Università degli Studi di Napoli Federico II

Facoltà di Scienze MM.FF.NN.
Corso di Laurea in Informatica

Tesi sperimentale di Laurea Triennale

Calcolo ad alte prestazioni basato su GPU
Un modello ibrido neurale-genetico per data mining in astrofisica

Relatori
Prof. Guido Russo
Dr. Massimo Brescia

Candidato
Andrea Solla
matr. 566/2867

Anno Accademico 2011-2012

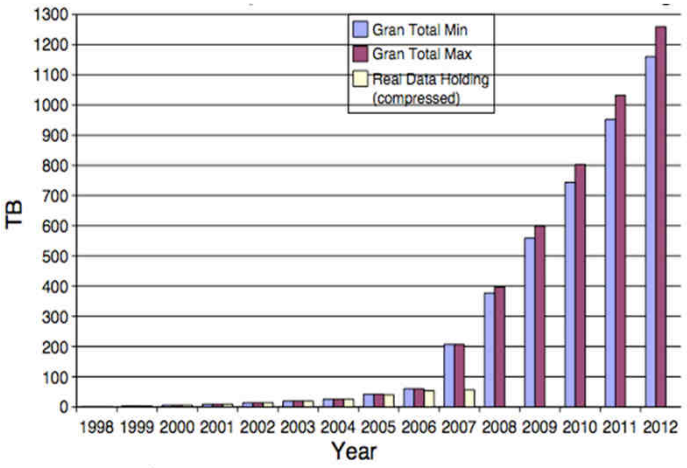
Contesto scientifico

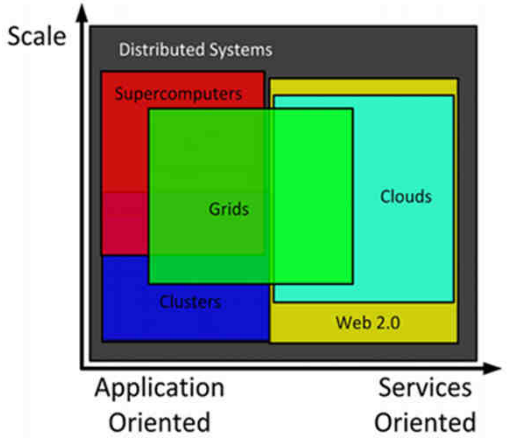
Problema

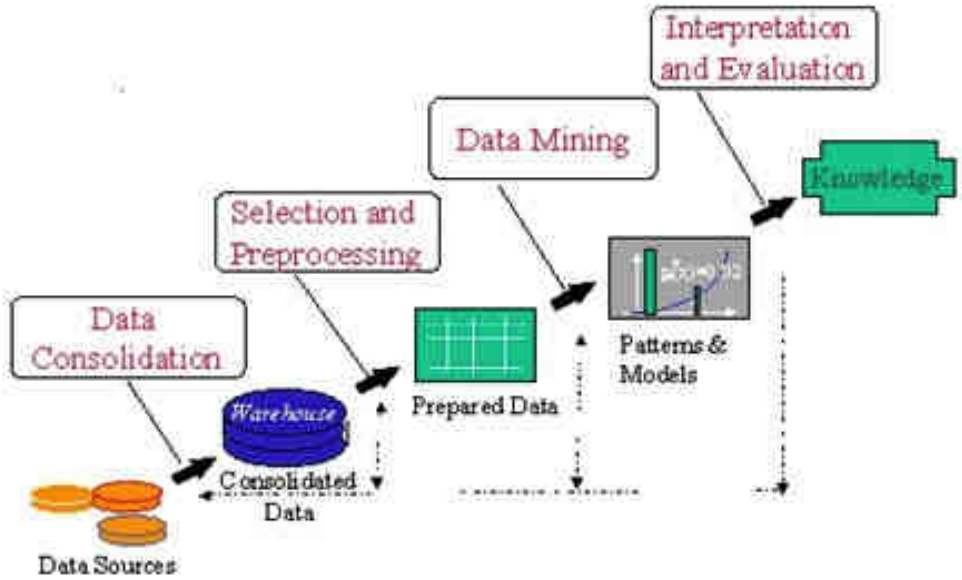
I dati derivanti da problemi scientifici sono aumentati in modo esponenziale negli ultimi anni, a differenza della potenza di calcolo che è aumentata linearmente.

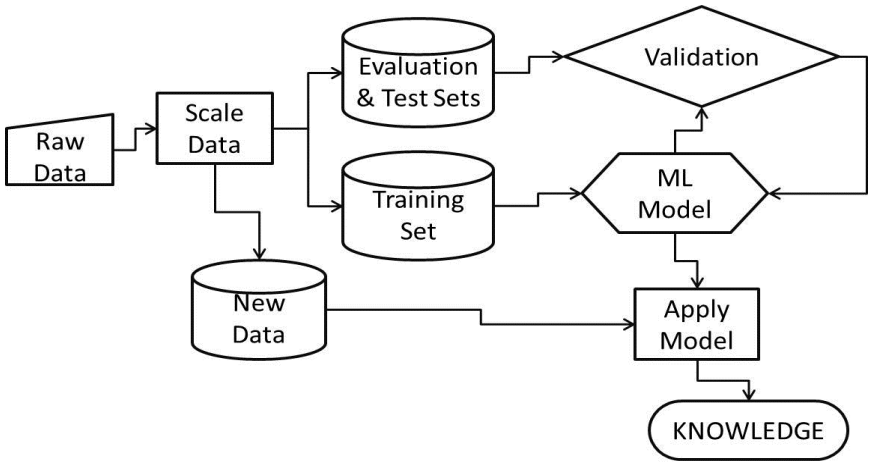
Questo «tsunami» di dati ha comunque portato alla nascita di nuove architetture per il data mining con calcolo ad alte prestazioni.

Le tecnologie e i servizi messi a disposizione da HPC, Grid e Cloud sono diventati strumenti fondamentali per la ricerca scientifica. Così come le tecniche di machine learning asservite al data mining.

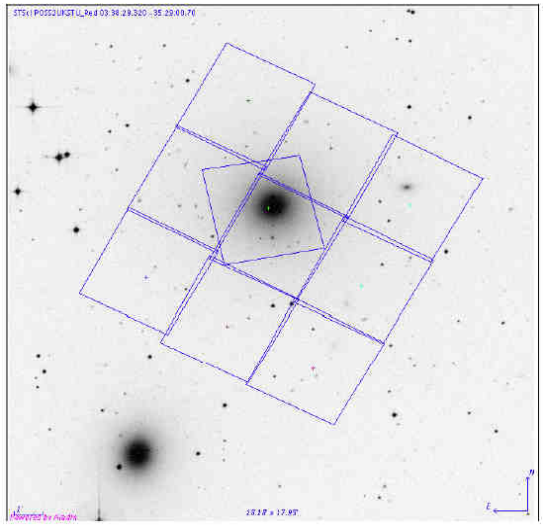






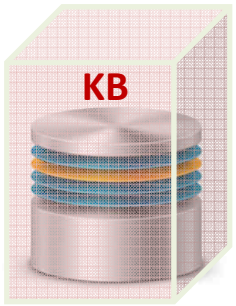


Casi d'uso in Astrofisica: Classificazione di ammassi globulari (GC)



Studio di popolazioni di ammassi globulari in galassie esterne a quella terrestre. In generale lo studio dei GC extra-galattici richiede l'uso di fotometria a grande campo e multi-banda. I GC in galassie lontane (diversi MParsec), appaiono come sorgenti luminose non risolte nelle immagini astronomiche da Terra (alta contaminazione da oggetti spuri).

- **7 parametri ottici (Magnitudini a varie aperture e FWHM immagine)** (feature 1-7);
- **4 parametri strutturali (raggi e brillanza)** (features 8–11);
- L'ultima colonna è la label della classe di appartenenza nota, **0 (no GC), 1 (yes GC)**;

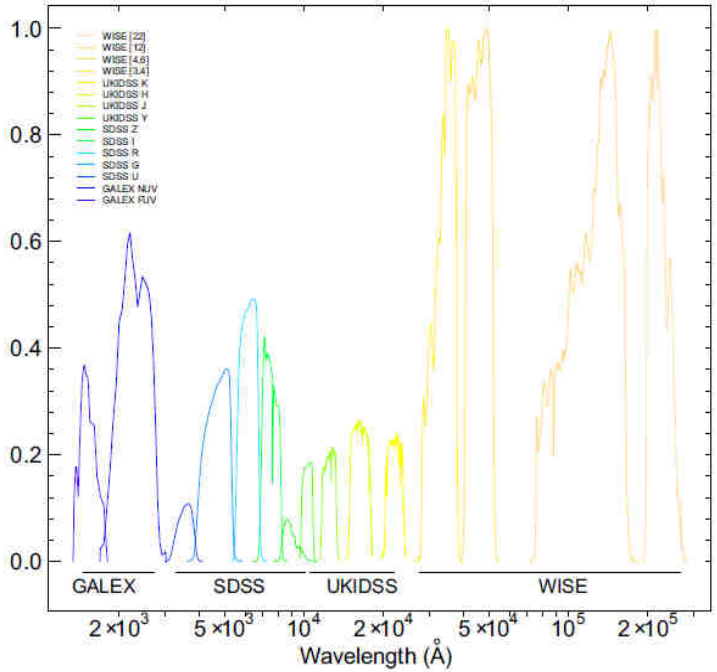


2100 training pattern ciascuno con 11 input + label classe

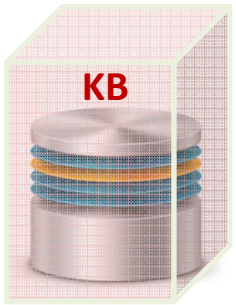
```
24.4753,26.7468,24.3789,0.0205,3.72,0.067,4.12,16.25,-0.1139,1.822,51.29,0
24.2342,26.5263,24.1632,0.0196,3.5,0.027,4.01,16.61,0.1321,1.856,35.38,0
23.1554,25.5964,23.1654,0.016,3.5,0.032,4.09,14.47,-0.3295,2.638,129.2,1
22.6316,25.3519,22.6808,0.0151,3.5,0.039,4.69,16.33,0.8065,5.002,80.45,1
22.4708,24.4951,22.4699,0.0216,3.5,0.066,3.45,12.81,-0.3912,-7.425,5.66,0
23.9033,27.5896,23.9168,0.0255,4.49,0.272,9.63,19.99,8.397,14.79,88.5,1
24.1972,26.4219,24.0978,0.0192,3.7,0.079,4.04,15.72,-0.1447,1.514,44.77,0
20.2423,22.1866,20.2963,0.017,3.5,0.03,3.23,6.68,-0.6999,-0.1492,1.899,0
23.5134,26.0983,23.511,0.0167,3.76,0.05,4.55,16.6,0.3777,4.75,105.8,1
…
```

Casi d'uso in Astrofisica: Regressione dei redshift fotometrici di Quasar

Si parla di redshift quando, nell'osservare lo spettro della luce emessa da galassie o quasar, questo appare spostato verso frequenze minori, se confrontato con lo spettro dei corrispondenti più vicini. Tale parametro può essere misurato o spettroscopicamente o in maniera fotometrica. Nel primo caso la stima è generalmente molto precisa, sebbene molto più onerosa in termini di quantità di tempo osservativo da impiegare con costosi strumenti astronomici.



- **11 parametri ottici (Magnitudini a varie bande)**;
- L'ultima colonna è il redshift spettroscopico noto;

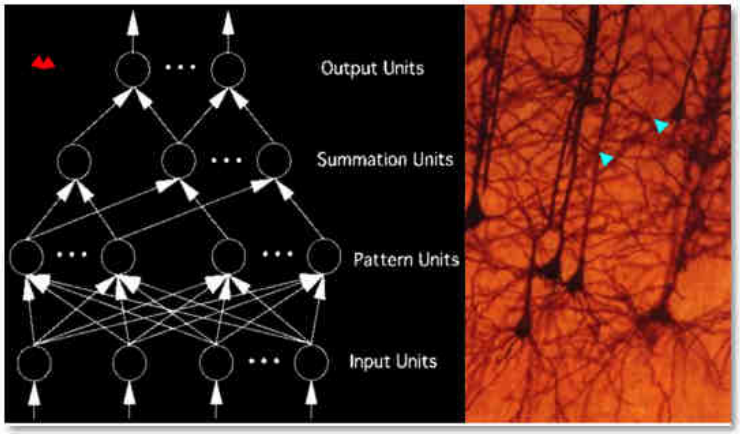


1499 training pattern ciascuno con 11 input + redshift spettr. noto

```
23.0451,22.3865,21.3956,20.6516,20.4471,0.6733,0.9188,0.8195,0.1957,3.77,0,0.557
22.6303,21.5566,20.427,19.5619,19.4067,1.0808,1.1,0.8933,0.1545,3.594,0,0.4399
23.7177,23.3864,22.6248,21.8677,21.729,0.286,0.7801,0.8125,0.0926,3.049,0,0.7268
24.9936,24.9237,24.3248,24.2199,24.0675,0.0466,0.1922,0.3948,0.1973,3.458,0,0.7148
23.6051,22.7975,21.7195,20.9121,20.7008,0.8584,0.9947,0.8709,0.2072,4.137,0,0.5319
22.5319,21.9514,20.975,20.146,19.9494,0.586,0.9315,0.8727,0.194,4.708,0,0.5667
22.0164,21.3014,20.556,19.8579,19.683,0.7201,0.708,0.7435,0.1828,5.8,0,0.4342
23.7532,23.1459,22.8099,22.4685,22.4437,0.6347,0.1696,0.4065,0.0269,3.455,0,0.3245
23.7147,22.8983,22.1731,21.6817,21.65,0.7623,0.594,0.5945,-0.0149,3.706,0,0.4362
22.3038,21.3017,20.3543,19.589,19.4856,1.0017,0.9237,0.797,0.1023,5.158,0,0.438
…
```

MLPGA Multi Layer Perceptron addestrato da un Algoritmo Genetico

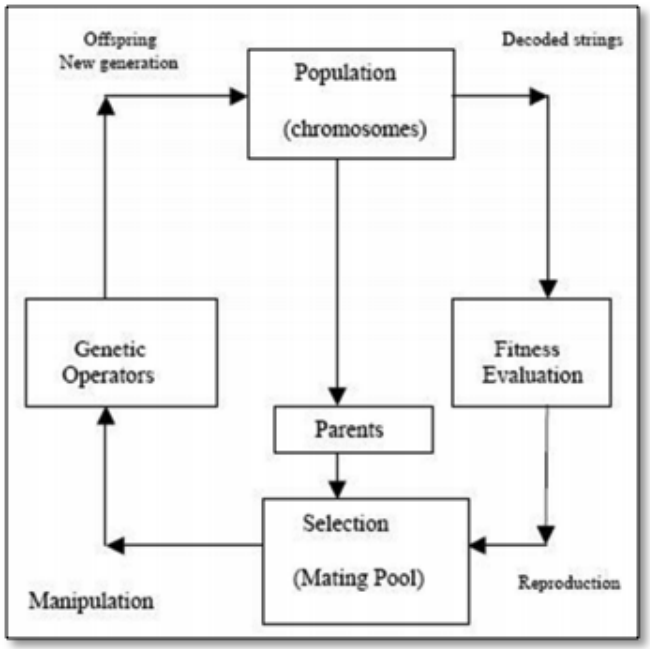
- Modello di machine learning supervisionato
- Prevede una fase di training con dati input + target noti
- Evoluzione dei pesi (sinapsi) basata su GA
- Capacità di generalizzazione su dati non noti
- Problemi di classificazione e regressione



Parametri in gioco

- Numero di livelli hidden della rete MLP;
- Numero di nodi del primo livello hidden;
- Numero di nodi del secondo livello hidden;
- Funzione di attivazione dei neuroni di ogni strato;
- Numero di iterazioni di training;
- Soglia di errore MSE di training;
- Distribuzione di probabilità per generazione popolazione iniziale;
- Numero di cromosomi della popolazione (numero di MLP per ogni ciclo di training);
- Funzione di selezione evolutiva;
- Numero di cromosomi di elitismo;
- Probabilità di crossover;
- Probabilità di mutazione;

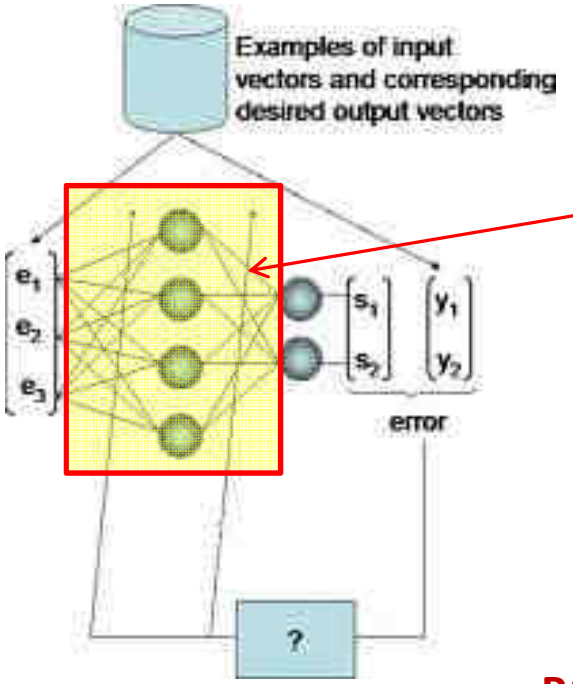
+

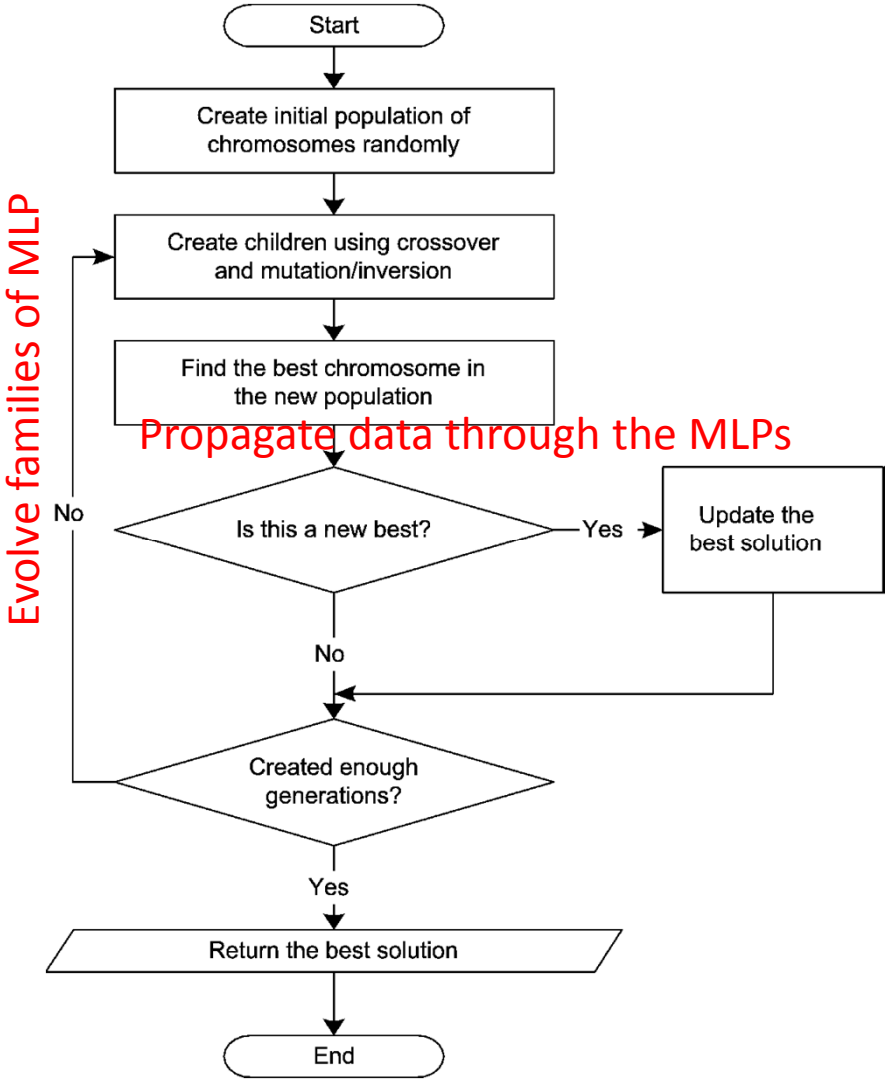


MLPGA seriale (CPU multi-core)

Il modello MLPGA, per costruzione, richiede un gran numero di iterazioni di training sui dati. Gli esperimenti di training sono quindi molto **lenti**. Inoltre il modello è intrinsecamente **non scalabile** rispetto ai dati input.

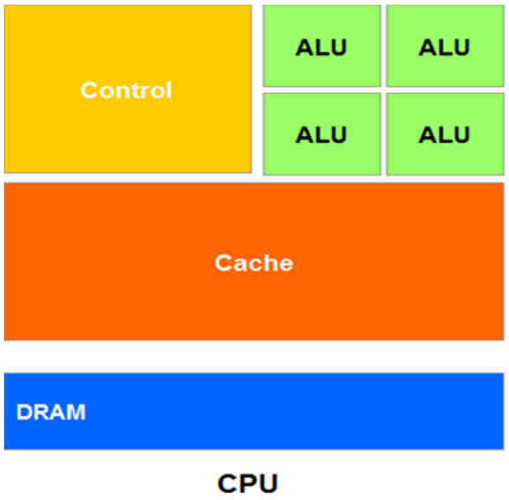
In particolare la fase computazionale più dispendiosa è la propagazione dei pattern input attraverso la rete. Ad ogni ciclo di training il GA evolve decine di reti MLP da testare sui dati.





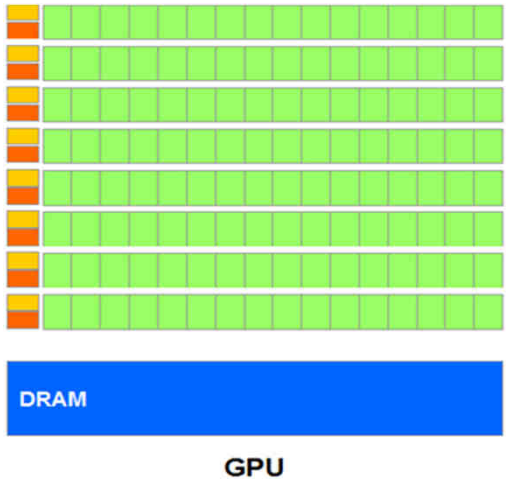
SOLUZIONE?
PARALLELIZZARE!!!

GPU+CUDA vs CPU => Fast MLPGA



Multi-core CPU

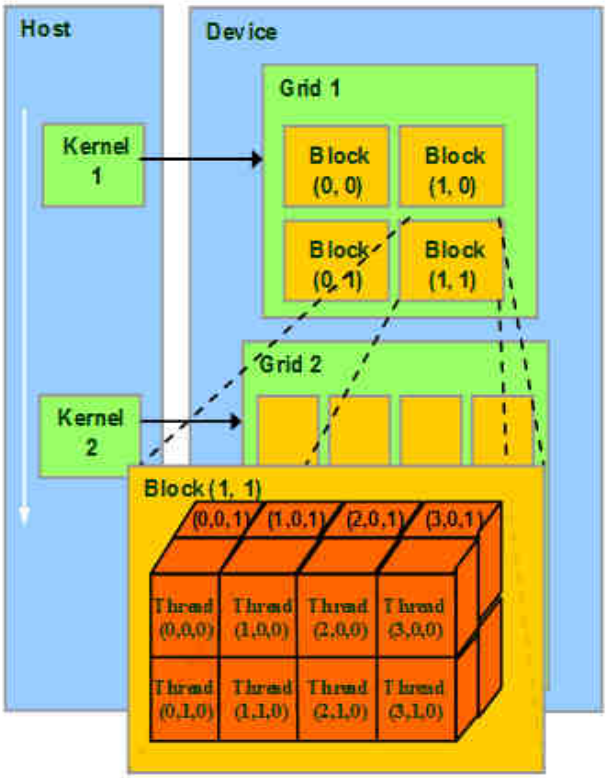
- Composta da pochi core, è progettata per massimizzare l'efficienza di codice sequenziale;
- Memoria cache di grandi dimensioni utile a ridurre i tempi di latenza per l'accesso ai dati o l'esecuzione di istruzioni complesse;
- Logica di controllo sofisticata, per la gestione avanzata del flusso delle istruzioni (pipelining e multi-threading ed esecuzione out-of-order).



Many-core GPU

- Composta da molti core (anche centinaia), è progettata per l'esecuzione di applicazioni parallele;
- Effettua contemporaneamente numerose operazioni semplici;
- Strutture di memoria con tempi di accesso spesso trascurabili;
- Logica di controllo più semplice (esecuzione sempre in-order);

GPU – Modello di Esecuzione



Esempio di una struttura a 5 dimensioni: *griglia* 2D (2x2) e *blocchi* 3D di *thread (4x2x2)*;

Un'applicazione CUDA è composta da parti seriali, eseguite dall'*host,* e da parti parallele (i *kernel)*, eseguite dal *device;*

Un *kernel* è definito come una *grid*, decomposta a sua volta in *blocchi* assegnati, sequenzialmente, ai vari *SM*;

Ogni *blocco* è formato da *thread,* l'unità computazionale fondamentale, ed eseguono tutti la stessa istruzione (Single-Program Multiple-Data);

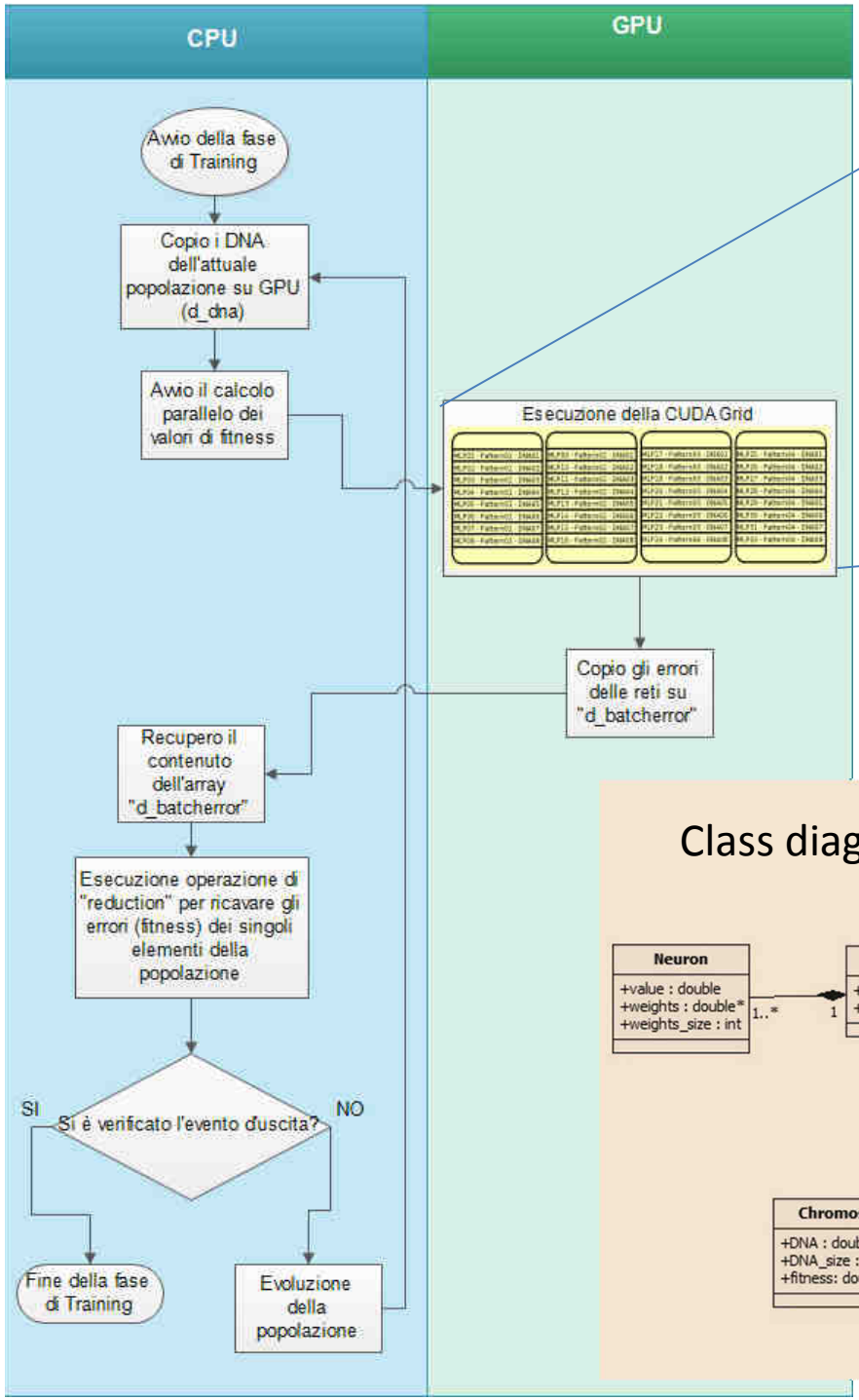
Ciascun *thread* può appartenere ad un solo *blocco,* ed è identificato da un indice univoco per tutta la durata del *kernel*;

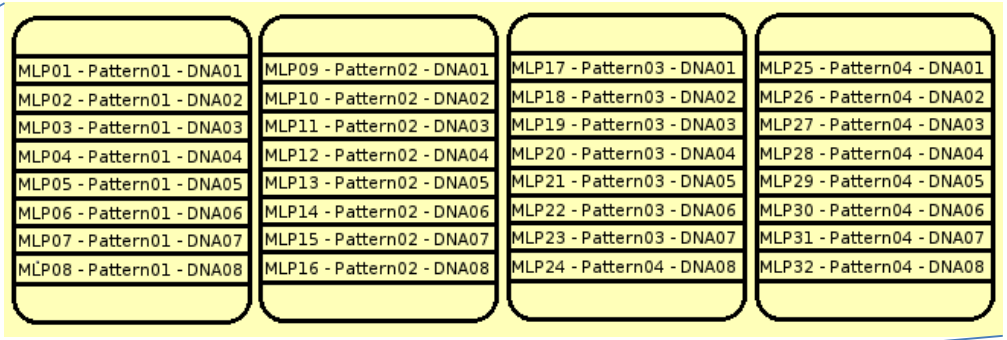
I *kernel* sono eseguiti sequenzialmente tra loro mentre *block* e *thread* sono eseguiti in parallelo;

Unità di memoria:

- *Registri*: read/write dal proprio thread;
- *Memoria locale*: read/write dal proprio thread;
- *Shared Memory*: read/write dai thread di uno stesso blocco;
- *Constant Memory* (cached): read-only da thread di una stessa grid;
- *Global Memory*: letta e scritta da tutti i thread nello stesso grid;

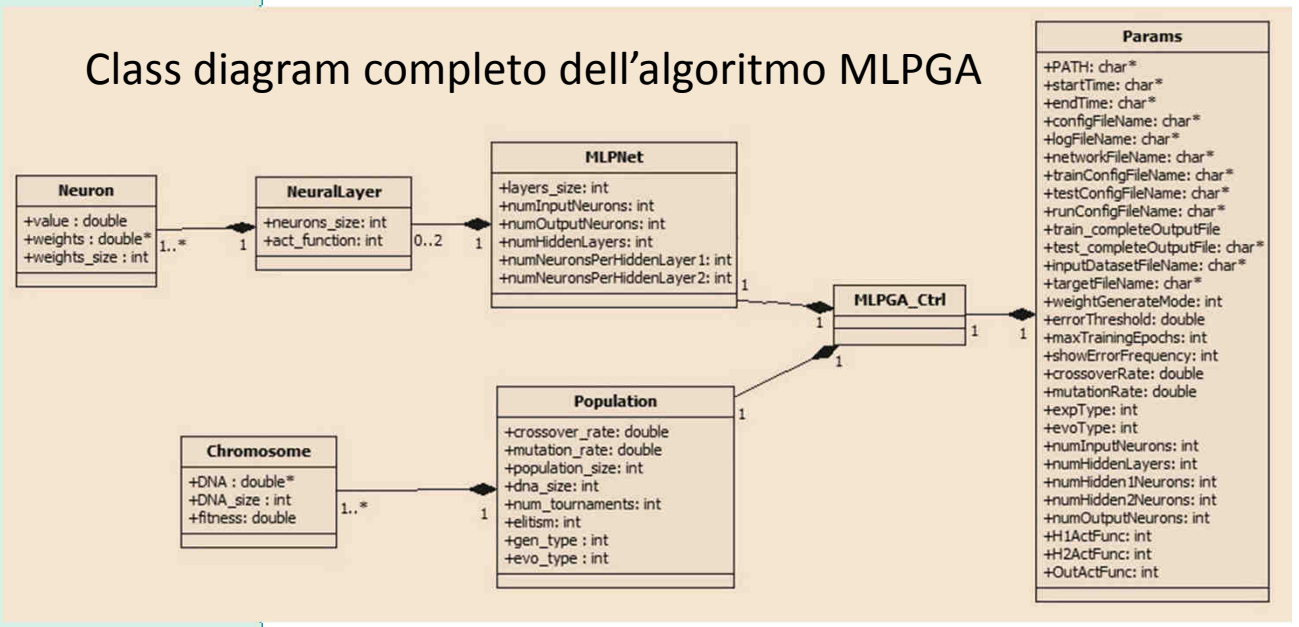
Flusso FMLPGA





Su GPU ad ogni iterazione si creano e processano tutte le N reti MLP (cromosomi del GA) in parallelo (CUDA GRID)

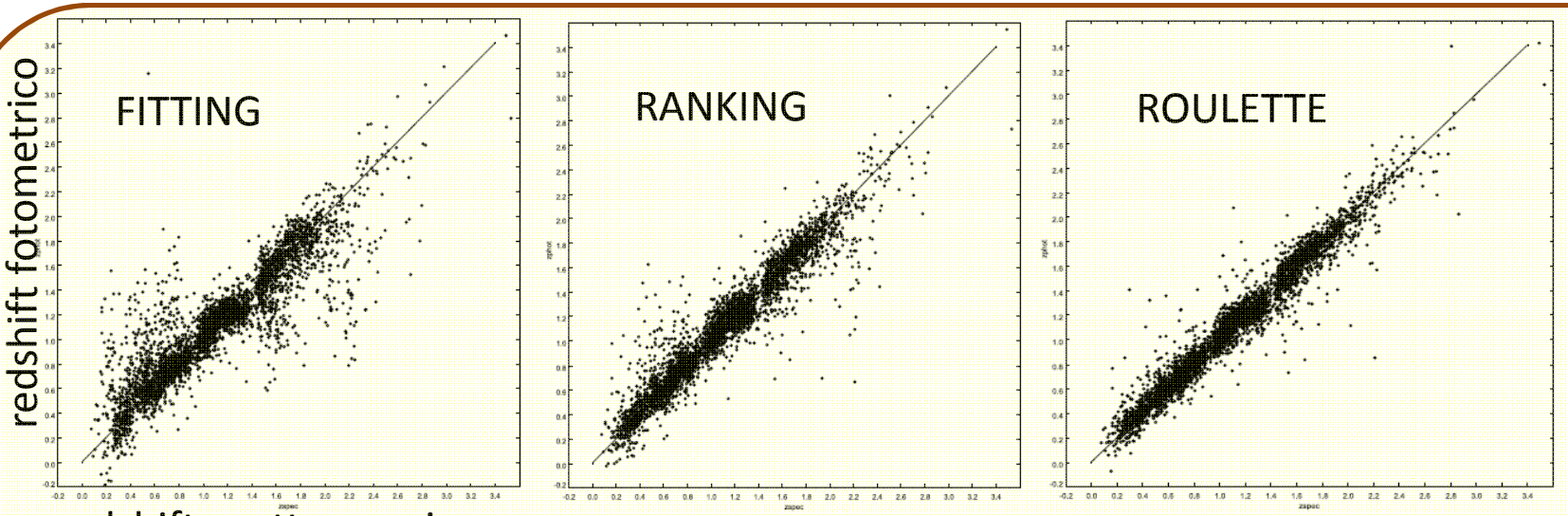
Class diagram completo dell'algoritmo MLPGA



TEST FMLPGA – Qualità scientifica

Per la classificazione, il risultato migliore è ottenuto con la FITTING. La qualità è superiore ai metodi astrofisici tradizionali.

Valutazione Qualità risultati di test Classificazione					
Funzione	iterazioni	Classe target	Accuratezza [%]	Purezza [%]	Contaminazione [%]
Roulette	50000		94.89		
		Not GC		97.01	2.99
		GC		92.64	7.36
Ranking	50000		93.18		
		Not GC		95.32	4.68
		GC		91.45	8.55
Fitting	50000		96.24		
		Not GC		98.41	1.59
		GC		94.67	5.33



Per la regressione, gli scatter plot zphot-zspec mostrano un ottimo risultato con la ROULETTE. La deviazione standard in questo caso è inferiore a 0.02

CONFRONTO MLPGA vs FMLPGA - Tempo di esecuzione

I test qualitativi consentono di fissare i migliori parametri del modello. Il confronto tra le due implementazioni, è stato condotto omologando i due dataset (1000 pattern e 11 features) e variando il numero di iterazioni e la funzione di selezione del GA.

Piattaforma di sviluppo: CPU Intel Xeon E5506 a 2.13 GHz 8-core e NVIDIA TESLA S2050 (4 GPU), con 448 core per ogni GPU
Test bench: CPU Intel i7 a 3.4 GHz quad-core e NVIDIA GPU GTX 460 a 336 core

In sostanza, la versione parallela del modello MLPGA permette di ottenere un guadagno medio in termini di velocità di esecuzione pari a 8x, rendendo efficiente e scalabile l'algoritmo.