TEMA 3:

ARQUITECTURAS CON PARALELISMO A NIVEL DE THREAD (TLP)

Lección 1: ARQUITECTURAS TLP.

Las arquitecturas TLP con una instancia del SO se pueden clasificar en:

- Multiprocesadores: Ejecutan varios flujos de instrucciones en paralelo en un computador con varios procesadores. Se pueden encontrar multiprocesadores en un chip, en una placa, en un armario o formados por varios armarios.
- Multinúcleos (multicores): Ejecutan varios flujos de instrucciones en paralelo en un chip de procesamiento con múltiples núcleos, cada uno en un núcleo distinto. Chip multinúcleo: multiprocesador en un chip.
- Núcleos (cores) multithread: Núcleo de procesamiento en el que se ha modificado su arquitectura ILP (Instruction Level Parallelism) para ejecutar flujos de instrucciones concurrentemente o en paralelo.

Multiprocesadores:

- UMA: En los primeros multiprocesadores, el acceso a memoria era uniforme, es decir, todos los procesadores tardaban lo mismo en acceder a todas las posiciones de memoria. Mayor latencia/poco escalable.
- NUMA: Para incrementar la escalabilidad, se distribuyeron físicamente los módulos de memoria compartida entre procesadores, permitiendo de esta forma configuraciones de cientos de procesadores manteniendo unas buenas prestaciones. En esta última configuración el acceso a memoria ya no es uniforme, acceden más rápido a aquellas que se encuentran en los módulos que tienen más cercanos. Menor latencia/escalable (pero requiere para ello distribución de datos/código).

Multinúcleos:

- Ejecutan varios threads en paralelo en un chip de procesamiento multicore.

Núcleos multithread:

- Etapas de la arquitectura ILP.
    - IF (Instruction Fetch): Captación de instrucciones.
    - ID (Instruction Decode): Decodificación de instrucciones y emisión a unidades funcionales.
    - Ex (Execution): Etapa de ejecución.
    - Mem (Memory): Etapa de acceso a memoria.
    - WB (Write-Back): Etapa de almacenamiento de resultado.

- Arquitecturas ILP:
    - Procesadores/cores segmentados: Ejecución de las instrucciones concurrentemente segmentando el uso de sus componentes.
    - Procesadores/cores VLIW (Very Large Instruction Word) y superescalares: Ejecutan instrucciones concurrentemente (segmentación) y en paralelo (múltiples unidades funcionales que emiten múltiples instrucciones en paralelo a unidades funcionales).
        * VLIW: Las instrucciones que se ejecutan en paralelo se captan juntas en memoria. Este conjunto de instrucciones forman la palabra de instrucción muy larga (denominación de VLIW). El hardware presupone que las instrucciones de una palabra son independientes (no tiene que encontrar instrucciones que puedan emitirse y ejecutarse en paralelo).
        * Superescalares: Encontrar instrucciones que puedan emitirse y ejecutarse en paralelo (tiene hardware para extraer paralelismo a nivel de instrucción).

- Clasificación de cores multithread.
    - TMT (Temporal Multithreading):
        * Ejecutan varios threads concurrentemente en el mismo core.
        * La conmutación entre threads la decide y controla el hardware.
        * Emite instrucciones de un único thread en un ciclo.
        * FGMT (Fine-grain multithreading): La conmutación entre threads la decide el hardware cada ciclo por turno rotatorio o por eventos de cierta latencia, combinado con alguna técnica de planificación.
        * CGMT (Coarse-grain multithreading): La conmutación entre threads la decide el hardware tras intervalos de tiempo prefijados o por eventos de cierta latencia. Clasificación:
            - *Estática:*
              La conmutación puede ser explícita (instrucciones añadidas al repertorio) o implícitas (instrucciones de carga, almacenamiento, salto).
              Ventaja: Coste cambio contexto bajo.
              Inconveniente: Cambios de contexto innecesarios.
            - *Dinámica:*
              La conmutación típicamente es por fallo en la última cache dentro del chip de procesamiento (conmutación por fallo de cache), interrupción (conmutación por señal) ...
              Ventaja: Reduce cambios de contexto innecesarios.
              Inconveniente: Mayor sobrecarga la cambiar de contexto.

    - SMT o multihilo (Simultaneous Multithreading):
        * Ejecutan, en un core superescalar, varios threads en paralelo.
        * Pueden emitir (para su ejecución) instrucciones de varios threads en un ciclo.

Hardware y arquitecturas TLP en un chip:

Hardware	CGMT	FGMT	SMT	CMP
Registros	**replicado** (al menos PC)	**replicado**	**replicado**	**replicado**
Almacenamiento	multiplexado	multiplexado, compartido, repartido o **replicado**	compartido, repartido o **replicado**	**replicado**
Otro hardware de las etapas del cauce	multiplexado	Captación: repartida o compartida; Resto: multiplexadas	UF: compartidas; Resto: repartidas o compartidas	**replicado**
Etiquetas para distinguir el thread de una instr.	Sí	Sí	Sí	**No**
Hardware para conmutar entre threads	Sí	Sí	**No**	**No**