Arquitecturas multiprocesadores

Pedro O. Pérez M., MTI

Multiprocesadores
Tecnológico de Monterrey

pperezm@tec.mx

08-2019

Arquitecturas de computadoras

- RISC
- CISC
- ARM
- GPU

Cache

- Coherencia del cache

Arquitecturas de procesadores

- Core i5
- Ryzen
- ARM
- GPU

Arquitecturas de computadoras

- Se define como el diseño conceptual y la estructura operacional de un sistema de computadoras, especialmente todo lo relacionado con la forma en que trabaja el CPU y cómo accede a la memoria.
- También suele definirse como la forma en que se interconectan los componentes de hardware, para crear computadoras según los requerimientos de funcionalidad, rendimiento y costo.

Arquitecturas de computadoras

La computadora recibe y envía la información a través de los periféricos, por medio de los canales. La CPU es la encargada de procesar la información que le llega a la computadora. El intercambio de información se tiene que hacer con los periféricos y la CPU.

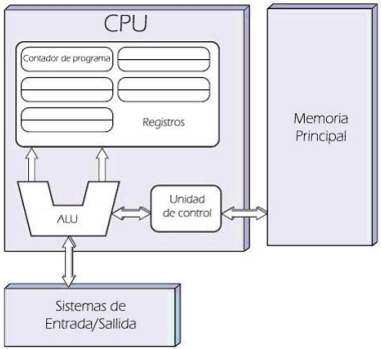


Figura: https://bit.ly/2NhEORD

Arquitecturas de computadoras

Existen muchas arquitecturas de computadoras, sin embargo mencionaremos solo las tres más importantes:

- RISC (Reduced Instruction Set Computing).
- CISC (Complete Instruction Set Computing).
- ARM (Advanced RISC Machine).
- GPU.

RISC

Un Computador con Conjunto de Instrucciones Reducidas (RISC) es un CPU generalmente utilizado en microprocesadores o microcontroladores con las siguientes características fundamentales:

- Instrucciones de tamaño fijo y presentadas en un reducido número de formatos.
- Sólo las instrucciones de carga y almacenamiento acceden a la memoria de datos.

Su objetivo es posibilitar la segmentación y el paralelismo en la ejecución de instrucciones y reducir los accesos a memoria. Ejemplos de esta arquitectura son los microprocesadores. PowerPC, DEC Alpha, MIPS, ARM y SPARC.

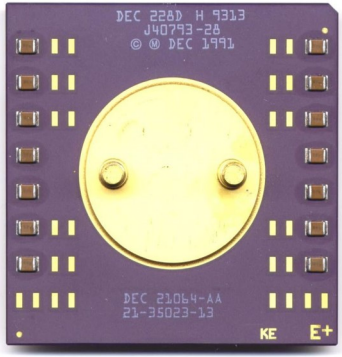


Figura: https://bit.ly/2zb9Gj4

CISC

Un Computador de Conjunto de Instrucciones Complejas (CISC) tienen un conjunto de instrucciones que se caracteriza por ser muy amplio y permitir operaciones complejas entre operandos situados en la memoria o en los registros internos, en contraposición a la arquitectura RISC.

- Este tipo de arquitectura dificulta el paralelismo entre instrucciones.

Ejemplos de esta arquitectura son los microprocesadores Motorola 68000, Zilog Z80, Intel x86 y AMDx86-64.

ARM

ARM es una arquitectura RISC de 32 bits y, con la llegada de su versión V8-A, también de 64 Bits, desarrollada por ARM Holdings.

- Se llamó Advanced RISC Machine, y anteriormente Acorn RISC Machine.
- La arquitectura ARM es el conjunto de instrucciones de 32 y 64 bits más ampliamente utilizado en unidades producidas.

Ejemplos de esta arquitectura son los microprocesadores Applied Micro Circuits Corporation X-Gene, DEC StrongARM, Freescale i.MX, Marvell Technology Group XScale, NVIDIA Tegra, Qualcomm Snapdragon, Texas Instruments OMAP, Samsung Exynos, Apple Ax, ST-Ericsson NovaThor, Huawei K3V2 e Intel Medfield.

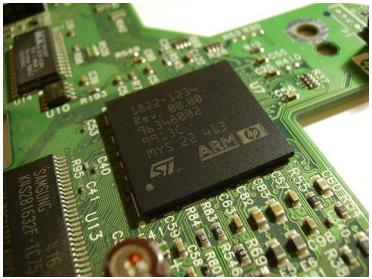


Figura: https://bit.ly/2TPTmhy

GPU

- La arquitectura NVIDIA GPU se basa es un matriz escalable de multiprocesadores. Un multiprocesador está diseñado para ejecutar cientos de hilos al mismo tiempo. Para administrar una cantidad de hilos tan grande se emplea una arquitectura única llamada SIMT (Single-Instruction, MultipleThread).
- Las instrucciones se canalizan para aprovechar el paralelismo a nivel instrucción dentro de un único subproceso, así como el paralelismo a nivel de subprocesos a través de multiprocesamiento de hardware simultáneo.

`https://bit.ly/1mKiP2s`

Cache

- La memoria cache es un componente usado por el CPU para reducir el costo promedio (en tiempo o energía) de acceso a los datos que se encuentran en memoria principal.
- La mayoría de los CPUs tienen varias memorias cache independiente, de datos o instrucciones, que se encuentran organizados en una jerarquía de varios niveles (L1, L2, etc.).
- La información que se cargan en la memoria cache depende de algoritmos sofisticados y ciertas suposiciones sobre el código del programa. El objetivo del sistema cache es garantizar que el CPU tenga el siguiente bloque de datos que necesitará ya cargado en memoria.

Coherencia del cache

Para mayor rendimiento en un sistema multiprocesador, cada procesador generalmente tiene su propio cache. La coherencia de cache se refiera al problema de mantener la coherencia de los datos de estas caches. El principal problema es lidiar con las escrituras de un procesador.

https://bit.ly/2TPTmhy

Coherencia del cache

Hay dos estrategias generales para tratar las escrituras en un cache:

- Escritura: Todos los datos escritos en el cache también se escriben en la memoria al mismo tiempo.
- Reescritura: Cuando los datos se escriben en un cache, se marca como "sucio" el bloque afectado. El bloque modificado se escribe en memoria solo cuando se reemplaza el bloque.

`https://bit.ly/2Hkqt86`

Core i5

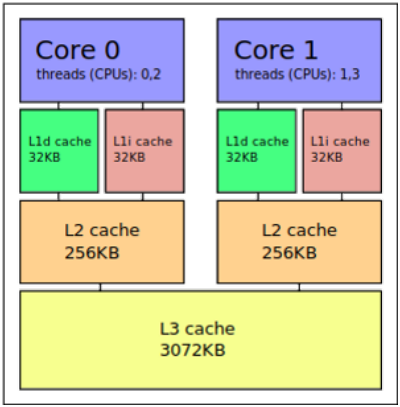
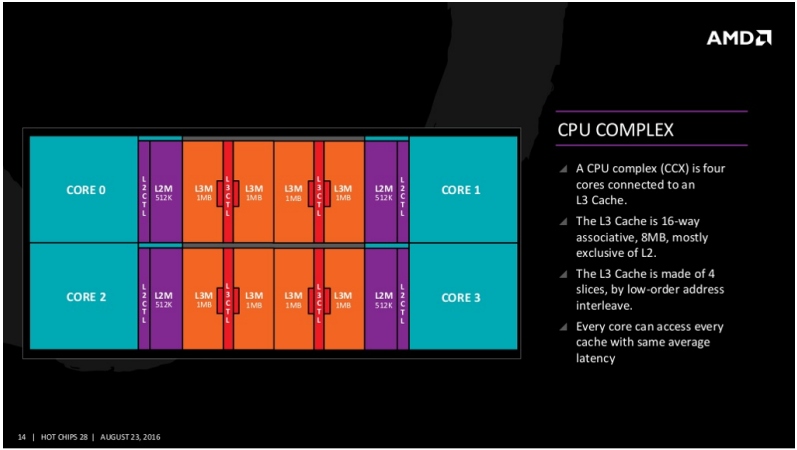


Figura: `https://bit.ly/2Zib3ai`

Ryzen


AMD
CPU COMPLEX
A CPU complex (CCX) is four cores connected to an L3 Cache.
The L3 Cache is 16-way associative, 8MB, mostly exclusive of L2.
The L3 Cache is made of 4 slices, by low-order address interleave.
Every core can access every cache with same average latency
CORE 0
CORE 1
CORE 2
CORE 3
L2M 512K
L3M 1MB
14 | HOT CHIPS 28 | AUGUST 23, 2016

ARM

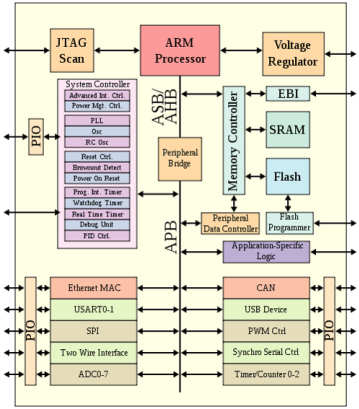


Figura: https://bit.ly/33LLvWH

GPU

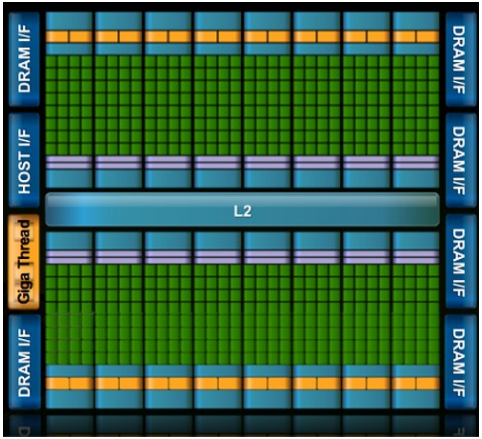


Figura: https://bit.ly/310mtnr

Definición

Una computadora paralela es un conjunto de procesadores que trabajan de manera cooperativa en la solución de un problema computacional. Esta definición es lo suficientemente amplia como para incluir supercomputadoras paralelas que tienen cientos o miles de procesadores, redes de estaciones de trabajo, estaciones de trabajo de múltiples procesadores y sistemas embebidos.

Metas

- El propósito principal del cómputo paralelo es acelerar la capacidad de procesamiento o, en otras palabras, aumentar la velocidad computacional.
- Incrementar el rendimiento, es decir, la cantidad de procesamiento que puede lograrse durante un intervalo de tiempo determinado.
- Mejorar el rendimiento de la computadora sin incrementar la velocidad de reloj.
- Dos o más ALUs en el procesador pueden funcionar simultáneamente para aumentar el rendimiento.
- El sistema puede tener dos o más procesadores operando al mismo tiempo.

Tipos de paralelismo

- Paralelismo a nivel de instrucción.
  - Nivel de bits.
  - Pipelining.
  - Super-escalar.
- Paralelismo a nivel de procesador.
  - Arreglo de computadoras.
  - Multiprocesadores.

Nivel de bits

- El paralelismo a nivel de bit es una forma de computación paralela basada en el aumento del tamaño de la palabra del procesador. Aumentar el tamaño de palabra reduce el número de instrucciones que el procesador debe ejecutar para realizar una operación en variables cuyos tamaños son mayores que la longitud de la palabra.
- Otra manera es el aumento del ancho del bus de datos externo. Por ejemplo, DDR1 transfiere a 128 bits por ciclo de reloj, mientras que DDR2 transfiere un mínimo de 256 bits por ráfaga.

Pipelining

- Un programa de computadora es, en esencia, una secuencia de instrucciones ejecutadas por un procesador. Si el paralelismo a nivel de instrucción, un procesador solo podría ejecutar menos de una instrucción por ciclo de reloj (Instruction Per Cycle, IPC < 1). Este tipo de procesadores se conocen como sub-escalares.
- Sin embargo, las instrucciones pueden reordenarse y combinarse en grupos que luego se ejecutan en paralelo sin cambiar el resultado del programa.

Pipelining

- Todos los procesadores modernos tiene tuberías de varias etapas. Cada etapa en la tubería corresponde a una acción diferente que el procesador realiza en una instrucción en esa etapa; un procesador con una tubería de N etapas puede tener hasta N instrucciones diferentes en diferentes etapas de ejecución (ciclo Fletch-Decode-Execute) y, por lo tanto, puede emitir una instrucción por ciclo de reloj (IPC = 1). Estos procesadores se conocen como escalares. Un ejemplo son los procesadores RISC que cuentan con cinco etapas: búsqueda de instrucción (IF), decodificación de instrucción (ID), ejecución (EX), acceso a memoria (MEM) y recuperación de registro (WB).
- **Considera que un Intel Core i7-7700K tiene un IPC, usando un solo hilo, de 174.**

Pipelining

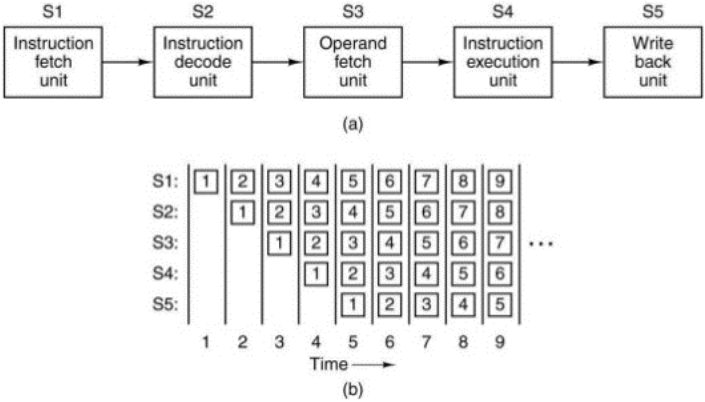


Figura: https://bit.ly/2MuP1PG

Super-escalar

- Sin embargo, la mayoría de los procesadores actuales tiene múltiples unidades de ejecución. Por lo tanto, si combinamos el IPC de un solo hilo con los múltiples hilos ejecución tenemos un mayor IPC. Estos procesadores se conocen como super-escalares.
- **De nuevo, tomemos el ejemplo del Intel Core i7-7700K, el cual cuenta con 8 hilos de ejecución. Este procesador super-escalar tiene un IPC de 872.**

Paralelismo vs. Concurrencia

- El cómputo paralelo está fuertemente relacionado con el cómputo concurrente. Es muy frecuente se empleen juntas, incluso que se combinen; sin embargo son muy diferente: podemos tener paralelismo sin concurrencia (como el paralelismo a nivel de instrucción) y podemos tener concurrencia sin paralelismo (como la multitarea).
- En el cómputo paralelo, una tarea se divide en varias sub-tareas muy similares que se pueden procesar de forma independiente y cuyos resultados se combinan para formar la solución final.
- En la computación concurrente, los diversos procesos no ejecutan tareas relacionadas; cuando lo hacen, las tareas pueden ser de una naturaleza variada y, muy a menudo, requieren alguna comunicación entre procesos durante la ejecución.

Cálculo de desempeño

La velocidad de un programa es el tiempo que tarda el programa en ejecutarse. Esto podría medirse en cualquier incremento de tiempo. SpeedUp se define con el tiempo que le toma a un programa ejecutarse secuencialmente (con un procesador) dividido por el tiempo que lleva ejecutarse en paralelo (con muchos procesadores o hilos).

Cálculo de desempeño

La fórmula para calcular el SpeedUp es:
Donde:

- S_p – La mejora obtenida al usar p procesadores.
- T_1 - El tiempo que tarda el programa en ejecutarse secuencialmente.
- T_p - El tiempo que tarda el programa en ejecutarse usando p procesadores

$$S_p = \frac{T_1}{T_p}$$