

14. Calcolatori Multicore

Evoluzione delle prestazioni hardware

I microprocessori hanno registrato una crescita esponenziale delle prestazioni attraverso miglioramenti nell'organizzazione, incremento della frequenza di clock e crescita del parallelismo. Le tecniche di parallelismo includono pipeline, architetture superscalari e multithreading simultaneo (SMT). Questa evoluzione comporta problemi: maggiore complessità della logica, aumento dell'area del chip e difficoltà di progettazione, realizzazione e debug.

Organizzazioni del chip

Esistono tre principali architetture: superscalare (singolo program counter, singolo banco registri), SMT (multipli program counter, banchi registri replicati) e multicore (multipli core superscalari o SMT con cache condivisa).

Complessità e consumo energetico

La potenza cresce esponenzialmente con la densità del chip e la frequenza di clock. Un rimedio consiste nell'allocare più spazio alla cache, meno densa e con consumo inferiore di ordini di grandezza. Nel 2015 si stimavano 100 miliardi di transistor in 300mm² con cache da 100MB e 1 miliardo di transistor per la logica.

La **Regola di Pollack** stabilisce che le prestazioni sono proporzionali alla radice quadrata dell'incremento in complessità: raddoppiare la complessità produce solo il 40% di prestazioni aggiuntive. Le architetture multicore offrono potenziale miglioramento quasi lineare, mentre un singolo core difficilmente utilizza efficacemente tutta la cache disponibile.

Prestazioni software

I vantaggi prestazionali dipendono dallo sfruttamento efficace delle risorse parallele. Anche una piccola quantità di codice seriale impatta significativamente: il 10% di codice sequenziale su 8 processori limita lo speedup a 4,7x anziché 8x. Esistono overhead dovuti a comunicazione, distribuzione del lavoro e mantenimento della coerenza della cache.

Organizzazione multicore

Le variabili progettuali includono: numero di core per chip, numero di livelli di cache e quantità di cache condivisa. Le configurazioni possibili sono: L1 dedicata con L2 condivisa, L2 dedicate, L1 e L2 dedicate con L3 condivisa.

Vantaggi della cache L2 condivisa

Riduzione dei miss totali, assenza di replicazione dei dati condivisi a livello L2, allocazione dinamica dello spazio cache ai core in base alla località dei thread, comunicazione inter-processo facilitata, problema di coerenza confinato a L1.

Vantaggi della cache L2 dedicata

Accesso più rapido alla memoria, migliori prestazioni per thread con forte località.

Architetture Intel

Intel Core Duo (2006): core superscalari, cache L2 condivisa da 2MB, cache L1 dedicata da 32KB, unità di controllo termico, APIC, logica di gestione della potenza.

Intel Core i7 (2008): core SMT, cache L3 condivisa da 8MB, cache L1 (32KB) e L2 (256KB) dedicate, controllore memorie DDR3 on-chip, QuickPath Interconnect con banda totale di 25,6GB/s.

Intel Core i7-990X: 6 core, cache L2 dedicate da 256KB, cache L3 condivisa da 12MB.

Intel Core i7-5960X (2014): 8 core, cache L2 dedicate da 256KB, cache L3 condivisa da 20MB, controllori DDR4, PCI Express a 40 lane.

ARM Cortex-A15 MPCore

Processore multicore omogeneo per mobile computing, server digitali e infrastrutture wireless. Componenti: Generic Interrupt Controller (GIC), unità di debug, timer generici per core, supporto trace, core A15, cache L1 dedicata per istruzioni e dati, cache L2 condivisa, Snoop Control Unit (SCU) per la coerenza L1/L2.

Prestazioni multicore e codice sequenziale

Lo speedup relativo decresce all'aumentare della percentuale di codice sequenziale. Con 8 processori: 0% sequenziale produce speedup ~8x, 2% produce ~7x, 5% produce ~6x, 10% produce ~4,7x, 15-20% produce speedup inferiore a 2x. Oltre una certa soglia di codice sequenziale, aggiungere processori non produce benefici significativi.