Introduzione al Many/Multi-core Computing

Sistemi Operativi e reti

Flavio Vella

Università degli Studi di Perugia - Corso di Laurea Magistrale in Informatica

6 giugno 2011

Parte I

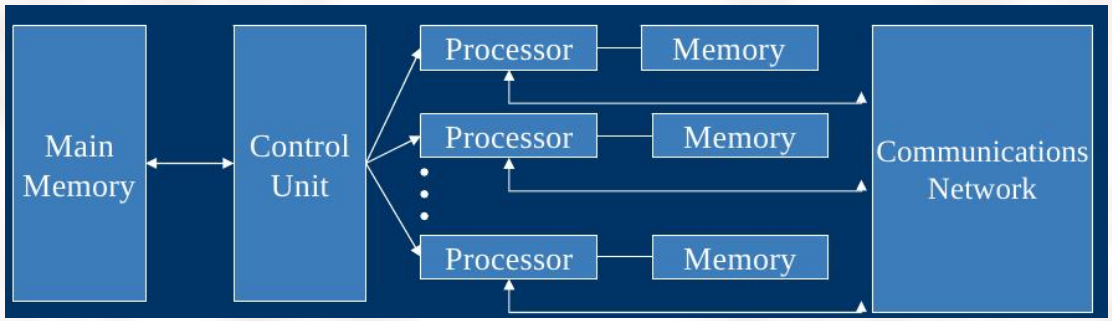
Architettura

Classificazione fra architetture

Flynn's taxonomy

- **SISD** Single instruction on Single Data- (es. architetture Von Neumann tradizionale)
- **SIMD** Single instruction on Multiple Data. (es. Processori vettoriali)
- **MISD** Multiple instruction on Single Data. (es. controller di volo dello Space Shuttle)
- **MIMD** Multiple instruction on Multiple Data. (es. architetture moderne multicore es Xeon Clovertown)

SIMD



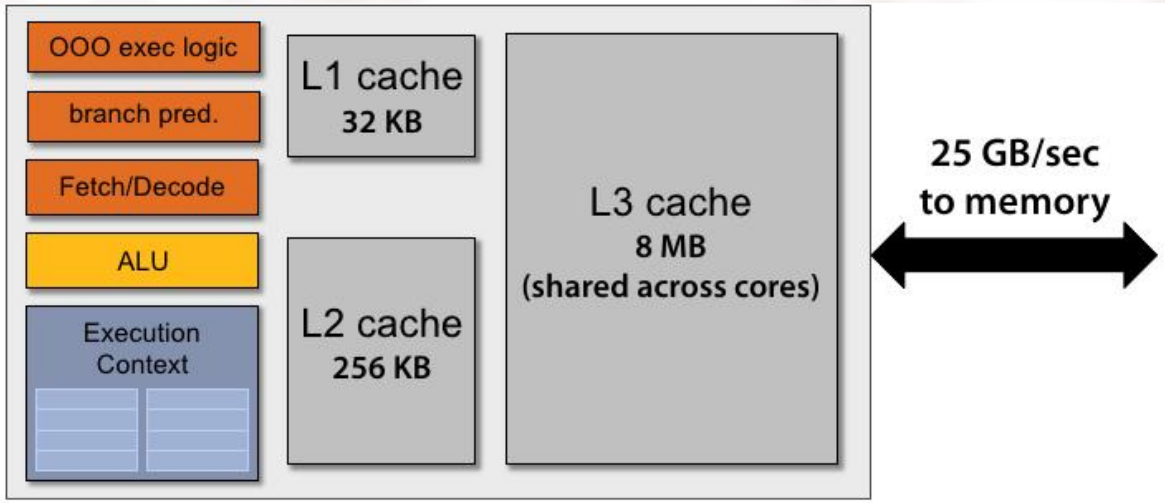
- Esegue una singola istruzione su più dati contemporaneamente utilizzando molte unità di calcolo.
- La fase di fetch e di decode dell'istruzione avviene una sola volta
- Esiste una sola control unit (CU) che gestisce il flow delle istruzioni di un dato programma.

Processori vettoriali

Definizione sono unità di calcolo che dopo aver eseguito il fetch e decode dell'instruzione la eseguono su più dati memorizzati in registri vettoriali.
I trasferimenti dati dalla memoria centrale verso i registri vettoriali (e viceversa) sono gestiti da l'unità di *load-store*.

MIMD

- Esegue istruzioni diverse simultaneamente.
- Ogni processore possiede una CU.
- Ad ogni processore può essere assegnato un task o una parte di esso.



Architettura delle GPU moderne: Nvidia G80 (2007)

Componenti principali

Le GPU sono composte da:

- Host interface: connette la periferica (device) via bus (PCIe) all'Host e ne gestisce la comunicazione (dati e istruzioni).
- Scalable Processor Array (SPA). Esegue le operazioni (programmabili) tramite i Texture Processor Cluster (TPC).
- Video RAM.

Componenti principali: Nvidia G80

Il TPC contiene: Geometry controller, due **Streaming Multiprocessors (SM)**, Texture unit e un Streaming Multiprocessor Controller (SMC).

In particolare un SM contiene:

- MT unit: multithreded instruction fetch and issue unit
- 8 Streaming Processor (SP): unità di calcolo vere e proprie (scalari).
- 2 Special Function Unit: usate per il calcolo di funzioni trascendenti.
- Cache delle istruzioni.
- Cache per la memoria costante.
- Memoria condivisa (16KB)
- 8192 registri (32 bit).

Componenti principali: Nvidia G80

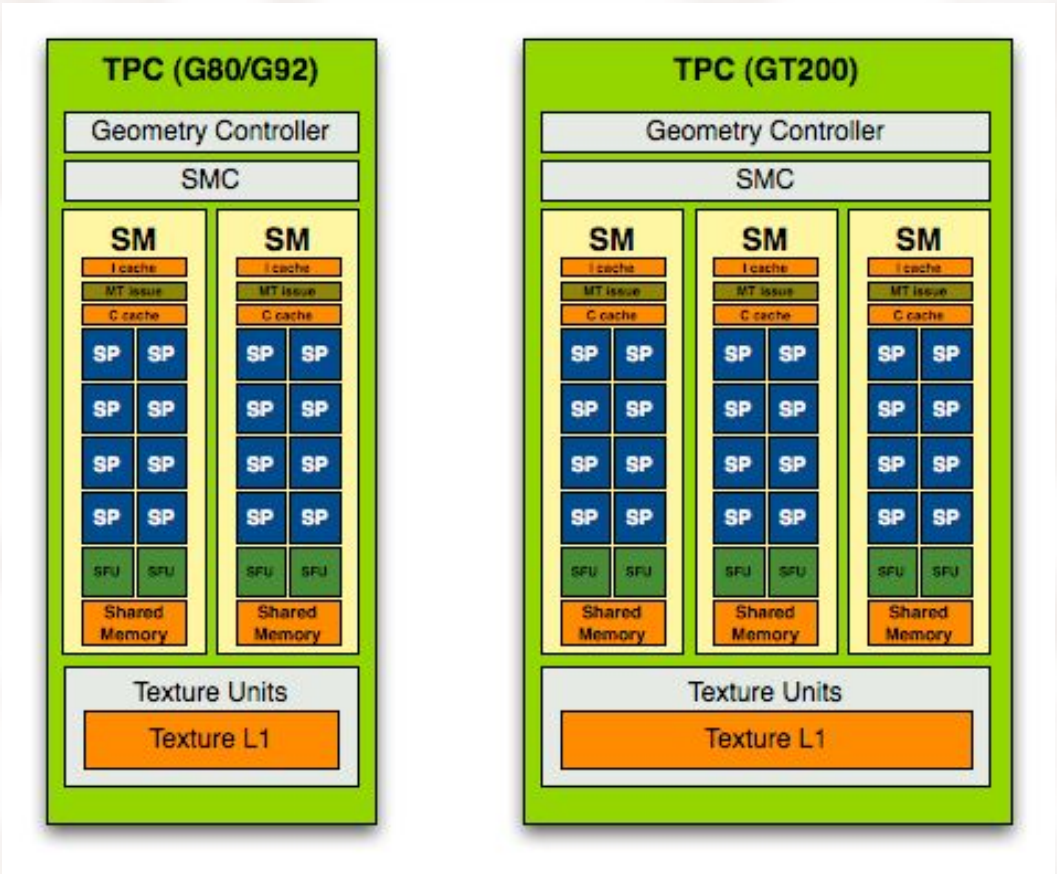
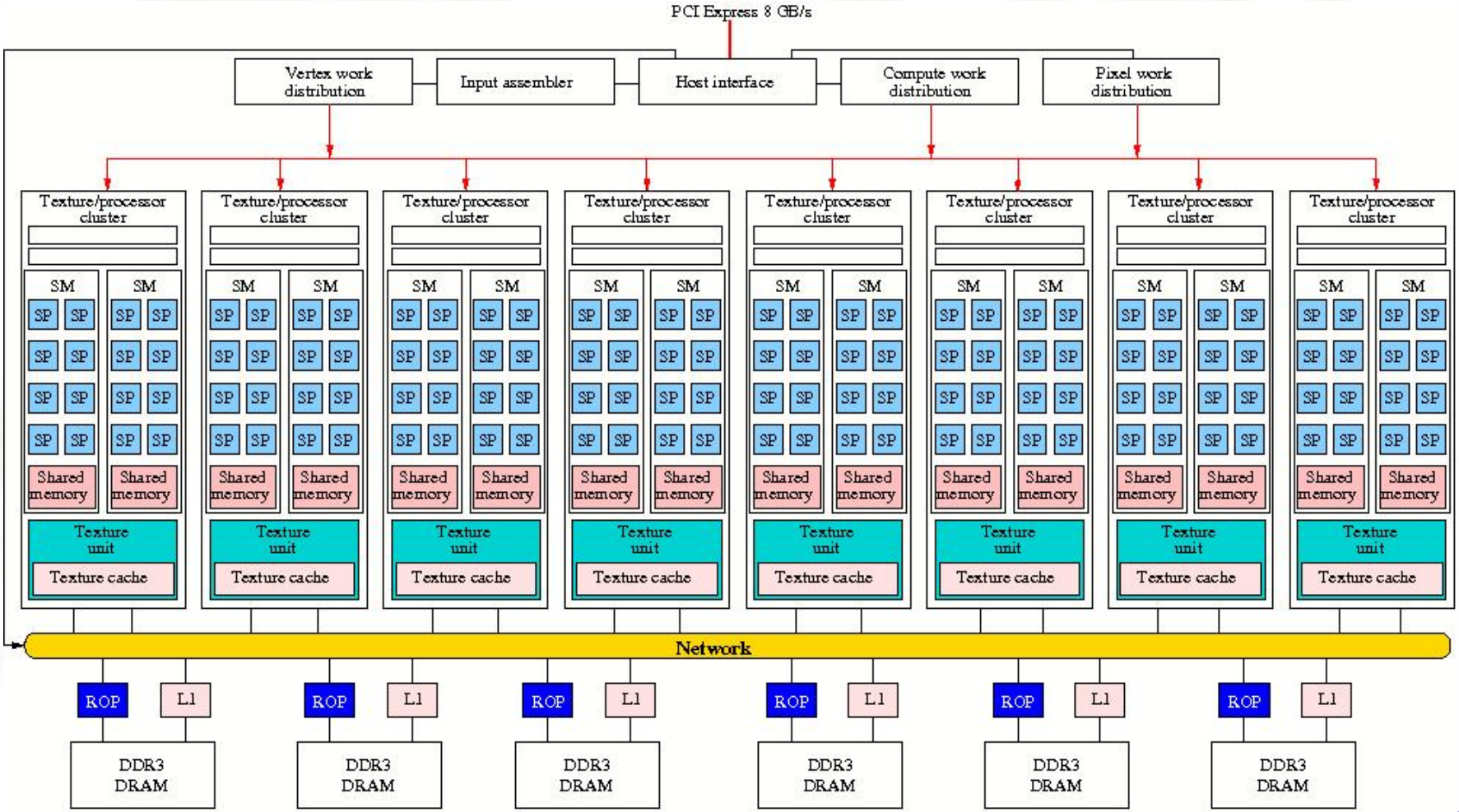


Figura: Texture Processor Cluster nel chipset G80 e GT200.

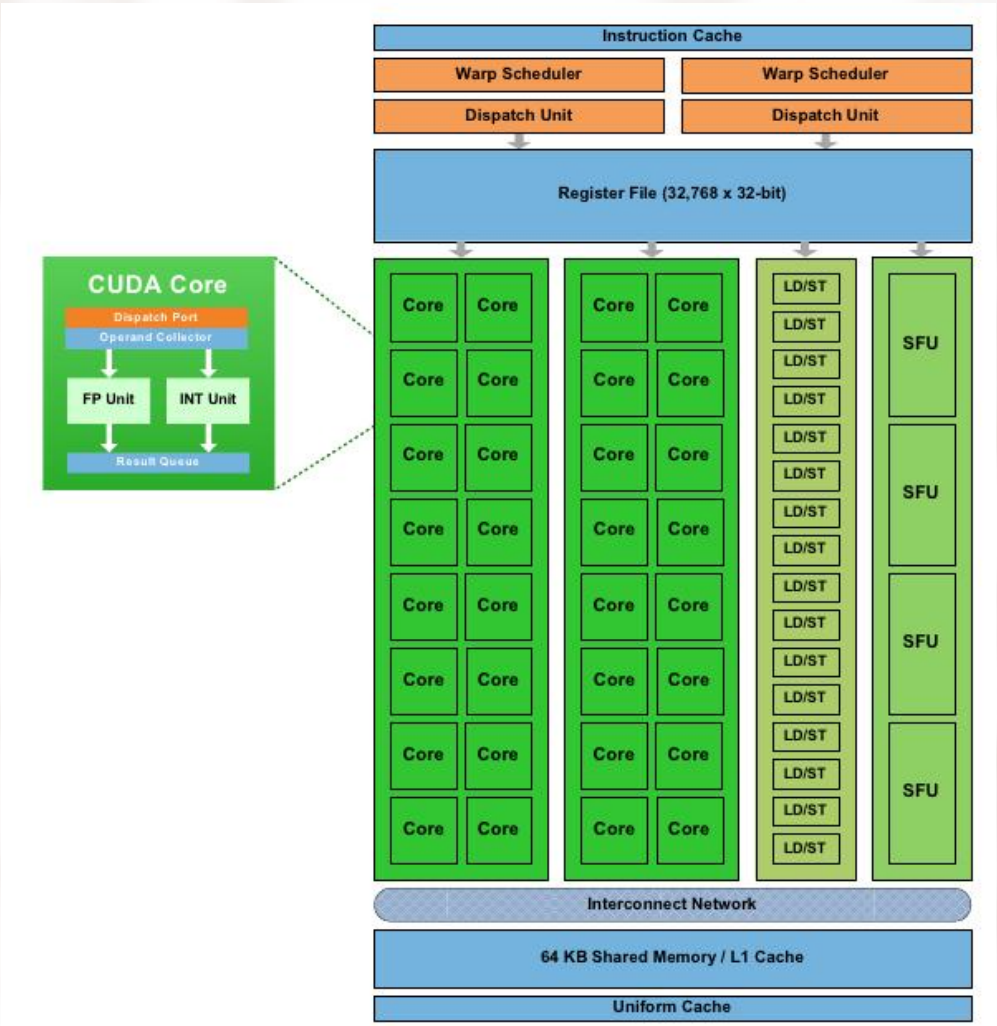
Architettura Nvidia G80



Memoria G80

- VRAM o global memory: velocità di accesso 80 GB/s. Dimensione max 512 MB
- Shared memory: 16 KB in blocchi da 1KB. 400 volte più veloce della VRAM.
- Registri o private memory: 600 volte più veloce della VRAM.

Architettura Nvidia Fermi 2011



Architettura Nvidia

GPU	G80	GT200	Fermi
Transistors	681 million	1.4 billion	3.0 billion
CUDA Cores	128	240	512
Double Precision Floating Point Capability	None	30 FMA ops / clock	256 FMA ops /clock
Single Precision Floating Point Capability	128 MAD ops/clock	240 MAD ops / clock	512 FMA ops /clock
Special Function Units (SFUs) / SM	2	2	4
Warp schedulers (per SM)	1	1	2
Shared Memory (per SM)	16 KB	16 KB	Configurable 48 KB or 16 KB
L1 Cache (per SM)	None	None	Configurable 16 KB or 48 KB
L2 Cache	None	None	768 KB
ECC Memory Support	No	No	Yes
Concurrent Kernels	No	No	Up to 16
Load/Store Address Width	32-bit	32-bit	64-bit

Maggiori dettagli `http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf`

Esecuzione delle istruzioni

L'esecuzione delle istruzioni sui dati sono organizzati in threads.

Ogni SM crea, gestisce, schedula e esegue threads in gruppi di 32 threads chiamati warps.

I threads dello stesso wrap partono dallo stesso program address ma seguono un proprio percorso indipendente.

La massima efficienza si ha quando tutti i threads di un wrap seguono lo stesso percorso (no branch).

Nvidia chiama questa architettura SIMT (Single Instruction Multiple Thread).

Maggiori dettagli OpenCL Programming Guide for the CUDA Architecture vers 3.1.