

The Terabyte Analysis Machine

Jim Annis, Gabriele Garzoglio, Jun 2001

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

The Terabyte Analysis Machine

Jim Annis, Gabriele Garzoglio, Jun 2001

Introduction

The goal of the **Terabyte Analysis Machine (TAM)** project is to explore and integrate emerging and stable technologies into an efficient computing environment for physicists and astronomers working on data intensive scientific analysis.

The studies in progress are in the areas of **Knowledge and Distributed Intelligence (KDI)** and **Large Distributed Archives** in Astronomy and Particle Physics (**ALDAP**)

The Cluster Environment

TAM is a cluster of (currently) 5 Linux boxes
(2xPIII 650 MHz, 1 GB RAM, 72 GB IDE disk)
connected to a 1 TB raid box via fibre channel

It uses GFS as File System.

Condor as Batch System.

The Distance Machine Framework

The Distance Machine framework allows specific analysis programs to transparently access large databases, making best use of sophisticated specific algorithms for indexing and partitioning the database.

As an example we have implemented an analysis engine for the k^{th} nearest neighbor search: ANN.

ANN 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

The Technology

ANN acts as a server that provides nearest neighbors to an analysis client using **corba** interfaces.

Due to the **client/server** architecture, ANN works in a **distributed** environment.

On TAM, many ANN servers can run in parallel, each one dealing with a different parameter space, topology and portion of the sky.

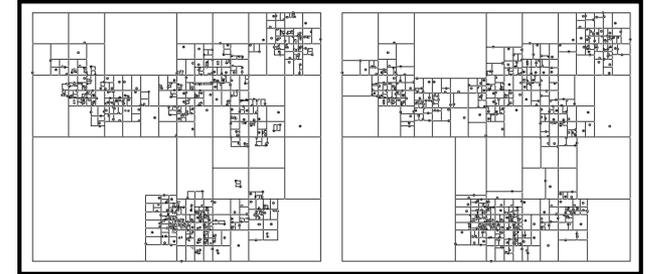
Astrotool can be used as an analysis client for ANN.

The Terabyte Analysis Machine

Jim Annis, Gabriele Garzoglio, Jun 2001

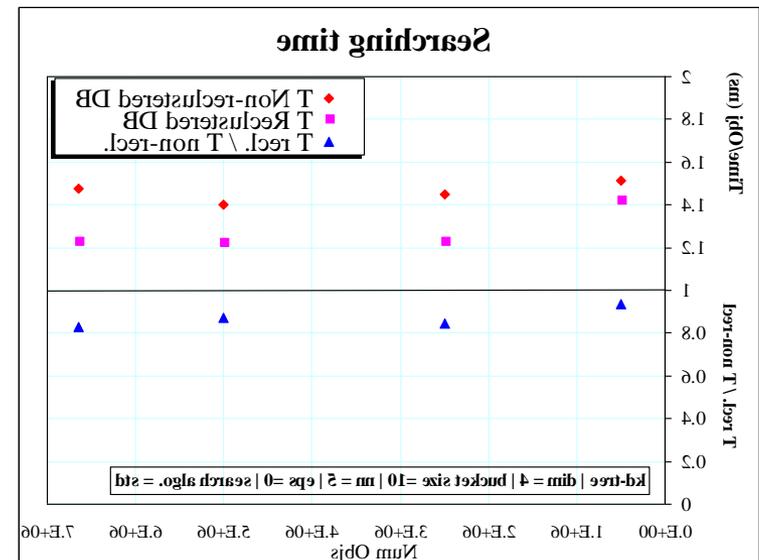
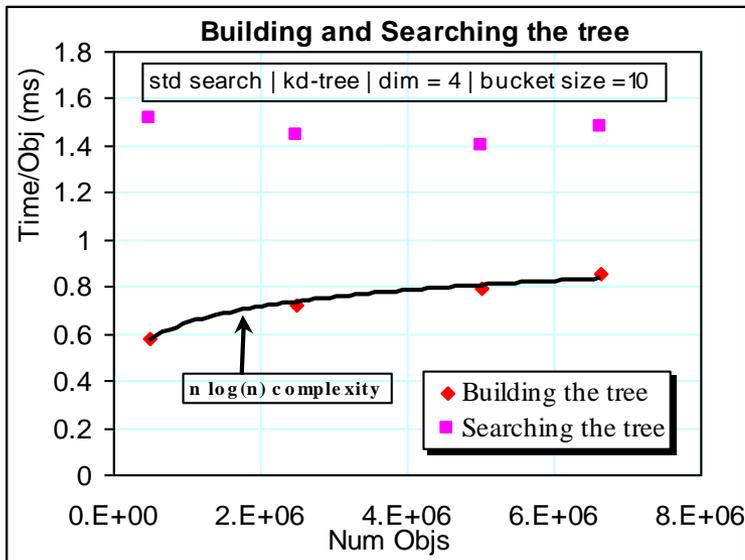
Performance Tests

The Distance Machine finds Nearest Neighbors using the kd-tree technique.



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

Re-clustering techniques improve performance



The Terabyte Analysis Machine

Jim Annis, Gabriele Garzoglio, Jun 2001

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)

The Distance Machine

- **ANN/Objectivity** (Garzoglio/Annis)
 - Adaptive smoothing kernel
 - Similarity search
 - Cluster search in high dimensions
 - Current use: spectroscopic cluster catalog (the Finger of God catalog)
- **Next project:** (summer student Peretz Partensky)
 - link DM/maxBcg to Kent image base
 - Run cluster finder on a objects on demand (Griphyn...)
 - Show nearest neighbors on images
- **Next instrument:**
 - kd-tree based range search
 - Range search the basis for most environmental studies

The Terabyte Analysis Machine

- The compute cluster

- Intended to hold tsObj files on disk
- Intended as Fermi/SDSS science analysis machine

- Performs well as cluster finder
- Non-EAG users: TAG, UC, Penn State, CMU

- Purchase in the works to double the machine size
- Upgrade cluster software to use QCD farm experience

- Suffering from a lack of computer professional support:
 - Only Annis and Garzoglio can bring the machine fully up
- Dan Yocum (EAG) will be brought up to speed/ownership
- GFS tutorial at Sistina soon