Sistemas de Cómputo

Un sistema de cómputo es un conjunto de elementos electrónicos organizados para permitir el procesamiento de información. A lo largo del tiempo los sistemas de cómputo han evolucionado notablemente, y en este material vamos a ver las cosas más importantes que han ocurrido durante esta evolución.

Evolución de los Sistemas de Cómputo

Los sistemas de cómputo constituyen una industria, y como en toda industria, existe la competencia entre los fabricantes, que buscan obtener más mercado mejorando las características de sus productos. En ninguna otra industria como en la computación es tan notable el crecimiento de las capacidades de los sistemas de cómputo y, simultáneamente, la reducción de costos y tamaños físicos de los componentes.

- Las sucesivas generaciones de sistemas de cómputo han ido creciendo en la velocidad de procesamiento y capacidad de memoria, así como en las capacidades de los discos y otras unidades de almacenamiento.
- Al mismo tiempo, se ha reducido notablemente el tamaño del sistema como un todo. Esto último se ha logrado a través de sucesivos avances en la **integración** de los componentes.
- La mayor integración ha facilitado la **economía de escala**, manteniendo o reduciendo en general los costos de producción.

Sin embargo, los diferentes componentes se producen con procesos de fabricación diferentes y que incorporan avances cada uno a su ritmo, por lo cual no siempre más rápido o más pequeño es mejor, sino que lo más importante es que las partes del sistema funcionen en armonía. De lo contrario, pueden presentarse desbalances que impidan que el sistema funcione en forma óptima.

Antecedentes históricos

En la antigüedad se crearon los que podríamos llamar sistemas de cómputo primitivos, aunque ingeniosos. Aquí citamos unos pocos ejemplos, como el ábaco chino; los quipus incas; los dispositivos de cálculo analógicos aparecidos en diferentes civilizaciones, como los que permitían calcular la torsión de los tensores de las catapultas romanas; el misterioso Mecanismo de Anticitera, un aparato astronómico encontrado entre los restos de un naufragio griego.

Más cerca de nuestros días, se crearon artefactos parecidos a las calculadoras y computadoras actuales, pero con la tecnología disponible en esos momentos, lo que lógicamente los hacían incómodos, o pobres en resultados, en comparación con las herramientas electrónicas de hoy.

El Mecanismo de Anticitera

La Pascalina

La Tabuladora de Hollerith

Entre estos proto-sistemas de computación es especialmente notable, por varios motivos, la máquina de Hollerith. Herman Hollerith trabajaba para la Oficina de Inmigración de EEUU hacia fines del siglo XIX, en momentos en que se formó una gran corriente inmigratoria desde Europa. La gran cantidad de personas que llegaban a radicarse, nunca vista antes, hizo desbordar el sistema de información nacional. Los responsables del censo poblacional se encontraban con un gran problema.

Mucho antes de Hollerith, un tapicero francés, Jacquard, había ideado un telar que se configuraba usando tarjetas perforadas. Alimentado con estas tarjetas, el telar creaba automáticamente el dibujo deseado. Inspirado en el telar de Jacquard, Hollerith creó un sistema de cómputo automático basado en tarjetas perforadas. Cada tarjeta representaba a un individuo. La tarjeta se dividía en campos que representaban los atributos o características personales del individuo (nacionalidad, fecha de nacimiento, sexo, estado civil, etc). Al llegar un individuo, el oficial de inmigración le presentaba un cuestionario y codificaba sus respuestas con una perforación en cierto lugar de cada campo.

La Tabuladora de Hollerith era un dispositivo que contabilizaba perforaciones en esas tarjetas. Podía ser programada para contar la cantidad de individuos por nacionalidad, por edad, por sexo, etc., o por varios de estos atributos simultáneamente. De esa manera el censo nacional pudo lograrse en muchísimo menos tiempo que con los anteriores métodos manuales.

La máquina de Hollerith es especialmente interesante porque sienta las bases del cálculo digital como se conocerá en los años siguientes (de hecho, las tarjetas perforadas siguieron utilizándose hasta muchos años después como medio de entrada, para codificar programas y datos), porque demostró el poder del cómputo automático con una aplicación concreta e importante, y porque, tomando su invento como punto de partida, Hollerith formó una importante empresa de computación que tuvo gran influencia en el desarrollo de la tecnología del siglo XX.

Primera Generación

Las primeras computadoras electrónicas usaban **bulbos, tubos de vacío, o válvulas**, como interruptores, implementando dispositivos que realizaban operaciones aritméticas y lógicas.

Dado el momento histórico en el cual aparecieron estos equipos, los objetivos con los cuales se creaban eran, con frecuencia, los usos militares. Las máquinas

de esta generación eran grandes instalaciones que ocupaban una habitación, y sus miles de válvulas disipaban una gran cantidad de calor, que debía combatirse con sistemas de aire acondicionado.

El **ENIAC** es un claro representante de esta clase de máquinas. Pesaba 30 toneladas, y ocupaba un recinto de 140 m^2. Era capaz de ejecutar 5000 operaciones de suma por segundo. El ENIAC usaba 18000 válvulas de vacío: cada dos días, en promedio, una de ellas fallaba, y debía ser reemplazada con un procedimiento que llevaba quince minutos.

El ENIAC no era una máquina de Von Neumann porque su programa no residía en memoria, sino que la computadora se programaba con un intricado sistema de interruptores manuales. Entre las máquinas de esta generación se encuentra la primera computadora de programa almacenado según el modelo de Von Neumann. Fue el IAS (siglas de **Institute for Advanced Study**), que usaba 1500 tubos de vacío y tenía 5 kB de memoria.

El tubo de vacío

El **tubo de vacío o válvula termoiónica** fue patentado por Edison y fue sucesivamente modificado para diferentes usos en electrónica hasta llegar a ser usado en las computadoras de la primera generación. Una de sus variedades, el **triodo**, tiene tres electrodos o terminales conectados al resto del circuito, llamados **cátodo, ánodo y rejilla o grilla de control**. En éstos, la corriente eléctrica se dirige siempre desde el cátodo al ánodo, pero únicamente circula cuando existe una determinada carga negativa en la grilla, que funciona como un interruptor.

De esta manera se puede controlar el flujo de corriente por un circuito y construir dispositivos que implementen funciones lógicas. Así, dos válvulas de este tipo, conectadas en serie, simulan una función lógica de conjunción o **AND**; dos válvulas conectadas en paralelo, simulan una disyunción u **OR**, etc. Con válvulas termoiónicas es posible además crear un dispositivo que mantenga permanentemente un cierto estado eléctrico, y que por lo tanto **puede almacenar un bit de información**.

La grilla de las válvulas necesita alcanzar una alta temperatura para poder gobernar el flujo de electrones. De ahí que el consumo de electricidad fuera altísimo y su funcionamiento sumamente lento. Unido esto a una alta tasa de fallos, las válvulas fueron rápidamente abandonadas en favor de una tecnología más conveniente, el **transistor**.

Memoria de núcleos

Las primeras implementaciones de la memoria principal (memorias de núcleos o **core memories**) fueron realizadas con pequeños anillos metálicos atravesados

por alambres. El flujo eléctrico que conducían estos alambres magnetizaba en forma estable los anillos, que almacenaban un bit de información cada uno. El sistema de memoria podía leer, más tarde, la **polaridad** magnética de cada anillo, y así se recuperaba el valor binario que había sido almacenado en ese bit.

Memoria Micro-SD

La tecnología de las memorias ha evolucionado espectacularmente desde la creación de las primitivas memorias de núcleos. Hoy, cincuenta años después, existen memorias de bajo costo, menor tamaño, mayor velocidad, y capacidad millones de veces superior. Una memoria Micro-SD de hoy, por ejemplo, puede alojar 32 GiB de información en una centésima parte del espacio ocupado por un 1 Kib de memoria de núcleos.

ENIAC

Presentado en 1946, **ENIAC** es reconocido como el primer computador digital, completamente electrónico, de propósito general. Usaba números representados en base 10. Tenía una capacidad de memoria de 1000 bits donde podía almacenar unos veinte números decimales de diez dígitos.

El ENIAC fue propuesto para cómputos de trayectoria de proyectiles, aplicación en la que logró reducir el tiempo de cómputo de una tabla de datos, de 20 horas a 30 segundos. Sin embargo, la guerra terminó antes de que pudiera ser realmente aplicado, por lo que se lo destinó a otros usos. Por este hecho, llamó la atención sobre la capacidad de las computadoras de ser destinadas a propósitos generales, en lugar de las máquinas de programa cableado que existían hasta entonces y que eran preparadas específicamente para una única tarea.

Clementina

¿Qué pasaba en nuestro país durante estas épocas? La actividad de la computación aquí no había comenzado. Recién a principios de los años 60 la universidad argentina decidió hacer una importante inversión, que fue la compra de una computadora de primera generación, bautizada aquí **Clementina**. El video adjunto cuenta interesantes detalles técnicos de la computadora, muestra cómo eran las personalidades involucradas por ese entonces en el proyecto científico y tecnológico argentino, y explica el contexto histórico en el que fue iniciado (y, lamentablemente, truncado) ese proyecto.

Segunda Generación

En 1948 los físicos habían descubierto que combinando, en ciertas proporciones, elementos que eran vecinos en la Tabla Periódica, se creaban nuevos materiales

con un desbalance de electrones; y que de esta manera se podía controlar el sentido de las corrientes eléctricas que atravesaban esos materiales. Así fue inventado un componente electrónico revolucionario, el **transistor**, que era básicamente un **triodo de estado sólido**, es decir, podía cumplir el mismo papel en un circuito que la válvula termoiónica de tres electrodos, pero era construido de una forma completamente diferente.

Esto significa que las mismas funciones lógicas de los interruptores, que en las computadoras de primera generación eran cumplidas por las válvulas termoiónicas, podían ser resueltas con dispositivos mucho más pequeños, de mucho menor consumo, con tiempos de reacción mucho menores y mucho más confiables. El impacto tecnológico y económico de este avance fue importantísimo y la computación "despegó". Fue posible aumentar la complejidad de las funciones, creando CPUs mucho más poderosas.

Decimos que esta segunda generación de computadoras fue construida con dispositivos **discretos**, es decir, separados, para distinguirla de la generación siguiente, donde esos dispositivos fueron **integrados**.

El transistor

El **transistor** consiste en tres pequeñas piezas, puestas en contacto entre sí, hechas de materiales especialmente preparados. Estas piezas consisten, cada una, de un núcleo de un cierto elemento, conteniendo impurezas de un segundo elemento diferente.

Si se contamina, en forma controlada, un núcleo principal de germanio o silicio, con una pequeña proporción de impurezas de boro, aluminio, arsénico, o galio, el material resultante tiene propiedades de conductividad especiales.

Según la combinación de los elementos, el material será de tipo **P** (atrae electrones de materiales cercanos) o de tipo **N** (emite electrones a los materiales cercanos). Disponiendo en capas alternadas tres pequeños bloques de materiales de tipos P y N se construyen dispositivos de tipo **PNP** o de tipo **NPN**. La capa central (llamada la **base** del transistor) actúa como la grilla de las válvulas: controla el paso de corriente a través de las otras dos capas (llamadas **colector** y **emisor** del transistor). El transistor resulta así un reemplazo eficaz de las antiguas válvulas termoiónicas.

Gracias a estas propiedades el transistor funciona como un interruptor, con el cual se pueden implementar todas las funciones lógicas necesarias en los circuitos de la CPU y otros componentes de las computadoras.

PDP-1

Los transistores abarataron las computadoras y redujeron su tamaño. Algunas de las desarrolladas en esta época recibieron el nombre de **minicomputadoras**.

El PDP-1 fue uno de los primeros computadores que pudieron ser accedidos masivamente por los estudiantes de computación. Tenía un **sistema de tiempo compartido (time-sharing)** que hacía posible la utilización de la máquina por varios usuarios a la vez. Tenía 144 KB de memoria principal y ejecutaba 100.000 instrucciones por segundo.

Tercera Generación

A mediados de los 60 se desarrollaron los **circuitos integrados o microchips**, que empaquetaban una gran cantidad de transistores en un solo componente, con importantes mejoras en el aspecto funcional y en la economía de la producción de computadoras. Aparecieron computadoras más baratas que llegaron a empresas y establecimientos educativos más pequeños, popularizándose el uso de la computación.

Los fabricantes comenzaron a producir familias de computadoras **compatibles**, comenzando con el System/360 de IBM. Los productos de la misma familia utilizaban el mismo lenguaje ensamblador, lo que permitía la portabilidad de los programas entre diferentes computadoras. El usuario podía hacer crecer su infraestructura de cómputo sin perder la inversión hecha en software.

También aparecieron las primeras **supercomputadoras**, como el Cray-1, en 1976, que ejecutaba 160 millones de instrucciones por segundo y tenía 8 MiB de memoria principal.

El **microprocesador** desarrollado por Intel reunió la mayor parte de las funciones de las computadoras en un solo microchip. La existencia del microprocesador favoreció la creación de una industria de las computadoras personales. En 1982 IBM propuso el PC (Personal Computer), del cual descienden la mayoría de las computadoras domésticas y de oficina que se usan hoy. Al contrario que las computadoras de hasta entonces, construidas con procedimientos y componentes propios del fabricante, y a veces secretos, la **arquitectura abierta** del PC estaba públicamente documentada, de manera que otras empresas podían libremente fabricar componentes compatibles con esta computadora.

Circuitos Integrados

Los **circuitos integrados** fueron el resultado de un proceso de fabricación completamente nuevo llamado **fotomicrolitografía**. Los ingenieros preparaban un diagrama del circuito deseado, con sus transistores, conexiones y demás componentes, y el diagrama se reducía por medios ópticos hasta un tamaño casi microscópico. La imagen resultante se grababa sobre un sustrato de silicio.

Repitiendo el procedimiento con diferentes diagramas, sobre sucesivas capas de materiales semiconductores y aislantes, se lograba la miniaturización de un circuito completo con miles de transistores en un espacio muy reducido. Los

modernos circuitos integrados reúnen miles de millones de transistores en menos de un centímetro cúbico de volumen.

El microprocesador

Un programador utiliza un microprocesador a través de su **conjunto de instrucciones** (aritméticas, de transferencia, de salto, etc.). Por otro lado, el microprocesador tiene un cierto funcionamiento, que el programador debe conocer, respecto de qué papeles cumplen sus registros, qué efecto tienen las instrucciones sobre esos registros, qué modos de acceder a los datos son posibles, etc. Este comportamiento puede llamarse un **modelo de programación** del microprocesador.

Este conjunto de instrucciones y ese modelo de programación, reunidos, son la **arquitectura del conjunto de instrucciones** del microprocesador, abreviadas **ISA** por **Instruction Set Architecture**.

Este ISA ha sido implementado de alguna forma en el microprocesador, es decir, los ingenieros han definido un cierto circuito formado por transistores, compuertas y otros componentes, que hace que el microprocesador funcione de esta manera. Pero a medida que avanzan las tecnologías disponibles para la fabricación de los procesadores, aparecen nuevas formas de implementar ese funcionamiento. Un nuevo modelo de microprocesador podría tener mayores capacidades, sin cambiar el ISA. Por ejemplo, podría tener más de una ALU para realizar los cómputos más rápidamente, en forma paralela; pero manteniendo el mismo conjunto de instrucciones y el mismo modelo de programación de antes.

Es conveniente que estos cambios queden invisibles al programador, porque él seguirá programando de la misma manera y seguirá corriendo sus programas sin necesidad de modificarlos.

La forma de implementar una arquitectura es llamada la **microarquitectura** del microprocesador. Una familia de microprocesadores puede evolucionar con cambios invisibles, cambiando su microarquitectura, sin cambiar la **arquitectura** y sin romper la compatibilidad con los productos anteriores.

Una familia de microprocesadores también puede ampliar su conjunto de instrucciones, pero manteniendo intactas todas las de los productos anteriores. Esta forma de compatibilidad se llama a veces retrocompatibilidad.

Intel I7

El microprocesador I7 es actualmente el procesador más avanzado para computadoras personales de la firma Intel.

- Su microarquitectura es sumamente compleja: tiene un modelo de memoria segmentado e instrucciones de longitud variable.

Intel I7

- Pertenece a una generación de procesadores donde, para enfrentar los problemas derivados de la microminiaturización, los diseñadores optaron por **replicar**, es decir, incorporar múltiples instancias de, las unidades de cómputo o **núcleos**.
- Cada uno de los núcleos, a su vez, puede ejecutar dos secuencias de programa independientes (dos **threads** o **hilos**).
- Cada núcleo tiene su memoria cache privada, dividida en cache de datos y de instrucciones, y además existe un segundo nivel de cache privada para datos e instrucciones a la vez.

Intel I7

- Además existe un tercer nivel de memoria cache compartida, donde se ubican datos que pueden ser necesitados por cualquiera de los núcleos.

Intel I7

Intel I7

- En la misma "pastilla" o unidad física del microprocesador se encuentra una unidad procesadora de gráficos o GPU. Este es un procesador especializado con una arquitectura especial, destinado a la generación de gráficos avanzados, pero que además puede utilizarse para cómputos paralelos de propósito general.

Tiempo para acceder a un dato

Como sabemos, no podemos utilizar un dato si no lo hacemos llegar primero al procesador o CPU; y el tiempo que tarda en llegar a un registro de la CPU, para poder operar sobre él, depende de dónde esté localizado este dato. Es interesante comparar los diferentes tiempos de demora en el acceso a un dato, o **latencia**, según en qué componente del sistema de cómputo se encuentra ese dato.

En la tabla adjunta tomamos como referencia un **ciclo de CPU**, es decir, el cambio de estado más pequeño posible en el circuito secuencial que implementa el procesador de la computadora. Los procesadores actuales utilizan pulsos de reloj de alrededor de 3 GHz, es decir, el reloj del sistema genera alrededor de 3.000.000.000 de señales por segundo; lo que da un tiempo de ciclo de unos 0.3 ns. Una instrucción de CPU puede llevar uno o varios ciclos para completarse, según la microarquitectura de la CPU. Pero la CPU, para ejecutar esa instrucción,

necesita tener a disposición, en sus registros, los datos sobre los cuales debe operar.

¿Cuánto lleva entonces acceder a un dato, en términos de esta duración básica de un ciclo? Si el dato está en memoria RAM, llevará unos 120 ns (unos cuatrocientos ciclos). Si está en el disco magnético, demorará unos 10 ms en llegar al procesador (unos 33000 ciclos).

Entonces, si una instrucción de CPU puede completarse en unos pocos ciclos, pero debe esperar ¡cientos o miles de ciclos! a que los datos atraviesen el sistema de memoria, habrá una **enorme** espera improductiva para la CPU. Peor aún, si el dato ¡debe llegar desde otro continente vía la Internet!

Si la CPU tuviera que esperar por los demás componentes, su **utilización** se reduciría ridículamente, y su gran velocidad de procesamiento quedaría completamente desperdiciada. Por esto es que se establece una **jerarquía de memoria**, con lugares de almacenamiento cuyas velocidades de acceso son cada vez mayores a medida que nos acercamos a la CPU en el sistema de cómputo. Por ejemplo, un dato que se encuentre en **memoria cache** ("cerca" de la CPU) tendrá una latencia de acceso mucho menor y será preferible a tener que accederlo desde la memoria RAM (más "lejana" en el camino de los datos).

Hay muchas otras medidas técnicas que toman los ingenieros de las modernas computadoras para resolver esta disparidad de los tiempos de acceso, tales como fabricar CPUs con varias unidades de cómputo ("**cores** o **núcleos**") que funcionan en paralelo, o diseñar complejos mecanismos de cómputo que reordenan y procesan varias instrucciones a la vez, de manera de ocultar esas grandes latencias de acceso.

Para comprender mejor, desde nuestra perspectiva de humanos, la importancia relativa de esos tiempos de respuesta, la tabla se **escala** al tiempo del ciclo de CPU. Es decir, los tiempos bajo la columna "Escalado" son aquellos que tardaría cada acceso **si un ciclo de CPU durara un segundo**.

Para completar la tabla, comparamos un ciclo de CPU con el proceso de **reboot** o reencendido de la computadora ("¿probó apagar y volver a encender el equipo?").

Memorias vs. CPU

En el gráfico comparamos las capacidades de memorias y procesadores en cuatro momentos relativamente recientes en el tiempo, que son cuando aparecieron cuatro especificaciones de memoria distintas: DDR (2002), DDR2 (2004), DDR3 (2007) y DDR4 (2013).

Leyenda

- Volts (V): voltaje de funcionamiento de las memorias

- Velocidad (MHz): velocidad de reloj de las memorias
- Densidad (Gb): capacidad de cada chip de memoria
- Transf (GB/s): velocidad de transferencia de la memoria
- SPECint CPU: valor del **benchmark** SPECint, que mide la capacidad de procesamiento de enteros, para procesadores comparables en cada año
- nCores: cantidad de **cores** o unidades de procesamiento en un mismo chip

Los datos están presentados como factores de escala de crecimiento, es decir, representan en qué medida cambió cada variable con respecto a 2002, que es el año en que apareció el primer estándar DDR. Por ejemplo, el factor de crecimiento de la densidad, o cantidad de Gb por chip de memoria, es 2 en 2004, porque en ese año aparecieron memorias DDR2 del doble de tamaño que las de DDR, y es 32 en 2013, porque en ese año su tamaño se multiplicó por 32 con respecto al valor del estándar DDR de 2002.

Notemos que podemos hacer click en las barras de color de la leyenda, al pie del gráfico, para ocultar una variable y estudiar cómo se relacionan las demás. Por ejemplo, si ocultamos la variable de la **densidad**, las demás variables muestran más claramente las relaciones entre ellas.

Al aparecer la primera computadora personal o **PC** en 1982, las memorias eran más rápidas que los procesadores. Sin embargo, en los últimos años los procesadores han evolucionado espectacularmente, y las tecnologías de memorias no han seguido la misma tendencia ascendente, convirtiendo a las memorias en un **cuello de botella**. Mientras los procesadores aumentaban su velocidad de procesamiento, las memorias se aceleraban en una proporción menor, lo que ocasionaba un desbalance cada vez mayor en los sistemas.

Si bien la velocidad de procesamiento de los procesadores venía aumentando desde fines del siglo XX a razón de un 50% por año, a partir de 2002 se encuentra la limitación del sistema de memoria que obliga a los diseñadores de CPUs a tomar decisiones de diseño especiales, como la inclusión de múltiples unidades de procesamiento o **cores**.

La consecuencia es que la mayoría de las computadoras actuales son máquinas paralelas, y la programación de las aplicaciones debe hacerse considerando este hecho para aprovechar el sistema de cómputo adecuadamente.