

The Terabyte Analysis Machine

Gabriele Garzoglio, Apr 2001

- Introduction
- The Cluster Environment
- The Distance Machine
- Scales
- Performance Tests
- Conclusions
- Links

Introduction

The **Terabyte Analysis Machine** is a cluster designed to allow research into the use of **large scientific databases**.

Our initial program is to explore efficient **parallel database access (KDI)** on **large data set (ALDAP)**, via **re-indexing** and **re-partitioning** techniques, into a “**griddified**” environment.

The Cluster Environment

The Distance Machine will manage in one year ~10 nodes using **Condor** as a batch system.

Possible SX database transfers will be managed via **GDMP**

The client-server applications of the Distance Machine framework communicate via **CORBA** interfaces.

The Distance Machine

As an example of a database re-indexing and re-partitioning in the TAM framework, the Distance Machine is an implementation of an analysis engine for the k^{th} nearest neighbor search.

The Distance Machine uses the SDSS (SX) database (John Hopkins University)

It extracts the relevant data from SX using the Tag Extractor utility developed by Koen Holtman (Caltech)

Scales

The SDSS database will have:

- **15 Tbytes image database**
- **500 Gbytes catalog database**
- 250 million objects
- **100 million galaxies** (example of the search for cluster of galaxies)
- 2.5 Kbytes object size
- 500 byte tag object size (example of tag with about 50 parameters of interest)
- **50 Gbytes tag database**

Moving the tag database:

- 50 Gbytes at 10 Mbytes/sec \Rightarrow 1.5 hours to transfer over network
- 50 Gbytes at 30 Mbytes/sec \Rightarrow 0.5 hours to write to disk

Dimensionality of the problem:

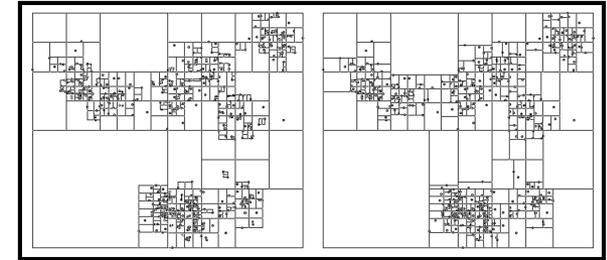
- 120 parameters (total) per object
- 5 parameters used for cluster finding: ra, dec, g-r, r-i, and i'

The Terabyte Analysis Machine

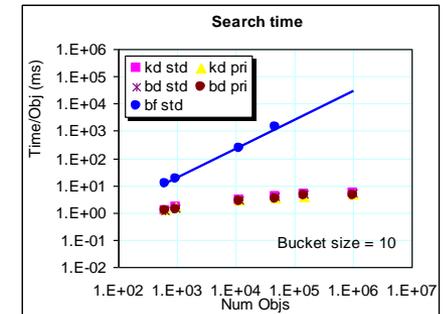
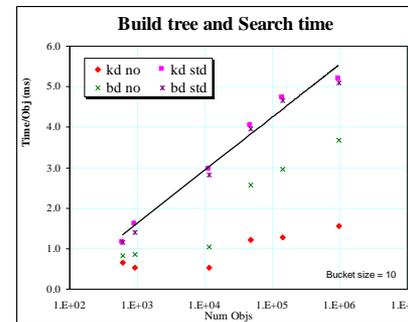
Gabriele Garzoglio, Apr 2001

Performance Tests

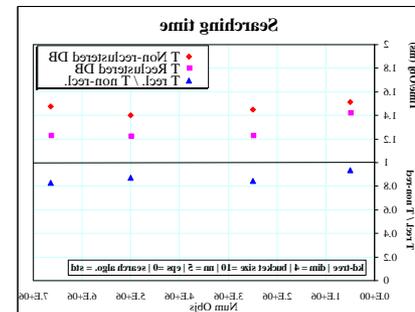
The Distance Machine finds Nearest Neighbors using the kd-tree technique, accessing directly the database.



This technique is $O(N \log N)$



Re-clustering techniques improve performance



Conclusions

The Distance Machine implements a framework to study how re-indexing and re-partitioning techniques increase performance on parallel access to large data sets.

It is a core engine for cluster finding analysis on the SDSS data.

The Distance Machine can be used as a test bed for the GriPhyN tools (virtual NN data sets).

It's flexible enough to be used on data other than astronomy (e.g. “out of the box”: study of Nearest Neighbors techniques for particle tagging)

Links

- <http://projects.fnal.gov/act/kdi.html>
- [http://sdss.fnal.gov:8000/~annis/astro
Compute.html](http://sdss.fnal.gov:8000/~annis/astroCompute.html)