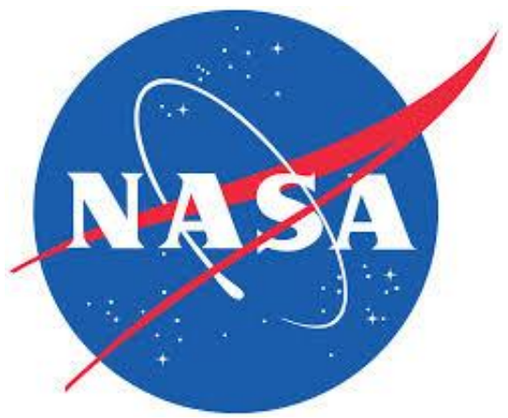
Instantaneous Archives

Walter Landry



Performance is a Feature

- Improving performance increases the breadth and depth of research users can do.
- We can never be too fast

How Fast is Fast Enough?

- Human reaction time: 200 ms

How Fast is Fast Enough?

- Human reaction time: 200 ms
- But many of these web interfaces talk to machines (IDL scripts, astro.py, TOPCAT, ...)

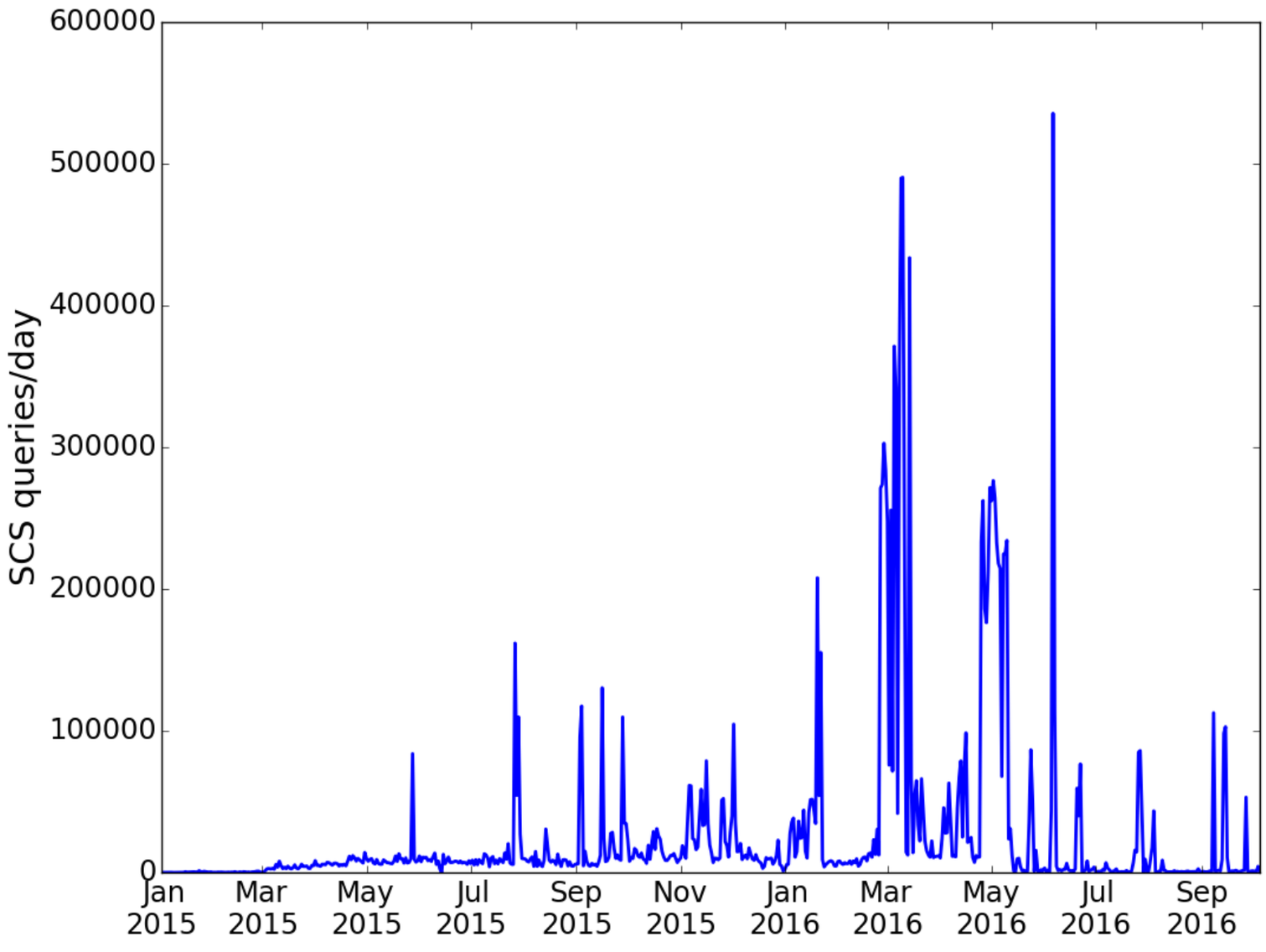
How Fast is Fast Enough?

- Human reaction time: 200 ms
- But many of these web interfaces talk to machines (IDL scripts, astro.py, TOPCAT, ...)
- Network latency to the end user
  - Other side of the world: 1000+ ms
  - Inside the building: 0.3 ms

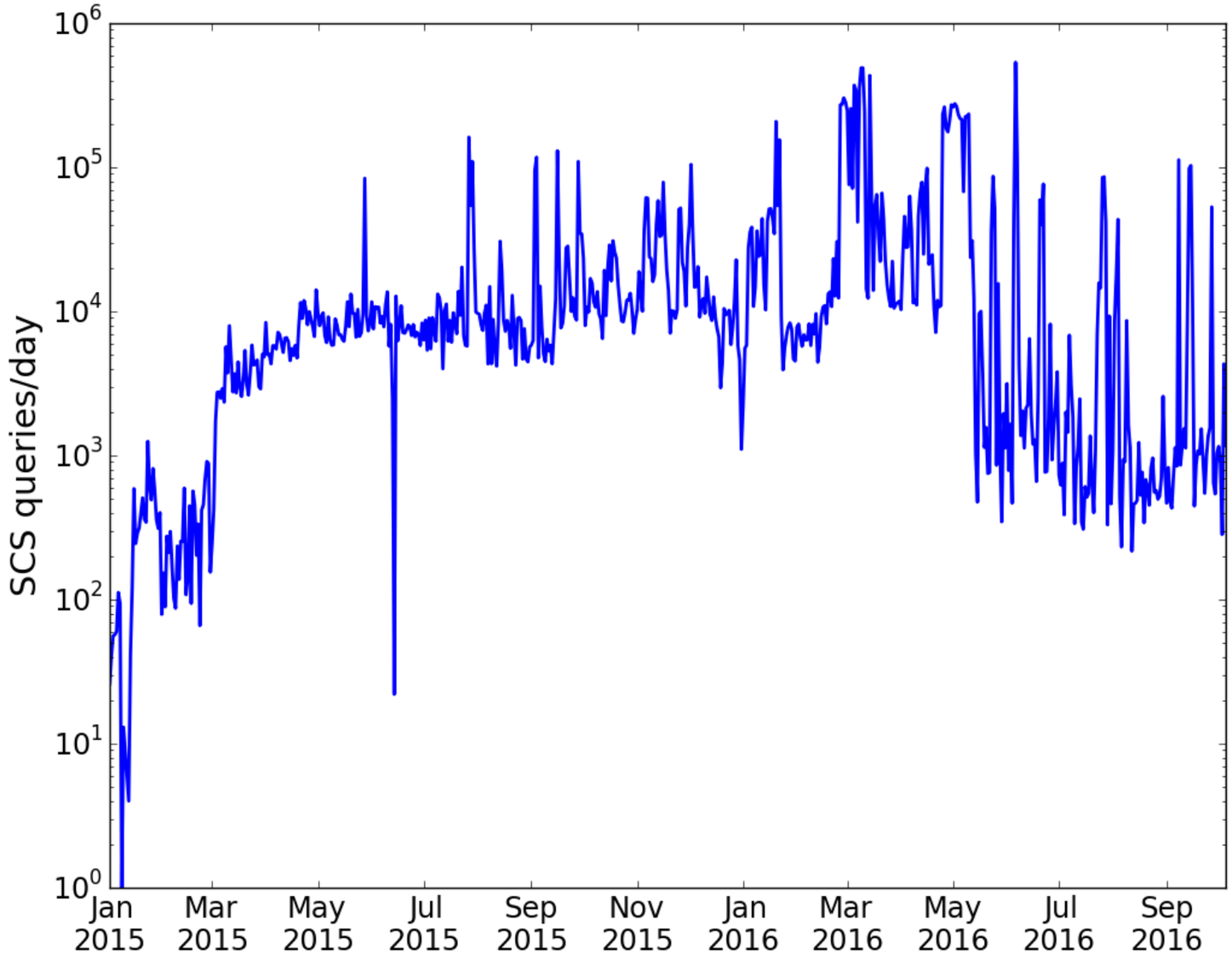
How Fast is Fast Enough?

- Human reaction time: 200 ms
- But many of these web interfaces talk to machines (IDL scripts, astro.py, TOPCAT, ...)
- Network latency to the end user
  - Other side of the world: 1000+ ms
  - Inside the building: 0.3 ms
- You also want to be fast enough so that users do not block each other.

Simple Cone Search Traffic



Simple Cone Search Traffic



Simple Cone Search Traffic

- Averaged over a day, the peak rate works out to 6 queries/second.
- We have 16 cores on our webserver, so if queries take more than about 3 seconds, we are going to fall behind.

What to Optimize For?

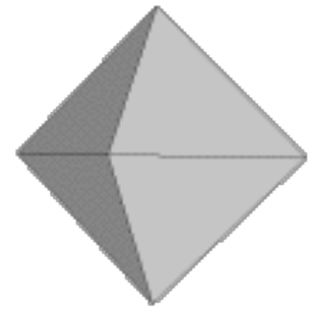
- Simple Cone Search is our most popular service by far.
- Most queries are small (<60 arcsec).
  - Cross-match the old-fashioned way
- The increased traffic is not a one-time incident. It is a consequence of making our data available through Virtual Observatory (VO) protocols (and especially TOPCAT).

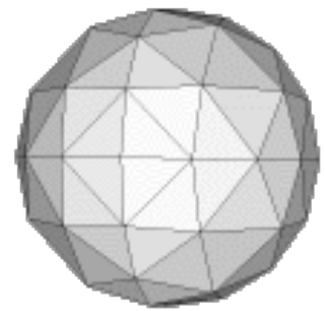
What to Optimize For?

- So we will be looking at small cone searches on the 2MASS All-Sky Point Source Catalog
  - 470 million rows
  - Big enough for realistic benchmarks
  - Small enough to copy around test machines.
- Current queries take about 1 second ... 

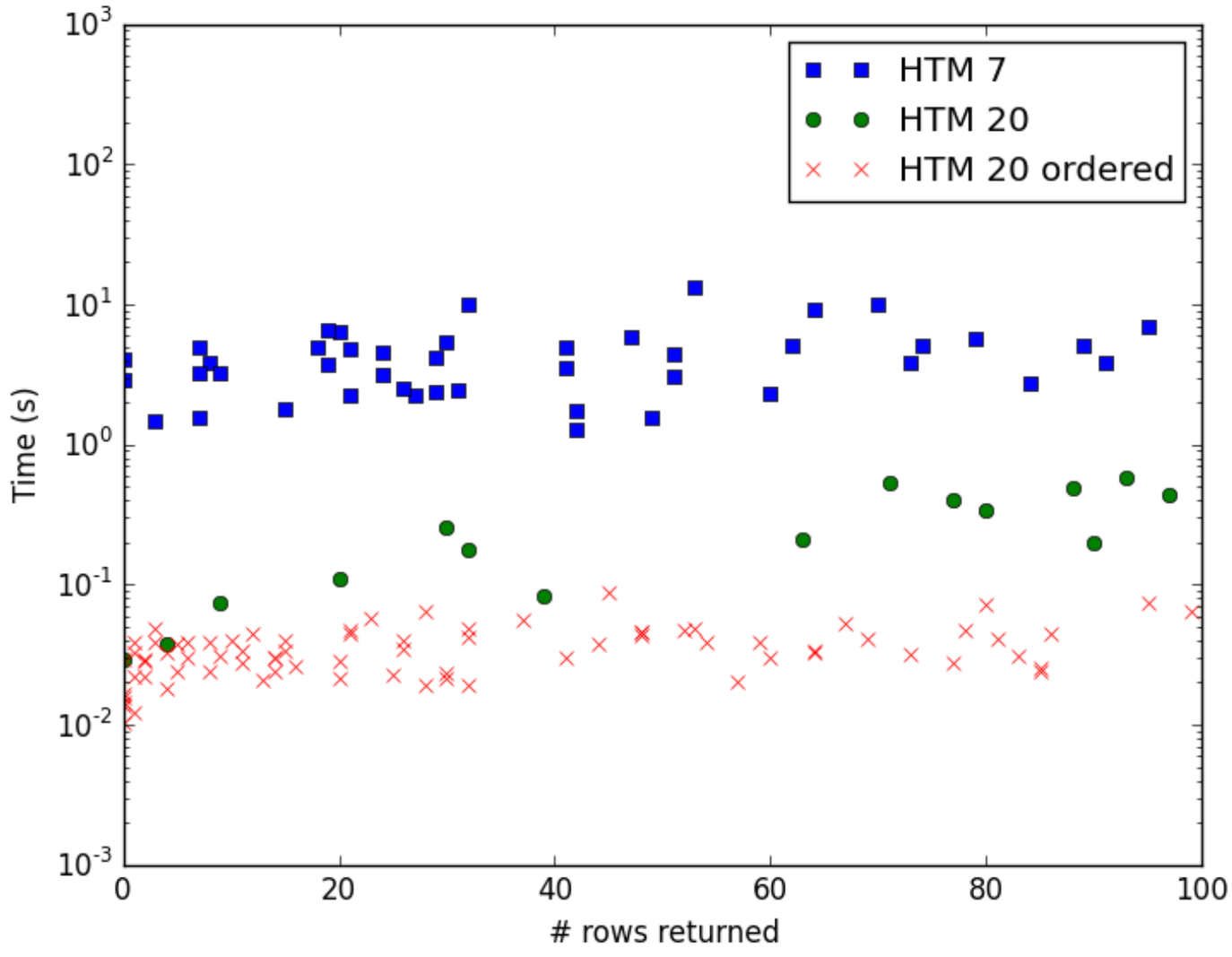
Database Bottleneck?

- We tile the sky with a Hierarchical Triangular Mesh (HTM).
- This gives us a number (HTM ID) for each part of the sky.
- HTM ID's next to each other tend to be close together spatially.
- For good performance, need to
  - Choose the right depth of hierarchy
  - Order the rows by HTM ID





2MASS Query Times Using a Direct SQL Connection



Alternate Spatial Indexing

- Cube-sphere (q3c)
  - Comparable to HTM for cone searches
  - Maybe faster for cross matches?

Alternate Spatial Indexing

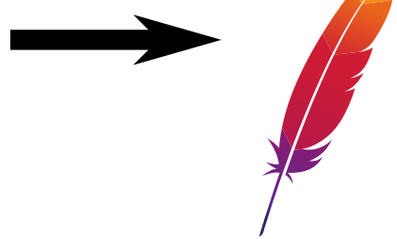
- Cube-sphere (q3c)
  - Comparable to HTM for cone searches
  - Maybe faster for cross matches?
- Healpix (h3c)
  - Slightly faster (10%) than q3c for cone searches

Alternate Spatial Indexing

- Cube-sphere (q3c)
  - Comparable to HTM for cone searches
  - Maybe faster for cross matches?
- Healpix (h3c)
  - Slightly faster (10%) than q3c for cone searches
  - Pathologically bad for polygons

Alternate Spatial Indexing

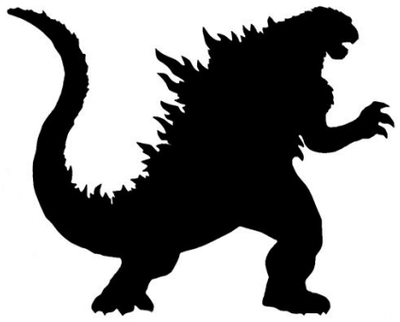
- RTree (PostGIS)
  - Handles polygon-polygon intersections
  - Fast enough for relatively small image tables (50 million rows)

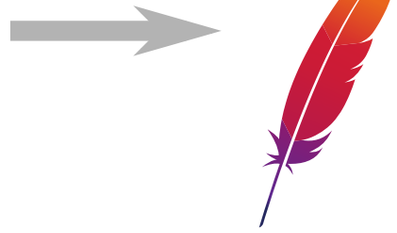


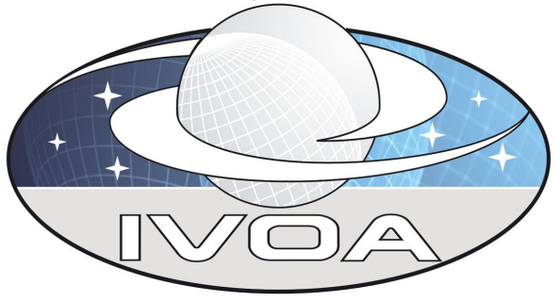
Architecture

What does 2MASS know about (210.802, 54.349)?

IRSA Architecture

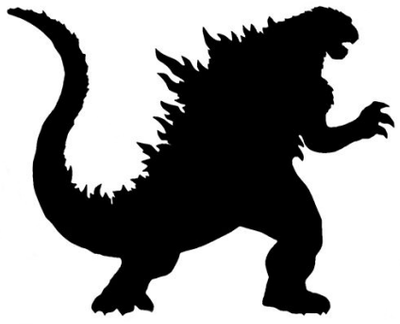




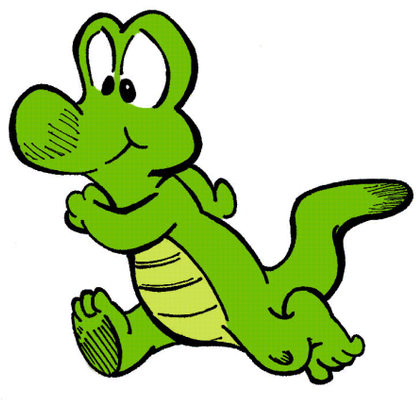


Apache starts a new cgi-bin process (IVOA) to answer the query

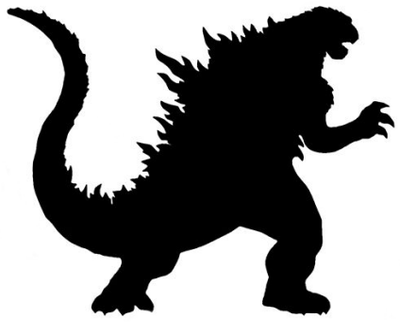
IRSA Architecture

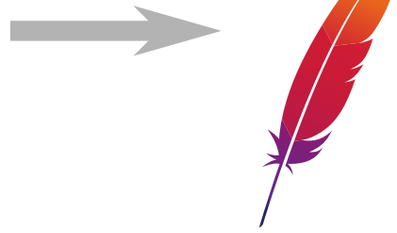


IVOA starts another process (Isisql) and asks for metadata about the 2MASS table

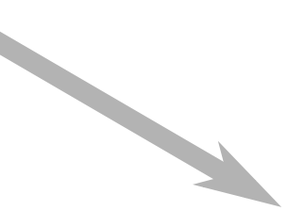


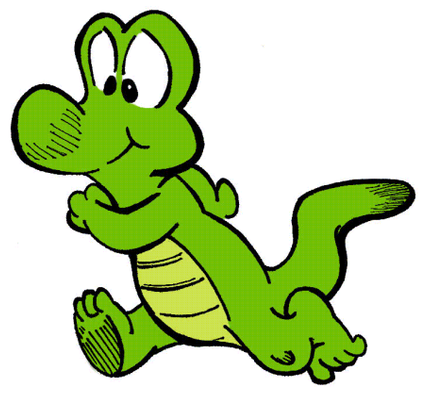
IRSA Architecture



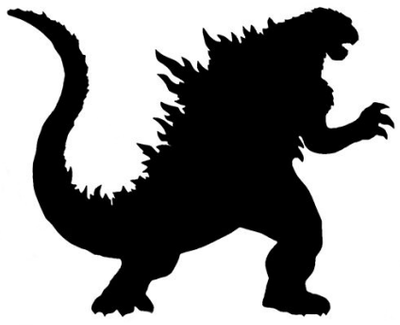


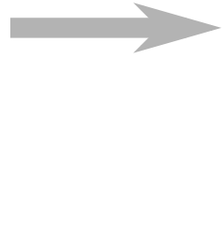
Isisql passes along the query to Oracle





IRSA Architecture

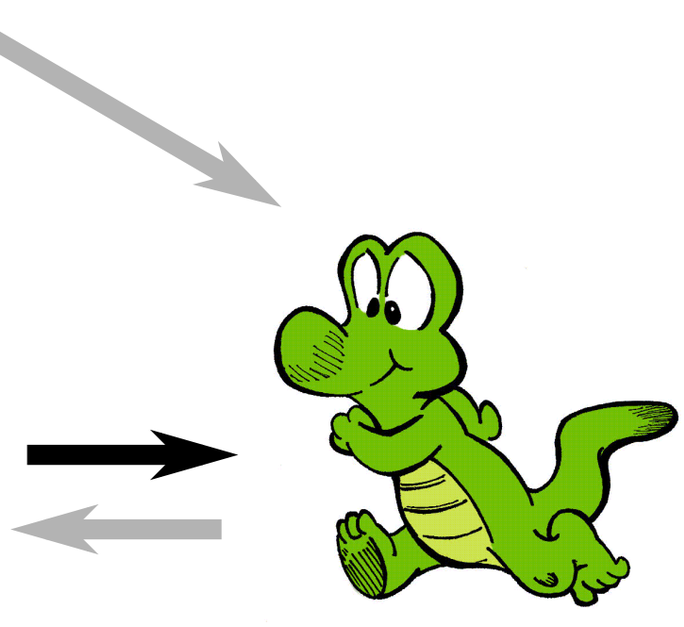




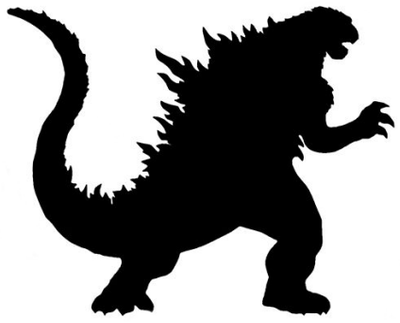
THE APACHE® SOFTWARE FOUNDATION

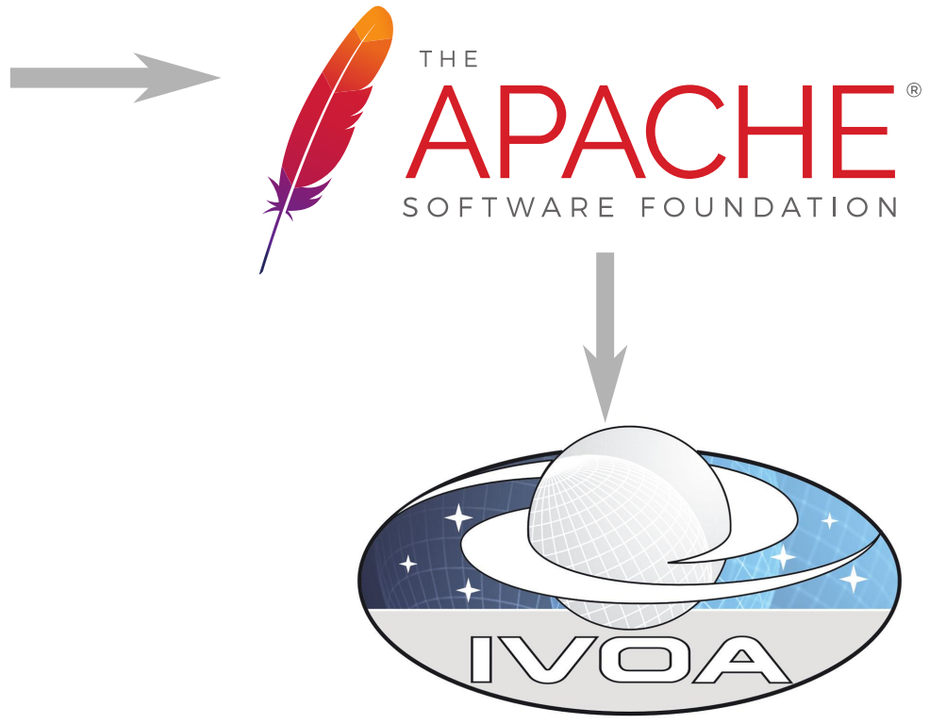
Oracle returns an answer

ORACLE® DATABASE



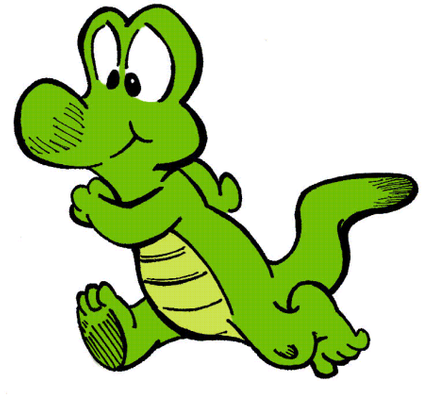
IRSA Architecture



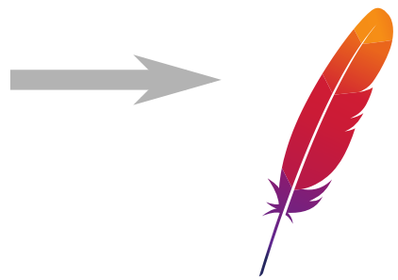


Isisql returns the metadata to IVOA

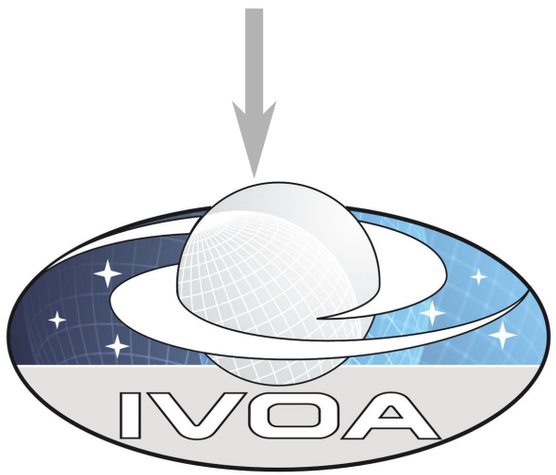
ORACLE
DATABASE



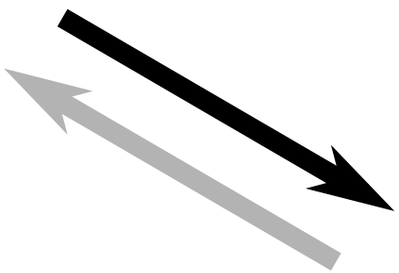
IRSA Architecture



THE APACHE® SOFTWARE FOUNDATION

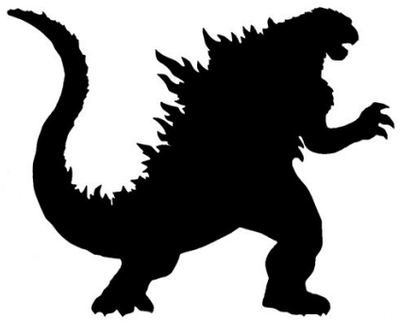


IVOA sends the real query to Isisql



ORACLE® DATABASE

IRSA Architecture



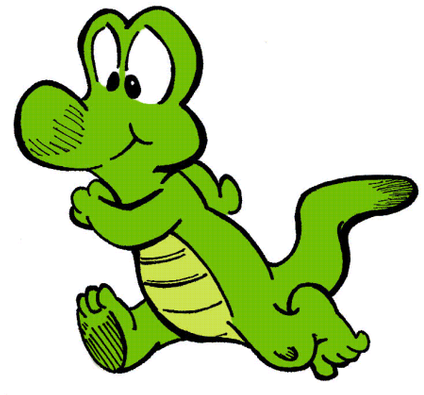
Isisql runs the query and stores it in a networked temp directory

THE APACHE SOFTWARE FOUNDATION

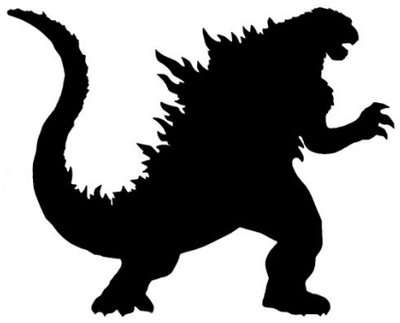
IVOA

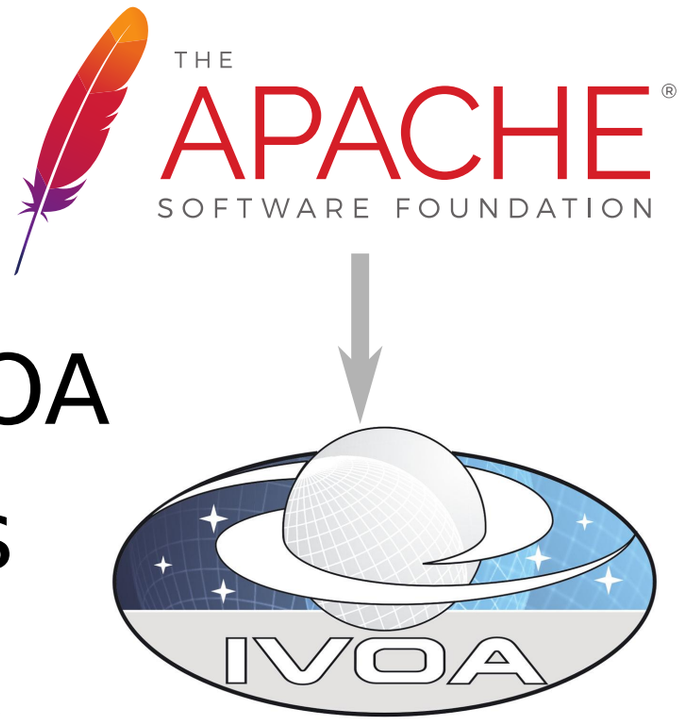
ORACLE DATABASE

ZFS
The Last Word in File Systems

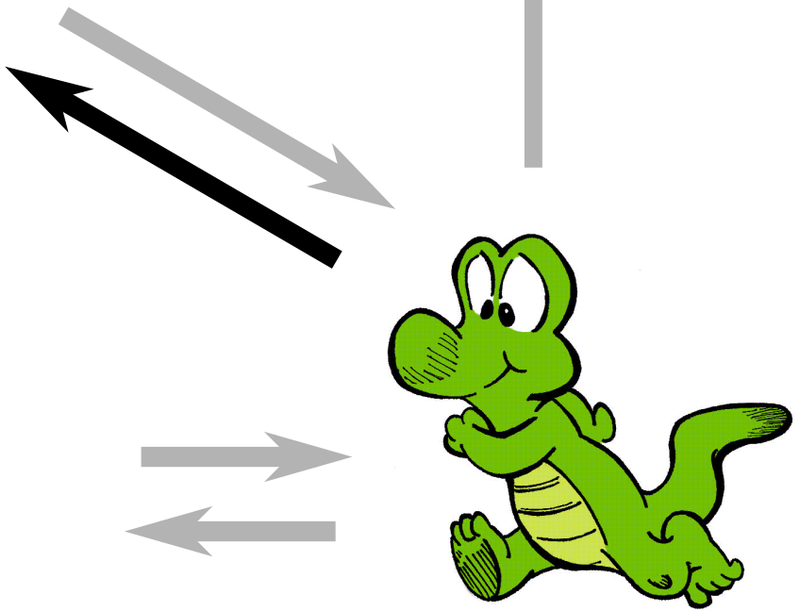


Architecture

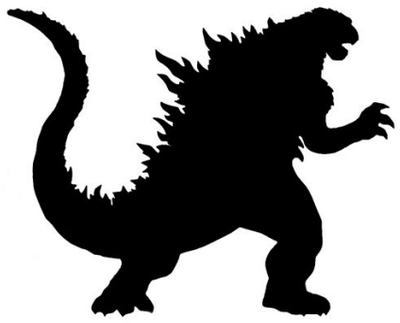




Isisql notifies IVOA that the result is ready, and IVOA retrieves it.



IRSA Architecture



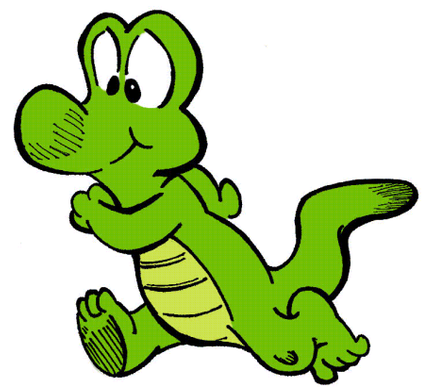
IVOA converts the result to the desired format (e.g. FITS, VOTable, HDF5) and sends it to Apache

THE APACHE SOFTWARE FOUNDATION

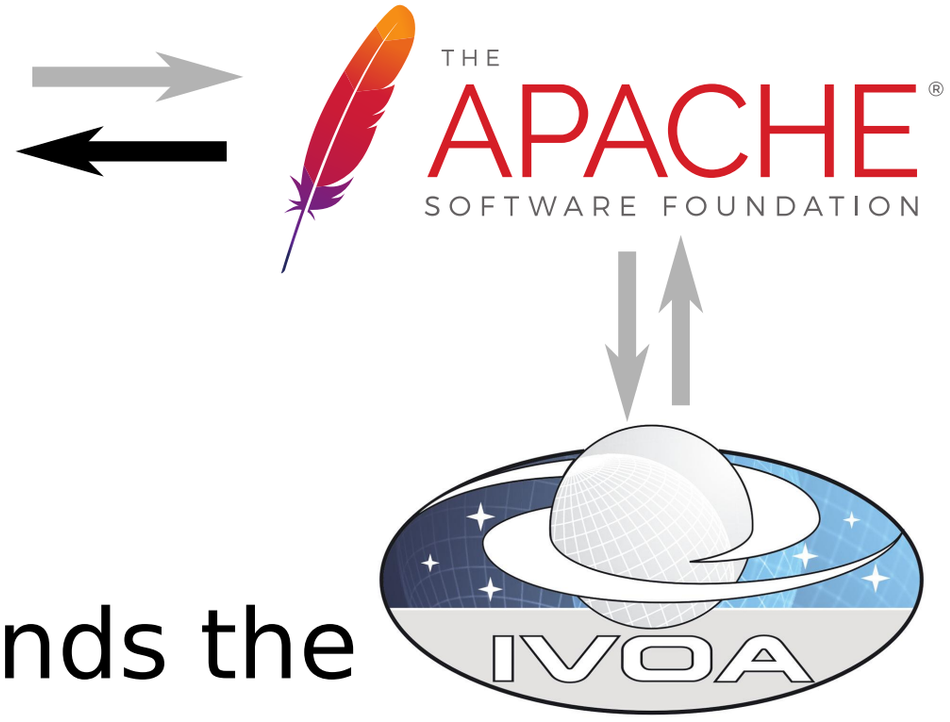
IVOA

ORACLE DATABASE

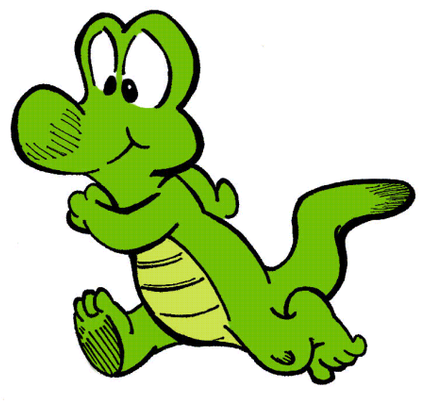
ZFS
The Last Word in File Systems



IRSA Architecture



Apache sends the result to the user



What takes Time?

1/3: Starting up processes

What takes Time?

1/3: Starting up processes

2/3: Making new connections to the database

What takes Time?

1/3: Starting up processes

2/3: Making new connections to the database

1/30: Running the user's query

Optimization Strategy

- Instead of using CGI programs with Apache, run an embedded, multi-threaded webserver to only service VO and other API calls.
  - This service can then keep a pool of database connections.
  - Do as much as possible in memory. For small cone searches, never write to the filesystem.

Choosing an Embedded Webserver

- Our stack is all C++ and C.
- That gives us a few options

Choosing an Embedded Webserver

- Our stack is all C++ and C.
- That gives us a few options too many

POCO
Webem
H2O
Mimosa
Proxygen
Kore
nxweb
Casablanca
Pistache
Served
libnghttpd2_asio
libmicrohttpd
Boost ASIO
Pion
Mongoose
cpp-netlib
libevent
node.native
Boost.http
CivetWeb
Simple Web Server
libhttpserver
libapache-mod-raii

Benchmarking the Webserver

- For each candidate, we wrote a simple service that waits 10 seconds and returns "Hello World".
- Then we used httperf to make 1024 concurrent connections.
- httperf -hog --num-conns=1024
  --num-calls=1 --rate=128
  --server localhost --port 8080

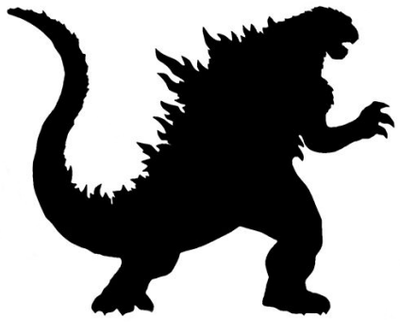
Benchmarking Results

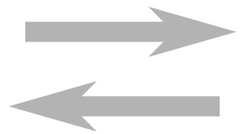
- This simple exercise broke a lot of the libraries
- Others just did not perform well with the load
- For this and other reasons, we chose libhttpserver
  - https://github.com/RipcordSoftware/libhttpserver
  - It is a C++ wrapper around libmicrohttpd, which is a well established, fast, embedded webserver library.

Connecting to the Database

- Once again, there are many libraries for connecting to databases
- Isisql connects to Oracle over ODBC.
- We chose the native Oracle C++ connection library (OCCI).
  - Probably the best performance???
  - Has pools built-in.

New IRSA Architecture



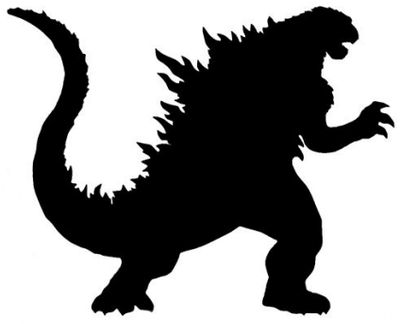


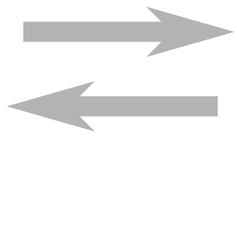
THE APACHE® SOFTWARE FOUNDATION

ORACLE®

DATABASE

New IRSA Architecture





- No process startup
- All in memory

Response Times

- 1024 Queries
- Randomly distributed throughout the sky
- 5" < Search Radius < 36"

Response Times

- 1024 Queries
- Randomly distributed throughout the sky
- 5" < Search Radius < 36"

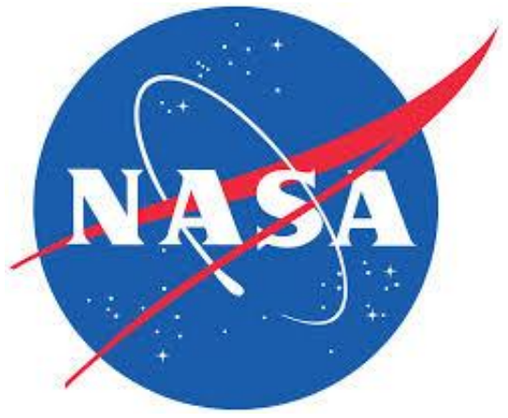
	Old	New
median	704 ms	79 ms

Conclusions

- Early results look promising
- Further significant gains would require progress in our database technology
  - In-memory databases? Plausible for 2MASS, not for all of our busy catalogs.

We're Hiring!

https://jobs.caltech.edu/postings/5580



Conclusions

- This only looked at the latency of small cone searches.
- Optimizing complex queries that require a table scan requires things like distributed databases.
  - LSST's Qserv
  - CitusDB
- Optimizing concurrent access requires fleets of frontends and databases.