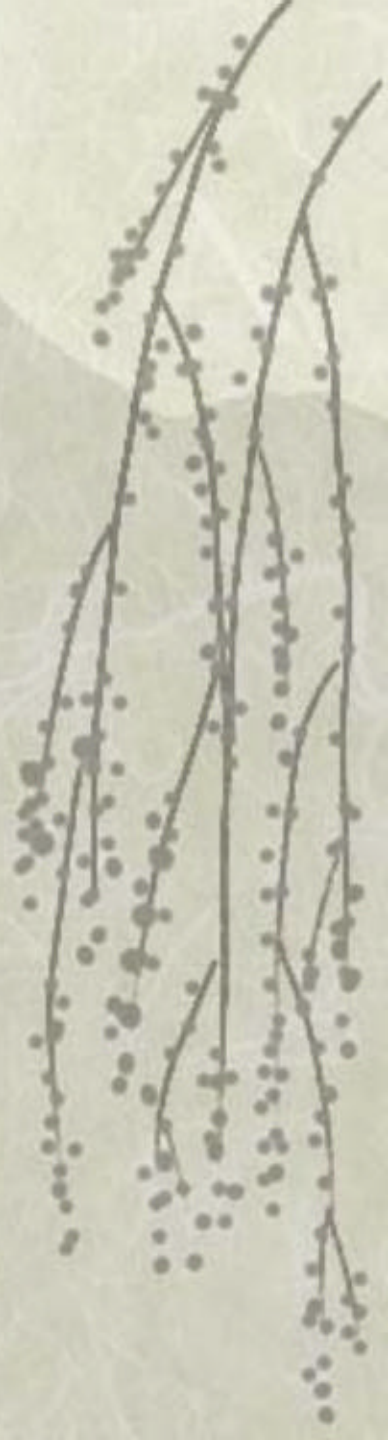
Databases for Astronomers - 2

Database Technology
and Pan-STARRS

6/27/03 mfw

Syllabus

- Database issues for Pan-STARRS
- Very large databases & related issues
  - Scalability
  - Performance
- Current Examples:
  - SDSS – Alex Szalay
  - Teraserver- Jim Gray
  - Distributed Databases



Where can Pan-STARRS use DB technology?

- Engineering & hardware diagnostics
- Data integrity & validation
- Archive maintenance
- Science data access

  - P/S will generate too much data, too quickly, for manual examination (*STSCI)*
  - A central access point/mechanism is needed to maintain control over data use & quality (*SDSS)*

Two kinds of information:

- Metadata: what was done, when, how
  - Exposure details (physical conditions)
    - Basically image header information, + system diagnostics
  - Processing steps
    - Record of what resulted as data transited the pipeline
  - => a transaction journal of what was done
    - Allow automated detection of subtle errors (using trends)
    - Centralized, standardized access interface to this record
- Science products
  - Random access mechanism for image-type data
    - Essentially a robust index
  - Catalog science
    - Data mining – correlations across large object sets

Pan-STARRS Issues:

- Metadata: highly dynamic (new entries for every exposure)
  - Indexing overhead
  - Rapid growth
  - Need rich analysis tools (computed trends, statistical quality control)
- Science data: large size (both image objects & catalog entries)
  - Access throughput (search, indexing, reload, updates)
  - Storage volume
  - External query issues

Very Large Database Issues

- Size/scalability:
  - Number of tables/records
  - Size of records
  - memory addressing , caching, etc
- Speed:
  - Computational: index traversal, comparisons
  - Record access – I/O bottlenecks
  - Index creation overhead
- Reliability:
  - Failure recovery, complexity

Systems design decisions

- One DB or several?
  - More DB = more overhead (=$$$)
  - Different dynamics argue for different designs
- Distributed storage or distributed servers?
  - Depends on future technology (I/O speed)
  - True distributed server technology still research topic

Some current examples

- SDSS skyServer – Alex Szalay & Jim Gray
- USGS Teraserver – Jim Gray
- Distributed Object database
- USNO-B as a database

Sloan Skyserver

- SQL server on Wintel, Microsoft support
- HP Itanium (64bit), 28GB memory,etc
- 350MB/sec data processing speed claimed
- 80GB database, 14M obj, 50K spectra
- Currently ~10m hits/yr, 40-50K/day total
- skyserver web site
- Further info, schema, "20 queries"

An interesting view of Astronomical DB issues

- Jim Gray - Databases Meet Astronomy

SkyQuery -A federated database server

- Combined NVO-style interface to 4 online databases (SDSS,2MASS, FIRST, INTWFS)
- Requires user to create SQL manually
- Basically a demo
- skyquery.html

USGS TerraService

- http://terraservice.net/
- 3.3 terabyte earth image store/server
- Goal to provide reasonable access to large data store over "slow" connection
- 4 node HP/Windows 2000 cluster
- 8-way MP servers, SAN
- 2TB/node (3 active) data + metadata (=6TB!)

More Terraserver details

- Database partitions run separate DBMS
- Simple schema, row contains metadata + pointer to image blob
- Search interface based on specifying image center
- Data loaded in subtiles of original 40kx40k satelite source
- Data is rescaled dynamically on request

Distributed Database Systems

- Current research at Uppsala University (Sweden, Witold Litwin)
- Scalable system of database processors (multi-computer cluster)
- SDDS overview, SDDS-2000-overview