GPU

- GPGPU
- Fermi architecture
- Parallel
- Cheap

GPU - costs

Data from NVIDIA	Performance	Performance/$	Performance/Kw
2xIntel Xeon X5550 7,000$	80 GFLOPS	11 GFLOPS/k$	146 GFLOPS/kW
2xTesla 2050 2xIntel Xeon X5550 11,000$	656,1 GFLOPS	60 GFLOPS/k$	656 GFLOPS/kW

Jaguar delivers 1,770 TFLOPS and consumes more than 7 MW of power. Nebulae delivers 1,270 TFLOPS and consumes 2.55 MW. That makes it twice as power-effcient as Jaguar.

Mimesis 2010

Aprile: architettura NVIDIA-Fermi, doppia precisione e ECC
Maggio: Progetto "Centro di Calcolo Scientifico"
Giugno: Presentazione del progetto a NVIDIA
possibilità di diventare centro di eccellenza
Luglio: NVIDIA fornisce una S2050 gratuitamente
Settembre: installazione di URANIA
Ottobre/Novembre: test con algoritmo sui plasmi
Dicembre: definizione del progetto DBwrap
Gennaio: applicazione del progetto a Fermi
Tosti e Dubois approvano preliminarmente
sottomissione NASA

Fermi

problemi di performance in analisi dati con sistemi attuali

due livelli:

- dati grezzi
- dati filtrati: miliardi di fotoni, 6 caratteristiche (32 byte)

DBwrap: interporre fra il db e l'interfaccia utente un sw che utilizzi array di GPU come cache e in grado di eseguire selezioni ed analisi statistiche sui dati

Fermi

- data selection
  - multi-dimensional range: time, energy, direction, quality
- livetime cube
  - prepare a cube matrix for convolution
- exposure map
  - effective area integrated in time for a given position
- likelihood
  - probability of obtaining the data given an input model

all steps are ineherently parallel!

>>> data mining

Astroinformatics

- data models
- data transformation and normalization methods
- indexing techniques
- information retrieval and integration methods
- knowledge discovery methods
- content-based and context-based information representations
- consensus semantic annotation tags,
- taxonomies
- ontologies
- and more

these enable data mining, information retrieval and fusion, and knowledge discovery from huge astronomy datasets

Data Mining

ASTRONOMY
fast growth in CCD detector size and sensitivity

today: 1 TB of data per day
near future: 30 TB of data daily

	10 years ago	now	10 years from now
average dataset size	100 Gbytes	10 Tbytes	10 Pbytes

Data Mining

How we can identify objects in scientific data sets and extract features or descriptors representing these objects

Astronomy

- Massive size of the data
- Data collected at different frequencies
- Real-time analysis
- Noisy data
- Temporal data analysis

Data Mining - Selection

- Simple dimensional range selection
- Statistical algorithm (likelihood)
- Clustering (kmean, ANN, GA, SVM, tree)

Data Mining - Riduction

Matrix: data item X values of the features for that item
dimension of the problem : number of features (N)
point in feature space of dimension N

dimension reduction: the identification of key features used to represent the objects or data items in a data set

- Principal Component Analysis
- Random projections
- Multidimensional scaling
- FastMap
- Self-organizing maps

Data Mining - Finding Patterns

- clustering
- classification
- regression
- association rules
- tracking
- anomaly detection

Table 1. Advantages and disadvantages of well-known machine learning algorithms in astronomy. These algorithms, and others, are described in more detail in §§2.4.1–2.4.4.

Algorithm	Advantages	Disadvantages
Artificial Neural Network	Good approximation of nonlinear functions Easily parallelized Good predictive power Extensively used in astronomy Robust to irrelevant or redundant attributes	Black-box model Local minima Many adjustable parameters Affected by noise Can overfit Long training time No missing values
Decision Tree	Popular real-world data mining algorithm Can input and output numerical or categorical variables Interpretable model Robust to outliers, noisy or redundant attributes Good computational scalability	Can generate large trees that require pruning Generally poorer predictive power than ANN, SVM or kNN Can overfit Many adjustable parameters
Support Vector Machine	Copes with noise Gives expected error rate Good predictive power Popular algorithm in astronomy Can approximate nonlinear functions Good scalability with number of attributes Unique solution (no local minima)	Harder to classify > 2 classes No model is created Long training time Poor interpretability Poor at handling irrelevant attributes Can overfit Some adjustable parameters
Nearest Neighbor	Uses all available information Does not require training Easily parallelized Few or no adjustable parameters Good predictive power	Computationally intensive No model is created Can be affected by noise and irrelevant attributes
Expectation Maximization	Gives number of clusters in the data Fast convergence Copes with missing data Can give class labels for semi-supervised learning	Can be biased toward Gaussians Local minima