Harnessing large numbers of processors from the Virtual Observatory

Distributed Systems Group

Garry Smith,
Univ. Edinburgh, Univ. Portsmouth.

29th July 2005.

Outline

- Overview,
- Introducing the algorithms: KDE Tree and NPT,
- Heterogeneity and the testbed,
- High-level requirements,
- Standards and existing software,
- Overview of the Broker,
- Summary.

Overview (1)

- The Virtual Observatory provides mechanisms for astronomers to use globally distributed services in a transparent and seamless manner for data mining and visualisation.
- EuroVO VOTech (DS6) project aims:
  - To develop new infrastructure, tools and data mining services for the Virtual Observatory.

Driven by science…

- Prof Bob Nichol, (ICG, Portsmouth) requires access to large numbers of processors to perform parameter sweeps,
- Initially focusing on two algorithms (NPT and KDE Tree)…

Distributed Systems Group

NPT algorithm

- Calculates the distance of galaxies in the Universe,
- 2pt function,
- 3pt function:
  - r: Shortest side of triangle,
  - q*r: Next shortest side,
  - Θ: Angle between previous two sides.
- Locate as many triangles that fit r,q,Θ configuration.
- Repeat many times for a bin,
- Reduce intermediate values to a single Q value for a bin,
- Have n bins, therefore n Q values.

Distributed Systems Group

Bob's NPT execution wish list (1)

- Input parameter generation:
  - –Create initial input parameters (bounds) data,
  - –Drag points on a chart to generate appropriate r,q,Θ values:
    - Currently generate using an IDL script,
  - –Large number of values make this an arduous task,
  - –Would like to reduce overhead for astronomer,
  - –Use new parameters values for new job submission.

Bob's NPT execution wish list (2)

- NPT produces intermediate raw data,
  - –Astronomer will want to check to this looks sane.
- Need to visualise intermediate and final data values in a meaningful way,
- Would like feedback mechanisms incorporated into graphical output:
  - –Create a new point on the chart:
    - Have the appropriate input values created automatically, based on the user selection,
    - Show the results back on the graph:
      - – Appropriate position, colour.

The KDE algorithm (1)

- A fast tree-based Kernel Density Estimation (KDE) code of Gray & Moore (2004).
- This is a command-line c code that takes a list of N-dimensional data-points and returns the probability density at position of each of the points.
- This probability is derived from optimally smoothing the data with an N-dimensional

Compute resources

- What happens when the existing compute resources an astronomer wishes to use are not/cannot be exposed in an IVOA compliant manner?
- We have access to a large number of heterogeneous resources:
  - Differences in platform, job scheduling system, some 'Grid' enabled, others not, different Grid middleware versions.
  - Produces complexity at the client.
- Why not 'just' install a common Grid middleware:
  - Inter-site politics, conflicting middleware versions, admin overhead, firewall rules,… ouch!
- Approach: We don't mandate the software a site should have, we just use whatever is available.

Distributed Systems Group

Testbed (1)

- Compute resources:
  - UK National Grid Service (NGS) www.ngs.ac.uk :
    - Compute clusters:
      - Accessed via Globus version 2.4.3
      - PBS for job execution
      - Storage Resource Broker (SRB) for storage
  - Univ. Portsmouth:
    - DSG Starbug Cluster:
      - Globus for remote site access,
      - Condor for direct submission from local users.
    - ICG:
      - Altrix and SGI machines: PBS, Condor (Internal), SSH
      - Workstations (Condor cycle stealing) coming soon!

Distributed Systems Group

Testbed (2)

- Compute resources (cont..):
  - –Univ. Westminster
    - Condor pools (2 clusters)
  - –Cambridge:
    - Cosmos:
      - – A 152 processor SGI Altix 3700 super computer,
      - – Shared memory, shared file system,
      - – No “Grid” access,
      - – Users log in via ssh to execute jobs to Platform LSF batch system.
- Others …

Distributed Systems Group

Requirements (1)

- From an astronomers point of view, Job submission should not be complicated, e.g:
  - Select algorithm, dataset, input parameters,
  - Select the required number of processors,
  - Click ‘submit’ and forget!
  - Resource location, scheduling decisions, process restart, data collecting, etc should be handled by a remote entity (a broker),
  - Disconnect client from the network, come back later to observe progress or collect results.
- A key aim is to abstract the astronomer from underlying complexity.

Distributed Systems Group

Requirements (2)

- Any solution should be flexible and allow new algorithms to be added with little effort,
- Standards compliance importance,
- Must integrate with the Virtual Observatory framework (in this case AstroGrid) in a consistent manner,
- No reinventing the wheel, so work with existing software.

Distributed Systems Group

Pertinent Standards (1)

IVOA Standards (Recommendations)

VOTable Format Definition Version 1.1:

- An XML language,
  - Flexible storage and exchange format for tabular data: Emphasis on astronomical tables,
- Allows meta data and data to be stored separately with links to remote data.

- Resource Metadata for the VO Version 1.01:
  - For describing what data and computational facilities are available, and once identified how to use them.
- Unified Content Descriptor (UCD) (Proposed):
  - A formal (and restricted) vocabulary for astronomical data.
- IVOA Identifiers Version 1.10 (Proposed):
  - Syntax for globally unique resource names.

Pertinent Standards (2)

- Job Submission Description Language (JSDL):
  - For a standard description of job execution requirements to a range of resource managers.
  - GGF Working group:
    - http://forge.gridforum.org/projects/jsdl-wg/
  - Currently in GGF public comment phase

AstroGrid components (1)

- MySpace: Distributed file store for workflows,results,
- Common Execution Architecture (CEA):
  - Codes need wrapping before use,
  - Take command line apps and present as a Web Service.
- Algorithm Registry:
  - Meta data from wrapped codes are published in a yellow pages, for searching.
- Portal:
  - Web interface for interacting with preceding services,
  - Workflow: Coordinate data flow/control of components within a larger system of work,
  - Submit jobs and observe status, and access files in MySpace.
- Dashboard/Workbench:
  - Interact with MySpace, Registry, CEA from any language that provides XML-RPC library. Web Start application.

AstroGrid components (2)

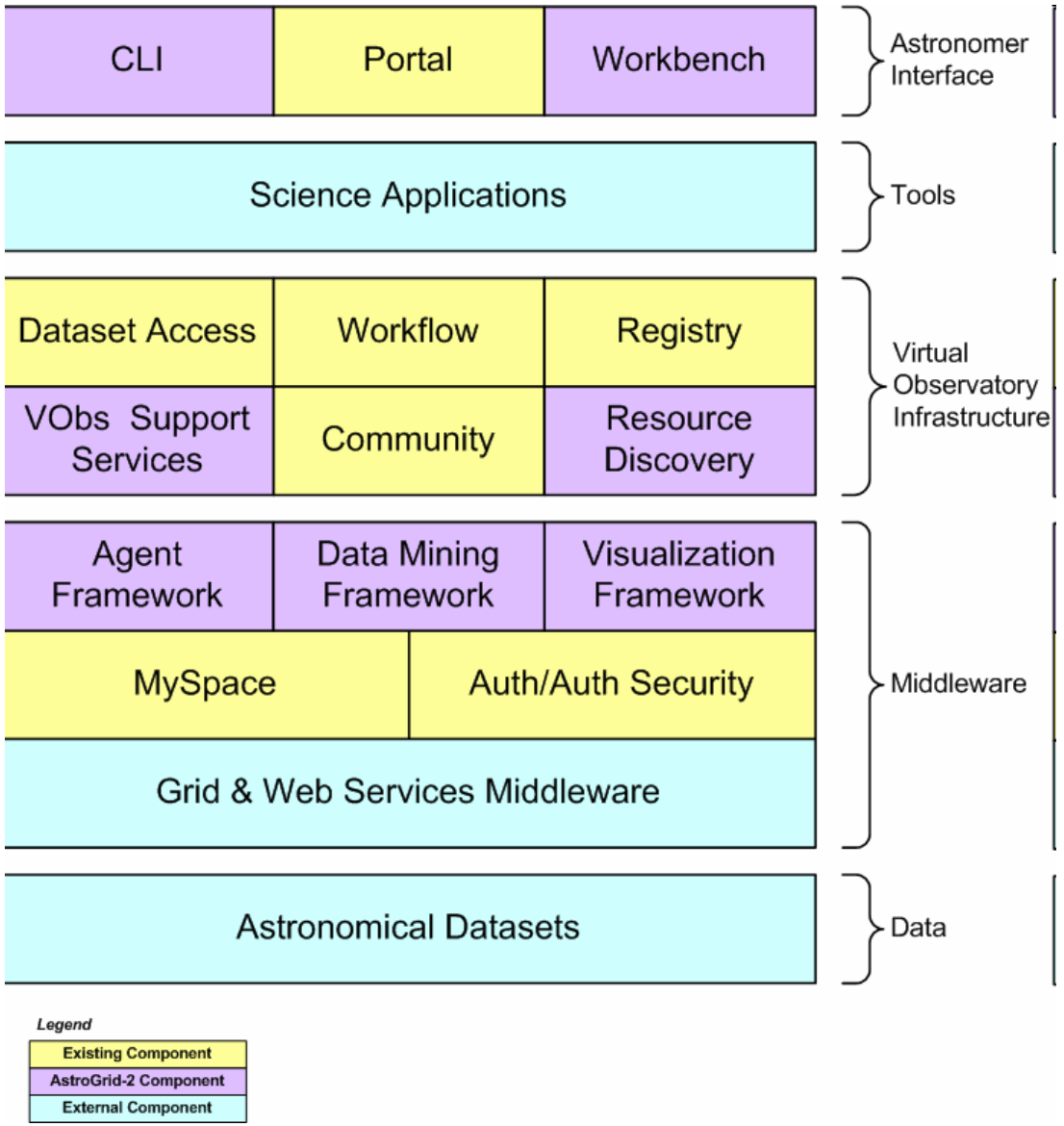


Figure courtesy of [linde2004]

Distributed Systems Group

Other software

- Globus – Grid middleware,
- Myproxy – X509 credential service,
- Condor –batch job submission and cycle stealing,
- Portable Batch System (PBS),
- Storage Resource Broker (SRB)– distributed file system,
- GridRM – provides a homogeneous way to obtain data from heterogeneous monitoring agents.

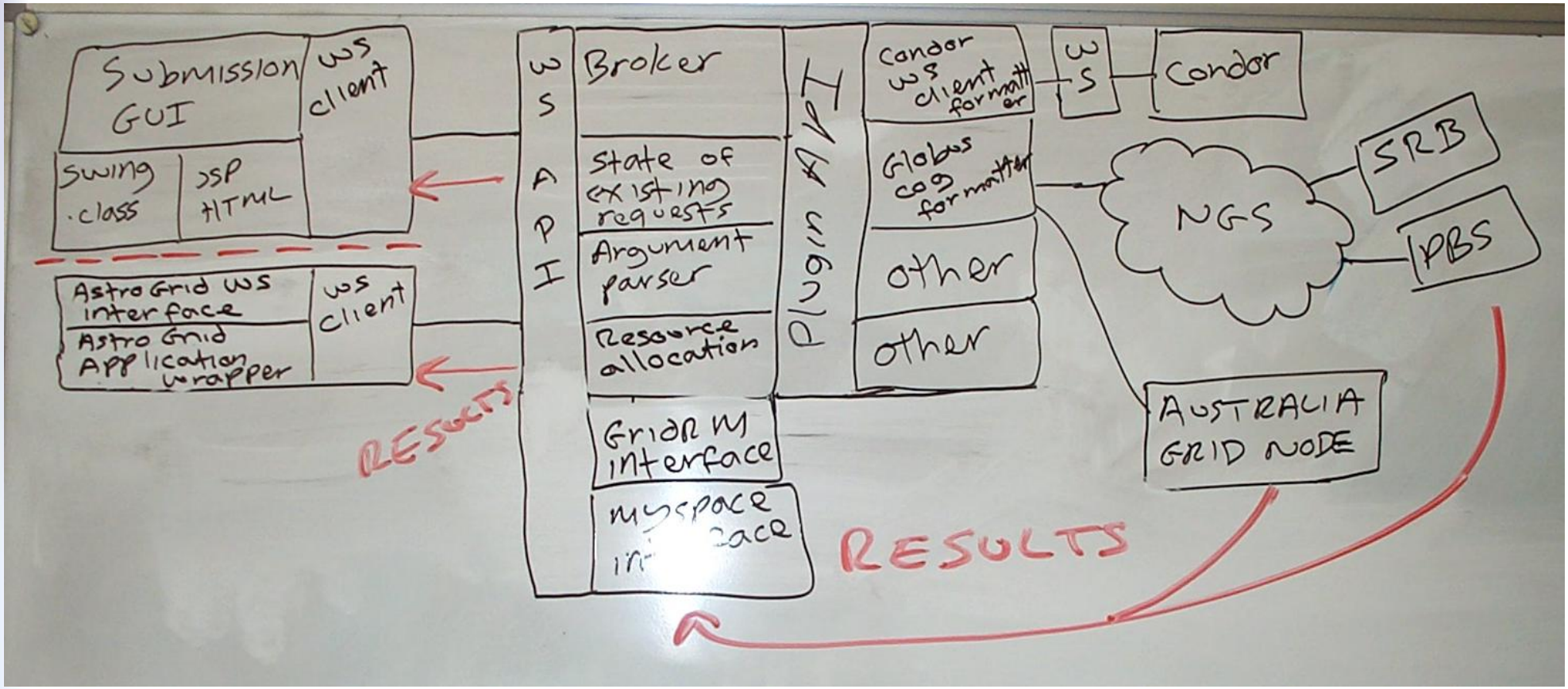
Distributed Systems Group

Broker

- Execute potentially thousands of sequential processes simultaneously, repeat multiple times,
- Utilise existing infrastructure at remote sites:
  - e.g. computational resources: Condor, Globus,
  - Transparent to the user.
- Locate suitable compute nodes (i.e. processor type, available libraries, CPU load, memory,
- Stage code and observe status of running processes,
- Combine results for further analysis, e.g as input to a post-mortem visualisation component in the AG workflow.
- Implementation is partly based on GridSAM from

Distributed Systems Group

Broker



JSDL Submission script

- Example client submission request in JSDL:

```
garry@chewy:~/jsdl> cat helloworld-myproxy.jsdl
<?xml version="1.0" encoding="UTF-8"?>
  <JobDefinition xmlns="http://schemas.ggf.org/jsdl/2005/06/jsdl">
   <JobDescription>
         <Application>
               <POSIXApplication
                   <Executable>/bin/echo</Executable>
                   <Argument>hello world</Argument>
               </POSIXApplication>
           </Application>
       </JobDescription>
       <MyProxy xmlns="urn:gridsam:myproxy">
           <ProxyServer>myproxy.grid-support.ac.uk</ProxyServer>
           <ProxyServerDN>/C=UK/O=eScience/OU=CLRC/L=DL/CN=host/myproxy.grid-
           <ProxyServerPort>7512</ProxyServerPort>
           <ProxyServerUserName>garrySmith</ProxyServerUserName>
           <ProxyServerPassPhrase>1234567</ProxyServerPassPhrase>
           <ProxyServerLifetime>2</ProxyServerLifetime>
       </MyProxy>
    </JobDefinition>
```

JSDL Submission script

- Client submission:

```
garry@chewy:~/jsdl> ~/bin/gridsam-client/bin/gridsam-submit -s
"http://l09.dsg.port.ac.uk:8080/gridsam/services/gridsam?WSDL" -

urn:gridsam:4028e4ed0539a289010539ab5e09000d
```

Getting status

```
garry@chewy:~/jsdl> ~/bin/gridsam-client/bin/gridsam-status -s
"http://l09.dsg.port.ac.uk:8080/gridsam/services/gridsam?WSDL" -j
urn:gridsam:4028e4ed0539a289010539ab5e09000d
Job Progress: pending -> staging-in -> staged-in -> active -> executed
--- pending - 2005-07-21 14:53:16.0 ---
job is being scheduled
--- staging-in - 2005-07-21 14:53:16.0 ---
staging files...
--- staged-in - 2005-07-21 14:53:16.0 ---
no file needs to be staged in
--- active - 2005-07-21 14:53:16.0 ---
job is being submitted through globus
--- executed - 2005-07-21 14:53:17.0 ---
globus job completed

--------------
Job Properties
--------------
urn:gridsam:globus:id=https://grid-compute.leeds.ac.uk:64015/23182/1121953988/
urn:gridsam:globus:rsl=&
(executable=/bin/echo)
   (arguments=hello world )
```

Distributed Systems Group

Summary

- A broker to submit parameter sweeps to the Grid, and other distributed resources, in a transparent way,
- Aim to interoperate with a wide range of job submission systems using a plug-in system,
- Do not mandate that version X of middleware Y must be installed,
- Protect the astronomer from boring ‘details’,
  - Select algorithm, inputs, number of processors, submit!

Standards based approach,

- Interoperability important: IVOA, Astro Grid, etc.

Work ongoing,…

References

- EuroVO, http://www.euro-vo.org,
- IVOA, http://www.ivoa.net,
- VOTech,http://wiki.eurovotech.org/bin/view/VOTech/WebHome,
- GridSAM http://www.lesc.ic.ac.uk/gridsam,
- GridRM, http://gridrm.org,
- Condor, http://www.cs.wisc.edu/condor/manual/v6.6,