 15, n.9, pp 8-21, 1982.

[Patt98] D.A. Patterson, http://www.cs.berkeley.edu/~pattrsn/252S98/, transparencias de su curso "Computer Architecture" en la Universidad de Berkeley en primavera de 1998.

[Rob02] Daniel Robaldo. "Modelo Pedagógico humanista (del nuevo humanismo). http://www.nalejandría.com/akademia/humanista/. Febrero 2002.

[Sae86] Oscar Sáenz (Director). "Pedagogía General. Introducción a la teoría y práctica de la Educación". Ed. Anaya. Madrid .

[Sar00] Jaume Sarramona. "Teoría de la Educación. Reflexión y normativa pedagógica". Ed. Ariel. Barcelona 2000.

[SiBN82] D. Siewiorek, G. Bell, A. Newell. Computer Structures: Principles and Examples. McGraw Hill, 1982.

[SiFK97] D. Sima, T. Fountain, P.Kacsuk. "Advanced Computer Architectures. A Design Space Approach". Addison Weslwy, 1997.

[Slat96] M. Slater, "The Microprocessor Today", IEEE Micro, Diciembre, 1996, pp. 32-44.

[SMMO95] J. Septién, H. Mecha, R. Moreno, K. Olcoz. La Familia del MC68000. Conexión y programación de interfaces. Síntesis, 1995.

[Stal03] W. Stallings. Computer Organization & Architecture. 5ª edición. Prentice Hall, 2000.

[Stal97] W. Stallings, "Organización y arquitectura de computadores", Prentice Hall, 4ª ed., 1997.

[StCo91] Harold S. Stone, J. Cocke, "Computer Architecture in the 1990s", IEEE Computer, Septiembre 1991, pp. 30-38.

[Ston83] H. Stone. Microcomputer Interfacing. Addison-Wesley, 1983.

[Taba95] D. Tabak, Advanced Microprocessors. McGraw Hill, 1995.

[Tane86] A. Tanenbaum. Organización de Computadoras: Un Enfoque Estructurado. 2ª edición. Prentice Hall, 1986.

[Tane99] A. Tanenbaum. Structured Computer Organization. 4ª edición. Prentice Hall, 1999.

[Tome81] I. Tomek. Introduction to Computer Organization. Computer Science Press, 1981.

[TOP99] TOP500 Supercomputing Sites, http://www.top500.org/, Noviembre 1999.

[Tred96] N. Tredennick, "Microprocessor-Based Computers", IEEE Computer, Octubre 1996, pp. 27-37.

[Tuc91] A,B,Tucker. Cumputing Curricula 1991. Communications of the ACM, 34,(6) Junio 1991. Pp68-84

[Wilk51] M. Wilkes. The Best Way to Design an Automatic Calculating Machine, Manchester University Computer Inag. Conference, pp 16-18, 1951. Reimpreso en Computer Design Development Principal Papers, E Swartzlandr Ed, Hayden, 1976.

[Wilk53] M. Wilkes. Microprogramming and the Design of the Control Circuits in an Electronic Digital Computer, Proc. Cambridge Philosophical Society, V 49, pp 230-238, 1953. Reimpreso en [SiBN82]

[Wilk95] M. Wilkes. Computing Perspectives. Morgan Kaufmann, 1995.

[Wilk96] B. Wilkinson. Computer Architecture: Design and Performance. Prentice Hall, 2ª edición, 1996.

[Yu96] A. Yu. The Future of Microprocessors, IEEE Micro, pp. 46-53, Diciembre 1996.

[Yuc91] Yu-Cheng Liu, The M68000 Microprocessor Familiy. Prentice Hall. 1991

[Zima91] H. Zima, "Supercompilers for Parallel and Vector Computers", ACM Press, 1991.