Web and Grid Services from Pitt/CMU

Andrew Connolly
Department of Physics and Astronomy
University of Pittsburgh

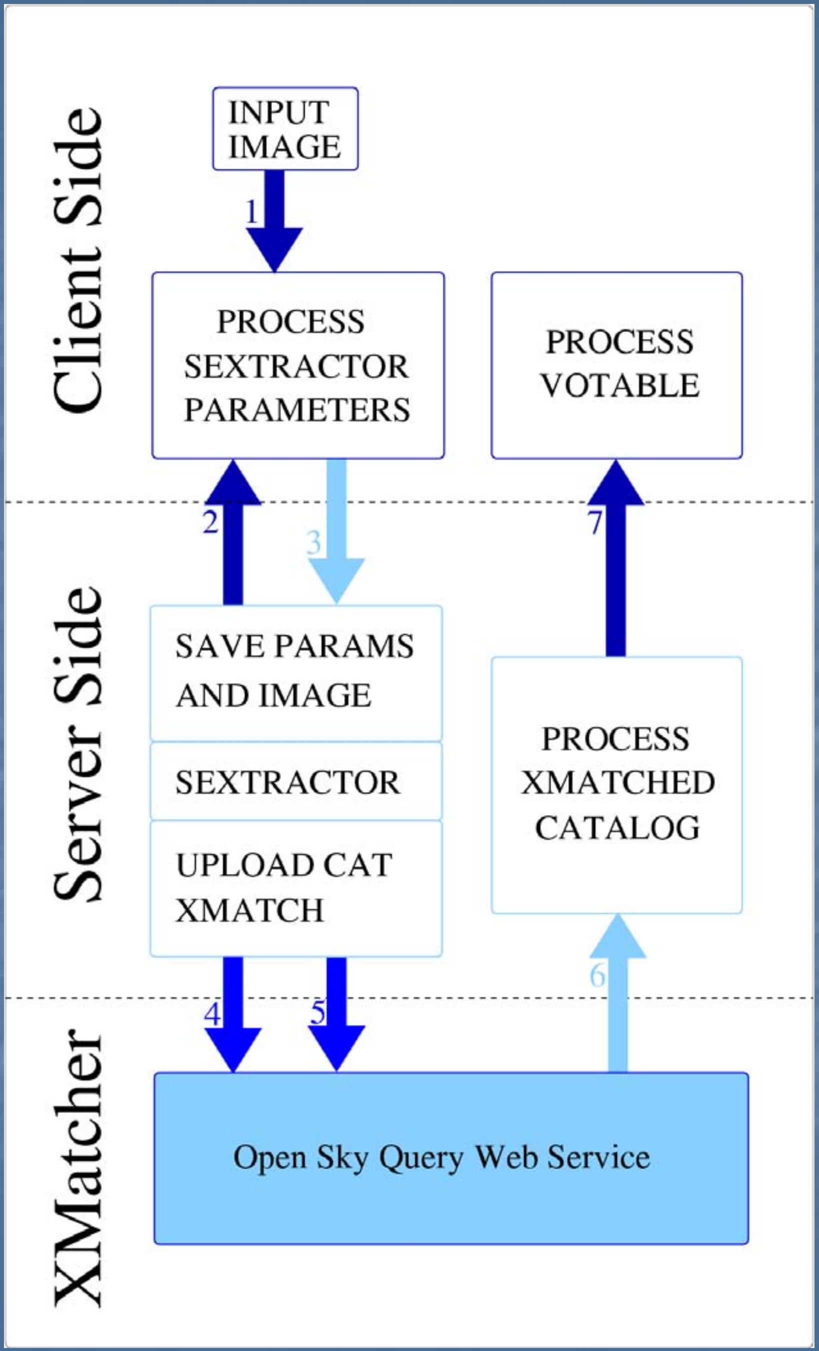
Jeff Gardner, Alex Gray, Simon Krughoff,
Andrew Moore, Bob Nichol, Jeff Schneider

Serial and Parallel Applications

- ITR collaboration: University of Pittsburgh and Carnegie Mellon University.
- Astronomers, Computer Scientists and Statisticians
- Developing fast, usually tree-based, algorithms to reduce large volumes of data.
- Sample projects:
  - Source detection and cross match
  - Parallel Correlation Functions
  - Parallelized serial applications (through GridShell on the Teragrid)
  - Object classification (Morphologies and general classification)
  - Anomaly finding
  - Density estimation
  - Intersections in parameter space

WSExtractor: Source Detection

- Wrapping existing services as webservices
  - Accessible through client and webservices
- Prototype Sextractor (and interface to other services)
  - Communication through SOAP messages and attachments
  - Full configurable through webservice interface
  - Outputs VOTables
  - Dynamically cross matches with openskyquery.net
  - Returns cross matched VOTables
  - Plots through Aladin and VOPlot WSext


Client Side
INPUT IMAGE
1
PROCESS SEXTRACTOR PARAMETERS
PROCESS VOTABLE
2
3
7
Server Side
SAVE PARAMS AND IMAGE
SEXTRACTOR
UPLOAD CAT XMATCH
PROCESS XMATCHED CATALOG
4
5
6
XMatcher
Open Sky Query Web Service

NVO

National Virtual Observatory Source Extractor

Home | Help | Upload Image | Plotter | Aladin

Welcome to the homepage of WSextractor

There are just six steps to getting your source catalog back.
If you are interested in testing out this service,
here is a test file that works.

Step 1: Specify the file you want to upload

/home/simon/images/756/ Browse...

Step 2: Select the catalog you would like to crossmatch with.

- ROSAT
- GALEX
- DLS
- RC3
- SDSS
- SDSSDR2

Step 3: Submit your file for processing

Submit

Contact US-VO help desk with any problems

National Virtual Observatory Source Extractor

Home | Help | Upload Image | Plotter | Aladin

SExtractor Output Fields

Step 4: Select the output fields you would like in your catalog.

- FLUXERR_AUTO
- MAG_AUTO
- MAGERR_AUTO
- FLUX_BEST
- FLUXERR_BEST
- MAG_BEST
- MAGERR_BEST
- KRON_RADIUS

Output Fields from SDSSDR2

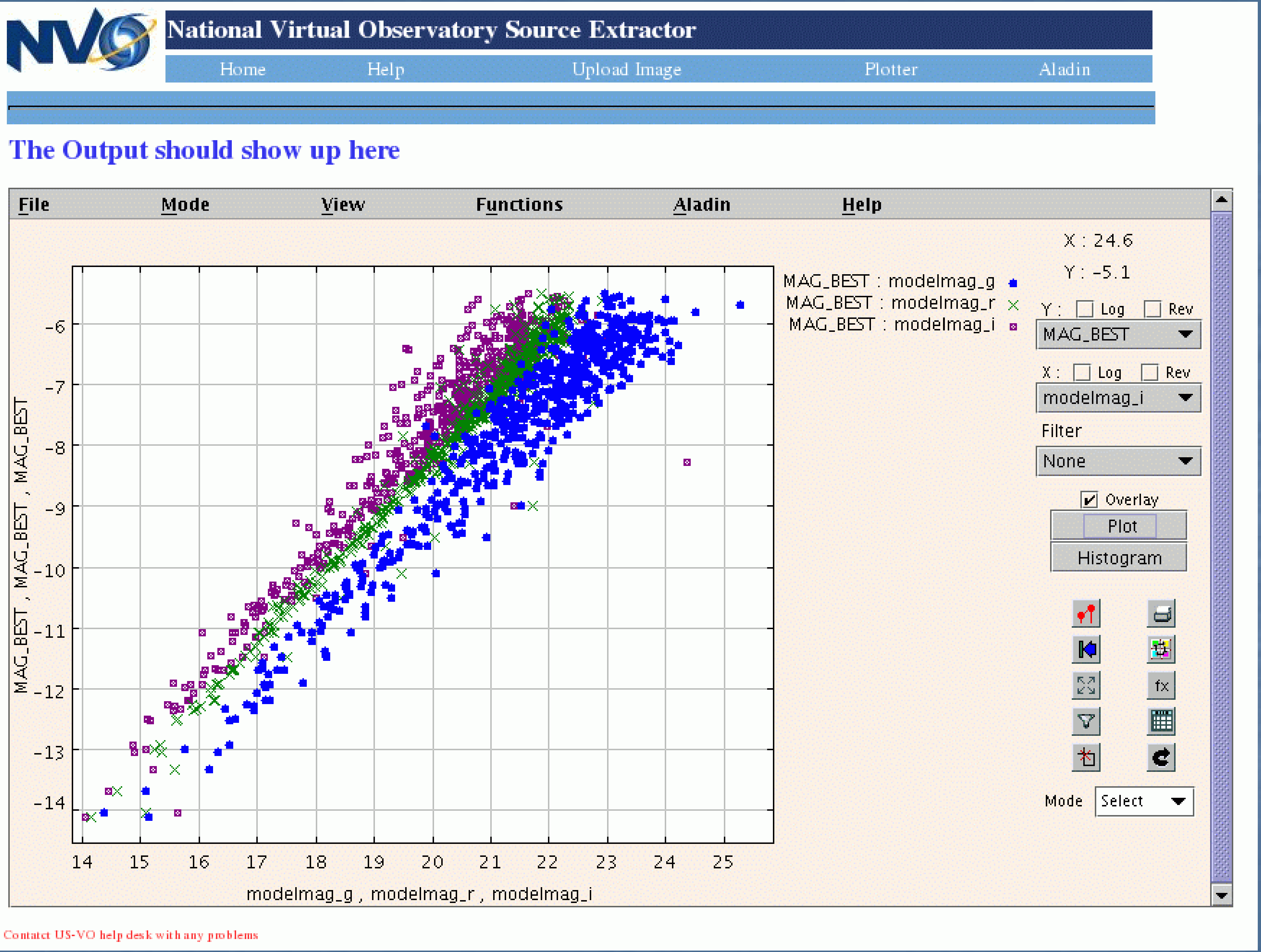
Step 5: Choose the columns you would like included in your CrossMatched catalog

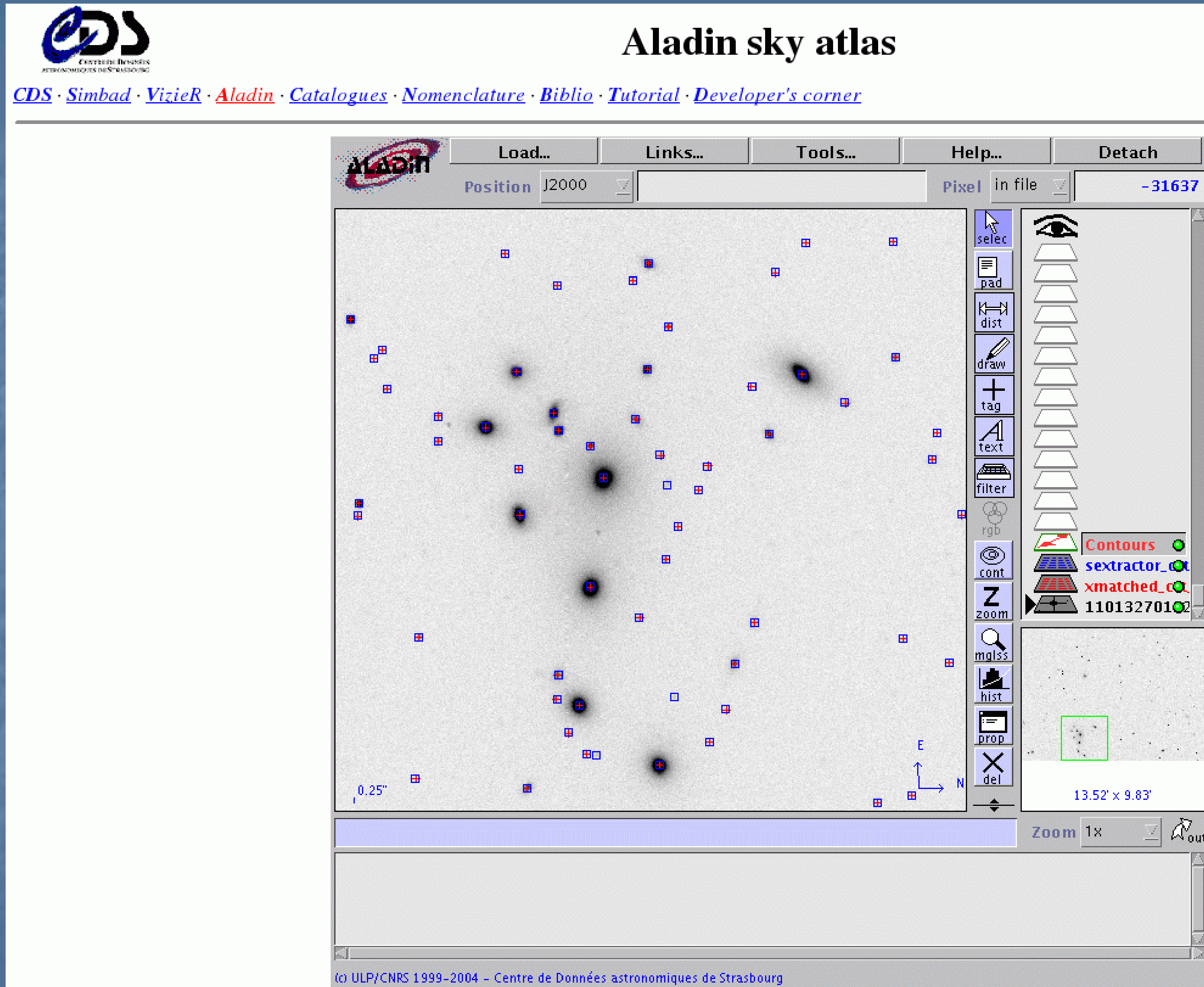
- probPSF_i
- probPSF_z
- status
- ra
- dec
- cx
- cy
- cz
- offsetRa_u
- offsetRa_g

Step 6: Go do it.

Go do it!

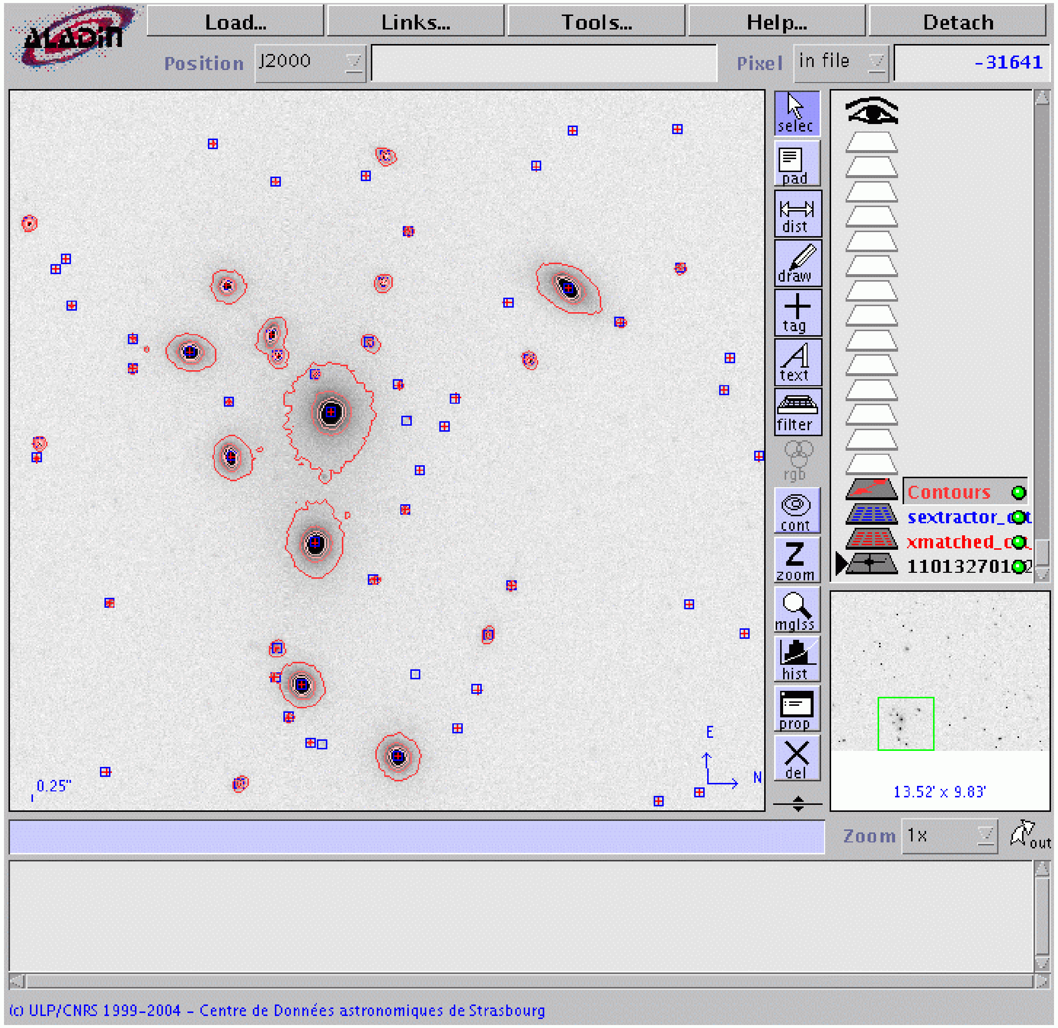
Contact US-VO help desk with any problems


NVO
National Virtual Observatory Source Extractor
Home
Help
Upload Image
Plotter
Aladin
The Output should show up here
File
Mode
View
Functions
Aladin
Help
X : 24.6
Y : -5.1
MAG_BEST : modelmag_g
MAG_BEST : modelmag_r
MAG_BEST : modelmag_i
Y : Log Rev
MAG_BEST
X : Log Rev
modelmag_i
Filter
None
Overlay
Plot
Histogram
fx
Mode
Select
-6
-7
-8
-9
-10
-11
-12
-13
-14
MAG_BEST , MAG_BEST , MAG_BEST
14
15
16
17
18
19
20
21
22
23
24
25
modelmag_g , modelmag_r , modelmag_i
Contact US-VO help desk with any problems


Aladin sky atlas
CDS · Simbad · VizieR · Aladin · Catalogues · Nomenclature · Biblio · Tutorial · Developer's corner
Load...
Links...
Tools...
Help...
Detach
Position
J2000
Pixel
in file
-31637
selec
pad
dist
draw
tag
text
filter
rgb
cont
zoom
mglss
hist
prop
del
Contours
sextractor_cat
xmatched_cat
1101327012
13.52' x 9.83'
0.25"
E
N
Zoom
1x
out
(c) ULP/CNRS 1999-2004 - Centre de Données astronomiques de Strasbourg

Aladin sky atlas

CDS · Simbad · VizieR · Aladin · Catalogues · Nomenclature · Biblio · Tutorial · Developer's corner



Issues

- VOTable and Java
- DataHandler type
  - Working with C and .NET
- Aladin as applet on server
  - Eliminates multiple transfers of images
- Concurrent Access – sessions with WS
- Value validation

Grid Services

- Harnessing parallel grid resources in the NVO data mining framework
- N-pt correlation functions for SDSS data:
    - 2-pt: hours
    - 3-pt: weeks
    - 4-pt: 100 years!
  - With the NVO, computational requirements will be much more extreme.
  - There will be many more problems for which throughput can be substantially enhanced by parallel computers.

The challenge of Parallelism

- Parallel programs are hard to write!
  - Parallel codes (especially massively parallel ones) are used by a limited number of "boutique enterprises" with the resources to develop them.
- Scientific programming is a battle between run time and development time:
  - *Development time must be less than run time!*
    - Large development time means they must be reused again and again and again to make the development worth it (e.g. N-body hydrodynamic code).
- Even the "boutiques" find it extraodinarily difficult to conduct new queries in parallel.
  - For example, all of the U.Washington "N-Body Shop's" data analysis code is *still* serial.

A Scalable Parallel Analysis Framework

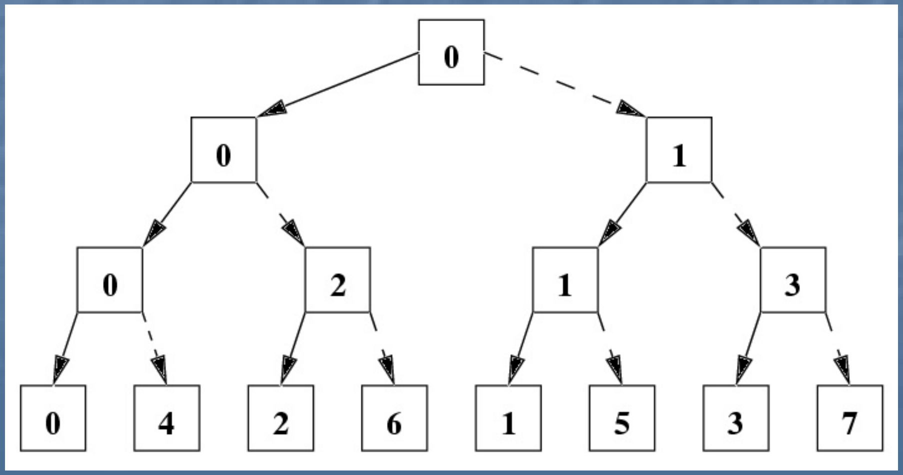
- Canned routines (e.g. "density calculator", "cluster finder", "2-pt correlation function") restrict inquery space.
- Bring the power of a distributed TeraScale computing grid to NVO users. Provide seamless scalability from single processor machines to TeraFlop platforms while not restricting inquery space.
- Toolkit:
  - Highly general
  - Highly flexible
  - Minimizes development time
  - Efficiently and scalably parallel
  - Optimized for common architectures (MPI, SHMEM, POSIX, etc)

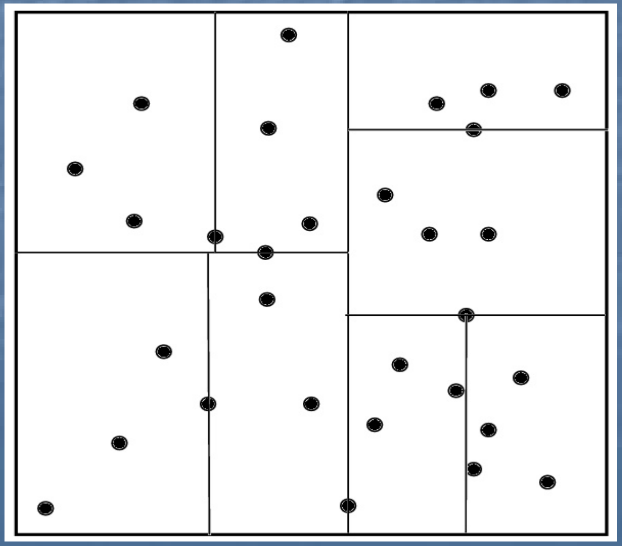
Methodology

- Identify Critical Abstraction Sets and Critical Methods Sets from work done on serial algorithm research by existing ITR.
  - Efficiently parallelize these abstraction and methods sets
  - Distribute in the form of a parallel toolkit.
- Developing a fast serial algorithm is completely different than implementing that algorithm in parallel.

Parallel Npt

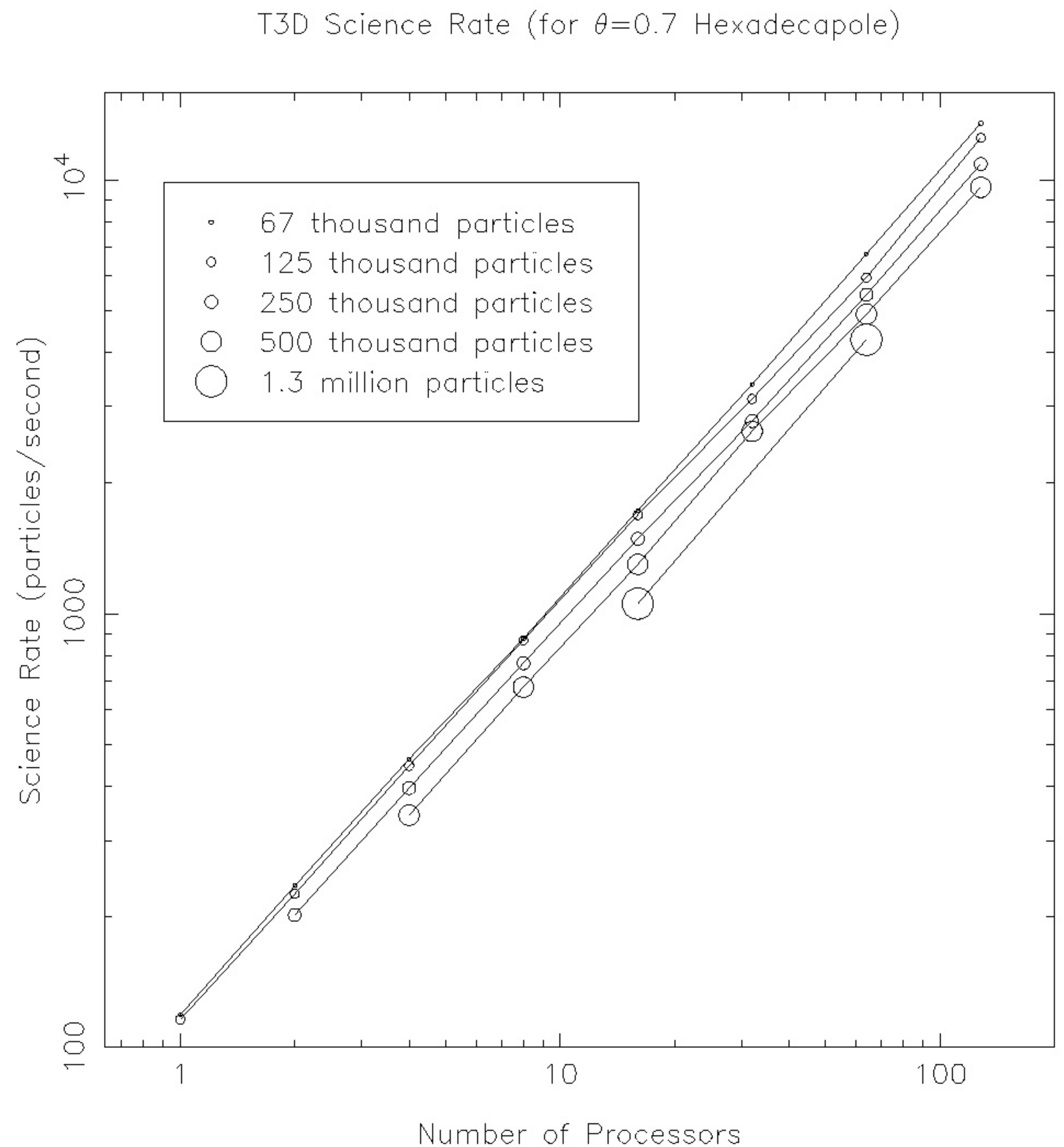
- Developing an efficient & scalable parallel 2-, 3-, 4-point correlation function code.
  - Development on *Terascale Computing System*
- Based on the parallel gravity code PKDGRAV:
  - Highly portable (MPI, POSIX Threads, SHMEM, & others)
  - Highly scalable





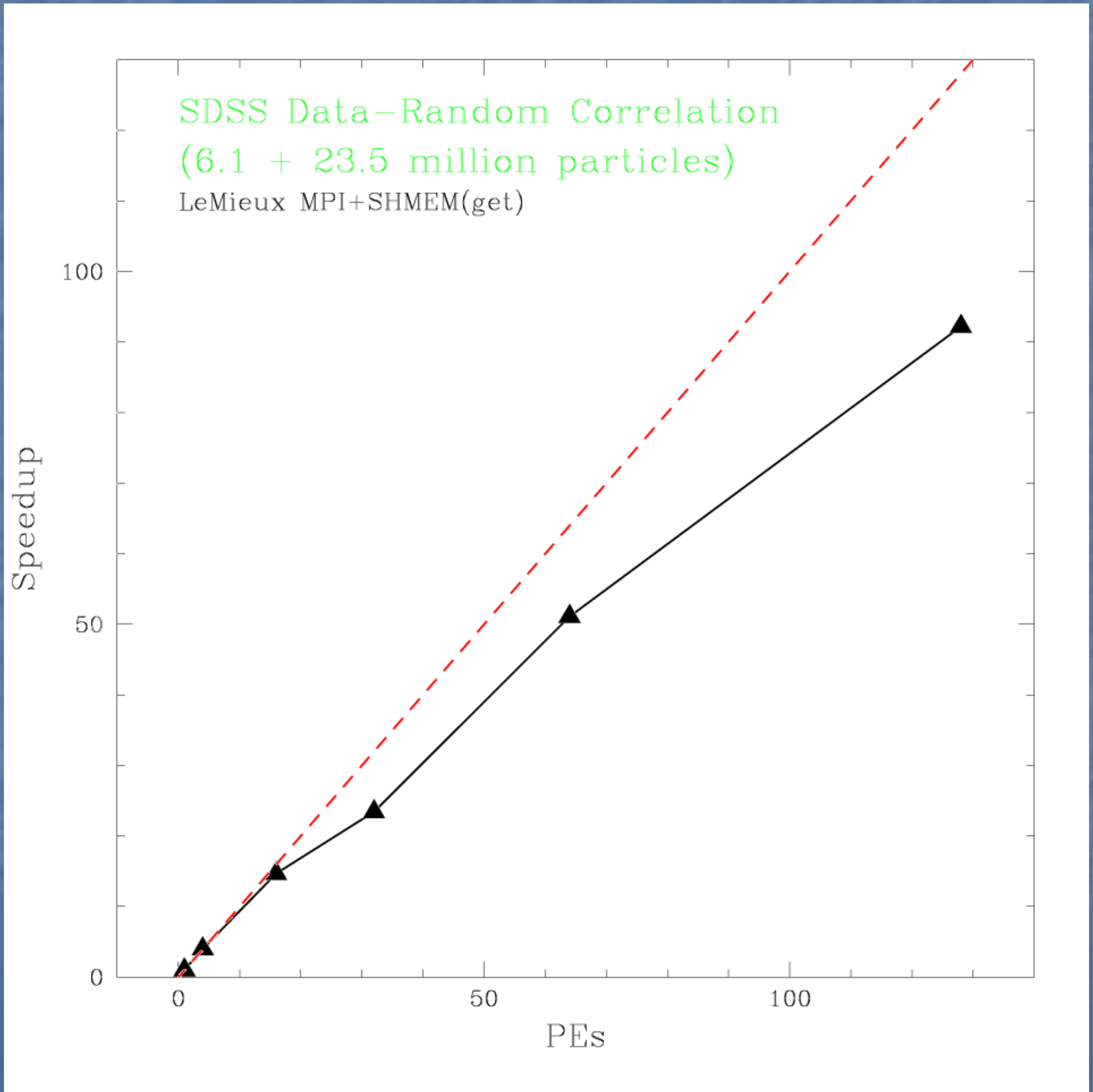
92% linear speedup on 512 PEs!

By Amdahl's Law, this means 0.017% of the code is actually serial.



Parallel Npt performance

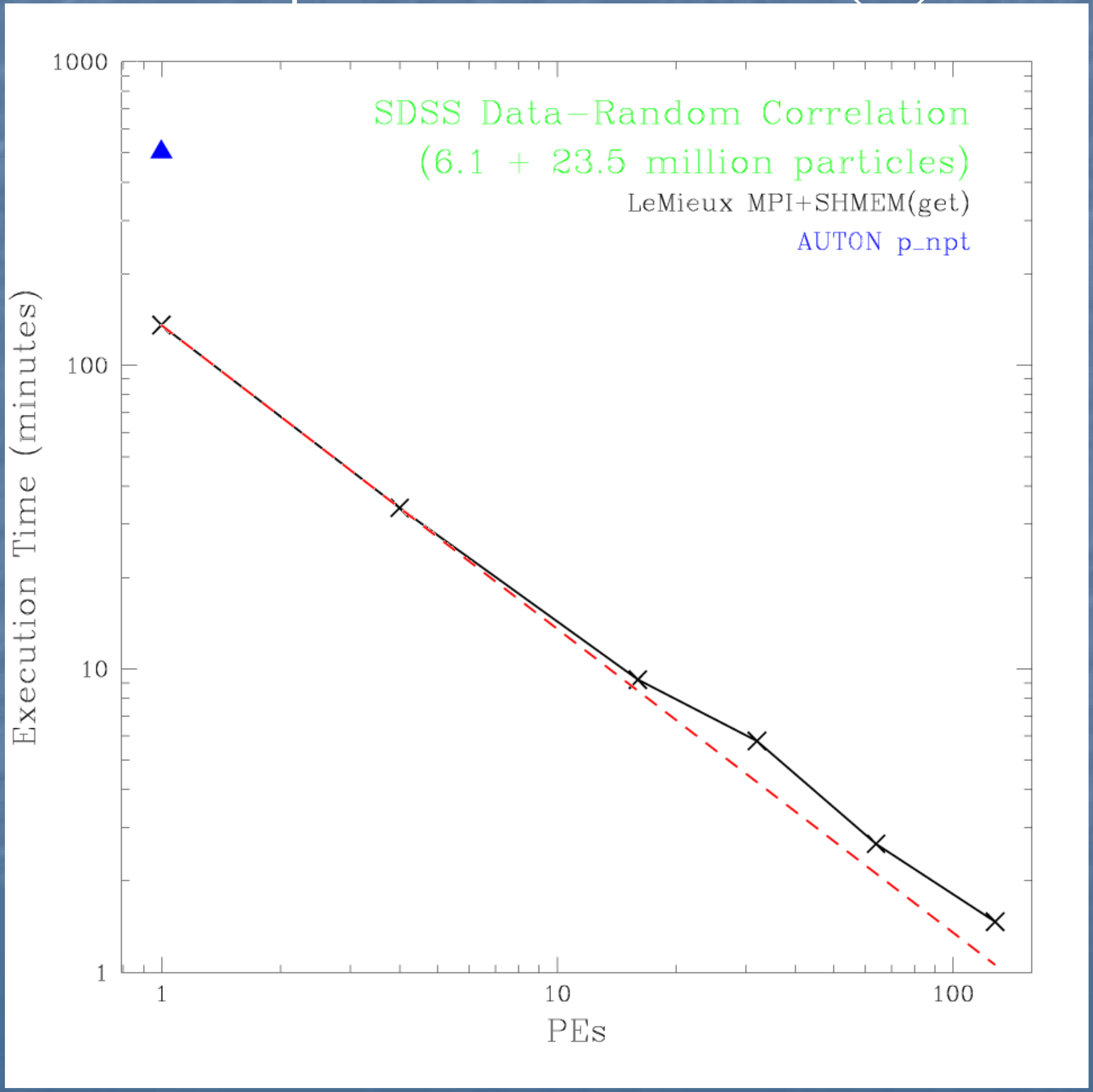
2-pt Correlation Function (2°)



So far, only 74% speedup on 128 PEs

Parallel Npt Performance

2-pt Correlation Function (2º)



Issues and Future directions

- Problem is communication latencies
- N-body inter-node communication small (npt large)
- Minimize off processor communication
- Extend to npt (in progress)
- Adaptive learning for allocating resources and work load
- Interface with webservices

- Not another IRAF