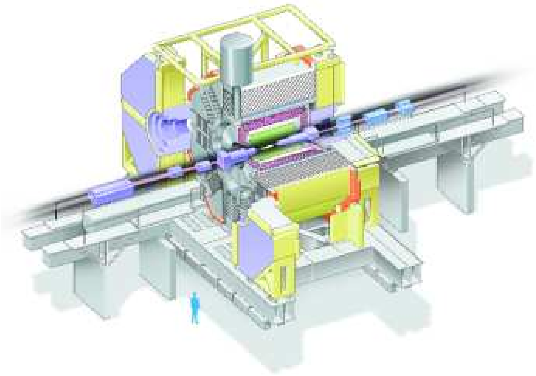
A Data Grid Environment and Testbed for the Analysis of Data from the Belle Experiment

Glenn Moloney
University of Melbourne



1–5 December 2003

The Australian HEP Data Grid Team

Who are we?

The Australian HEP Data Grid Team

Who are we?

- *Physicists:*
  - Experimental Particle Physics: University of Melbourne
  - Falkiner High Energy Physics: University of Sydney

The Australian HEP Data Grid Team

Who are we?

- *Physicists:*
  - Experimental Particle Physics: University of Melbourne
  - Falkiner High Energy Physics: University of Sydney
- *Computer Scientists:*
  - GRIDS Lab: University of Melbourne
  - Computer Science: University of Adelaide

The Australian HEP Data Grid Team

Who are we?

- *Physicists:*
  - Experimental Particle Physics: University of Melbourne
  - Falkiner High Energy Physics: University of Sydney
- *Computer Scientists:*
  - GRIDS Lab: University of Melbourne
  - Computer Science: University of Adelaide
- *High Performance Computing:*
  - MARCCentre (HPC): University of Melbourne
  - Internet Futures Group: Australian National University
  - Australian Partnership for Advanced Computing (APAC)
  - Victorian Partnership for Advanced Computing (VPAC)
  - GrangeNet: Australian 10Gb Academic Research Network
  - IBM Singapore

What are our activities?

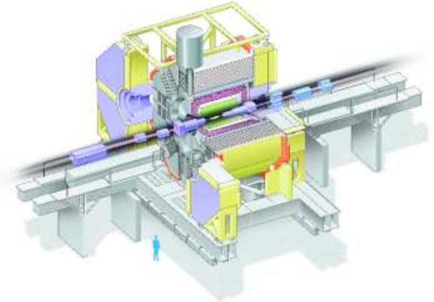
Atlas

- Participate in Atlas Data challenges
  - with *HPC* centre at Melbourne

What are our activities?

Atlas

- Participate in Atlas Data challenges
  - with *HPC* centre at Melbourne



Belle

- Introducing Grid techniques to:
  - Belle physics analysis
  - Monte Carlo generation

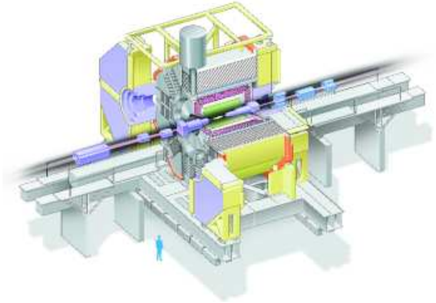
What are our activities?

Atlas

- Participate in Atlas Data challenges
  - with *HPC* centre at Melbourne

Belle

- Introducing Grid techniques to:
  - Belle physics analysis
  - Monte Carlo generation



We have funding for:

- Post-doc for 2 years: *Lyle Winton* (OzBelle Grid)
- System Programmer: *Robert Sturrock* (Atlas Data Challenges)
- Funded by Australian Research Council and *Expertise Program* of the Victorian Partnership for Advanced Computing

Australian Belle Data Grid Testbed

- *“Simple”* Data Grid tools could provide real benefits for physicists *now*:
  - Data Catalogue (Replica Catalogue)
  - *Network-aware* scheduler

Initially, we aimed to:

- Use standard middleware products wherever possible
- Develop simple tools to fill the gaps
- Start *real* data analysis ASAP.

Then move on to:

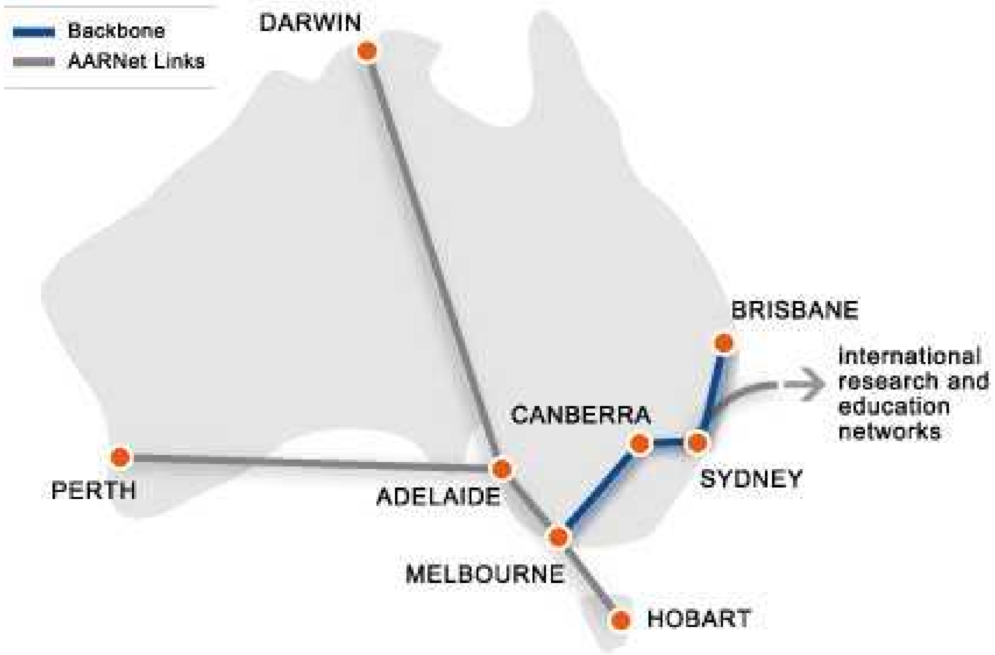
- Trial and incorporate more sophisticated tools for:
  - Scheduling
  - Data Replication and Caching
    - *EDG, LCG, SRB, . . .*
  - Monitoring and Simulation

    (In collaboration with CS colleagues)

What have we got to work with?

Network Infrastructure in Australia:

- Australian Aacademic Research Network (AARNET)
- GrangeNet: Multi-gigabit network to support grid and advanced research projects

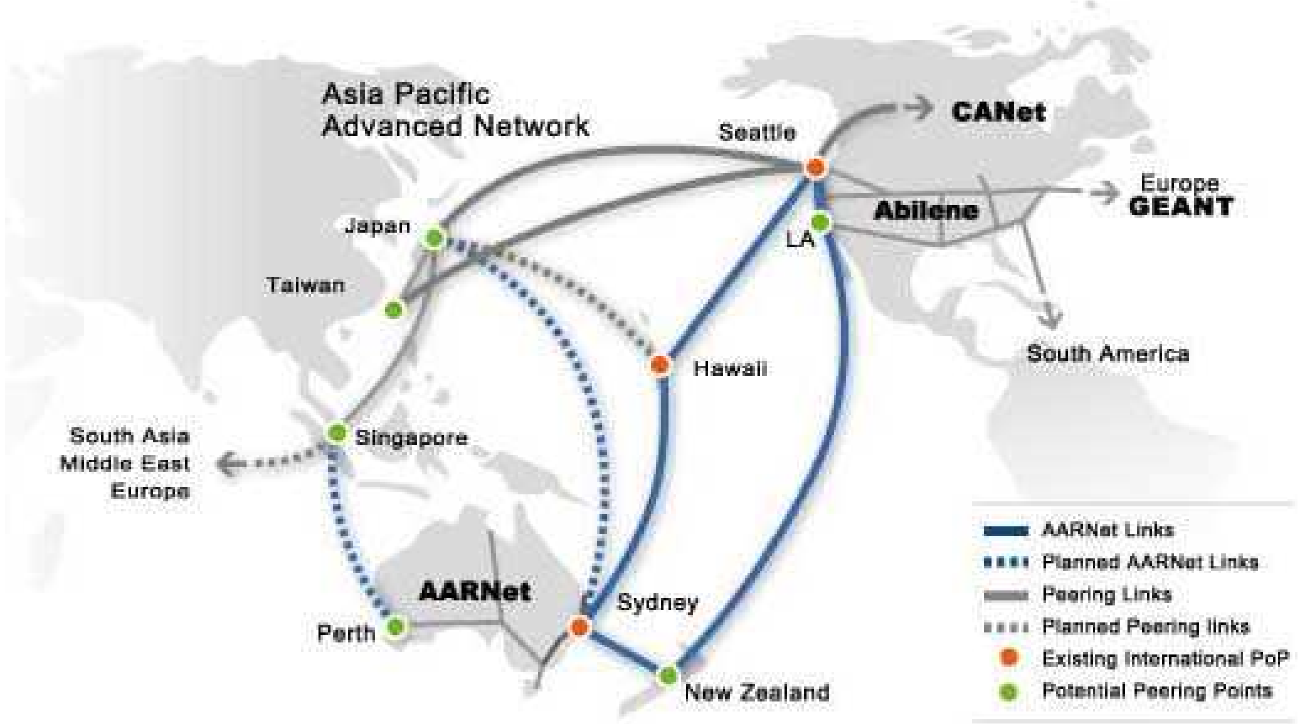


- Active 2003
- 10 Gigabit backbone between:
  - Melbourne
  - Sydney
  - Canberra
  - Brisbane

Future Upgrades to International Links

Planned upgrades to international research and education links

- 10Gb to US
  - *within 12 months*
- 10Gb to Japan
  - *Later*
- 100Mb to Singapore
  - Being installed now



What have we done?

- Installed Globus at each Facility:
  - Melbourne, Sydney, Canberra, Adelaide
  - Mix of Globus 2.0, 2.2 and 2.4
  - Certificate Authority in Melbourne
  - Battled with bugs and undocumented features

Lyle Winton

The Belle Analysis Software: *BASF*

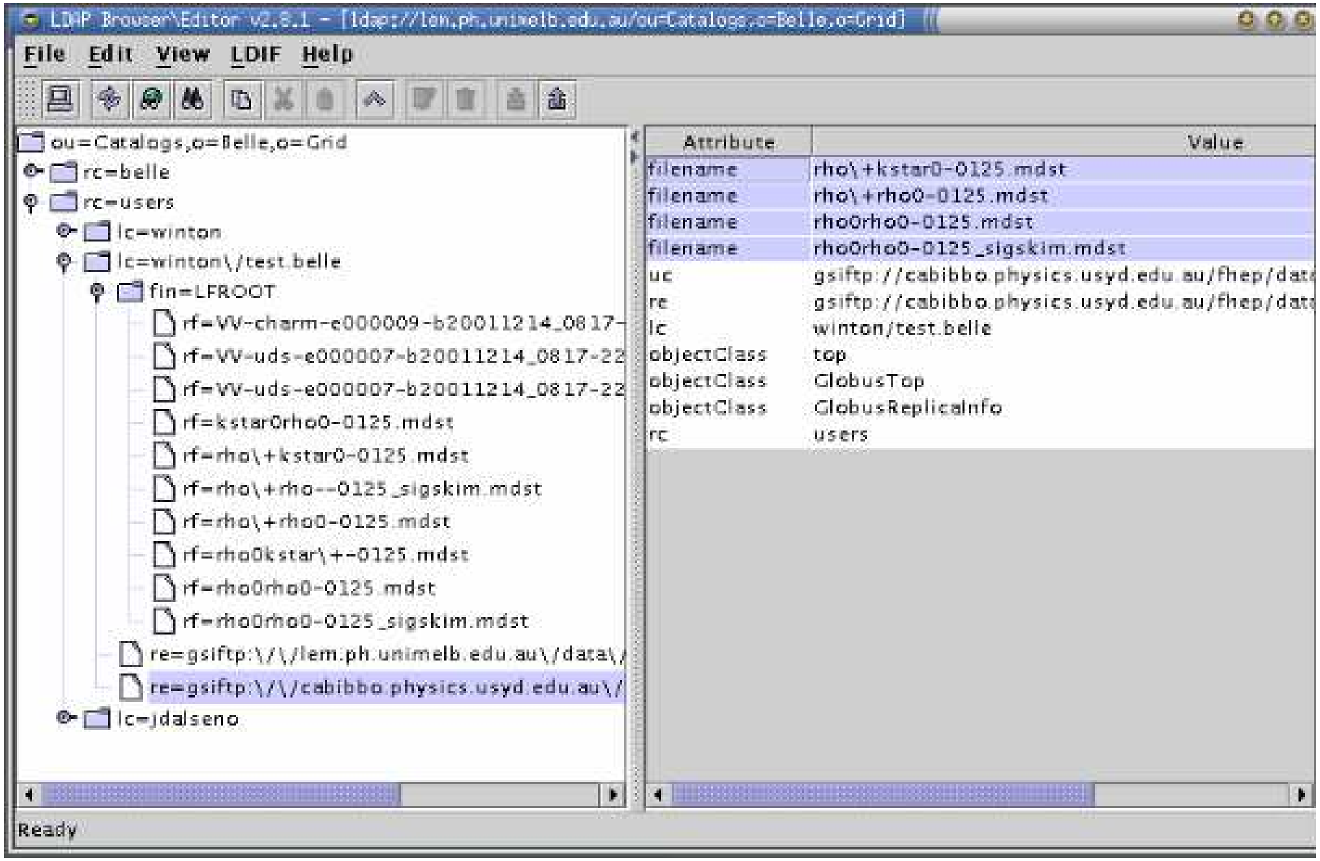
- Enable BASF to read and write Grid URIs directly
  - A new IO module for BASF: *fpdagrid.so*
  - Able to *stream* data across network
    - Removes *dead–time* from data transfer
  - A *simple* solution which initially
    - does *not* require data migration support from middleware

Lyle Winton

Replica Catalogue

- Replica Catalogue (virtual data directory)
  - LDAP based

Lyle Winton



LDAP Browser\Editor v2.8.1 - [ldap://lem.ph.unimelb.edu.au/ou=Catalogs,o=Belle,o=Grid]

```
ou=Catalogs,o=Belle,o=Grid
  rc=belle
  rc=users
    lc=winton
    lc=winton\/test.belle
      fin=LFROOT
        rf=VV-charm-e000009-b20011214_0817-
        rf=VV-uds-e000007-b20011214_0817-22
        rf=VV-uds-e000007-b20011214_0817-22
        rf=kstar0rho0-0125.mdst
        rf=rho\+kstar0-0125.mdst
        rf=rho\+rho--0125_sigskim.mdst
        rf=rho\+rho0-0125.mdst
        rf=rho0kstar\+-0125.mdst
        rf=rho0rho0-0125.mdst
        rf=rho0rho0-0125_sigskim.mdst
      re=gsiftp:\/\/lem.ph.unimelb.edu.au\/data\/
      re=gsiftp:\/\/cabibbo.physics.usyd.edu.au\/
    lc=jdalseno
```

Attribute	Value
filename	rho\+kstar0-0125.mdst
filename	rho\+rho0-0125.mdst
filename	rho0rho0-0125.mdst
filename	rho0rho0-0125_sigskim.mdst
uc	gsiftp://cabibbo.physics.usyd.edu.au/fhep/data
re	gsiftp://cabibbo.physics.usyd.edu.au/fhep/data
lc	winton/test.belle
objectClass	top
objectClass	GlobusTop
objectClass	GlobusReplicaInfo
rc	users

- Meta-Data - easily added to LDAP directory

Grid-RC-tools

Lyle Winton

- Convenience for putting data into Replica Catalog
- Developed to emulate Unix directory structure commands

```
>  grid-rc-cd  winton/mcset1
>  grid-rc-mkdir  newcollection
RC Password:  ********
>  grid-rc-ls -l
drwxr-x  Lyle_Winton     2002-11-18_03:36               0  .
-rw-r--  Lyle_Winton     2002-11-18_03:35       503589128  myfile3.mdst
-rw-r--  Lyle_Winton     2002-11-18_03:35       516000000  myfile4.mdst
-rw-r--  Lyle_Winton     2002-11-18_03:35       167506804  myfile5.mdst
>  grid-rc-cp  -local myfile1.mdst  .  gsiftp://remote/dir/
>  grid-rc-cp  gsiftp://remote2/dir/  myfile2.mdst
>  grid-rc-cp  myfile2.mdst  gsiftp://remote3/adir/
>  grid-rc-rm  myfile3.mdst
>  grid-rc-location  *.mdst
/users/winton/mcset1/myfile1.mdst:  gsiftp://remote/dir/myfile1.mdst
/users/winton/mcset1/myfile2.mdst:  gsiftp://remote2/dir/myfile2.mdst
  gsiftp://remote3/adir/myfile2.mdst
/users/winton/mcset1/myfile4.mdst:
/users/winton/mcset1/myfile5.mdst:  http://somehost/otherdir/myfile2.mdst
>  grid-rc-setattr  description=MC D*D*Ks  myfile?.mdst
>  grid-rc-find -r /users/winton(size>=1000)
```

GQSched: Grid Quick & Dirty Scheduler

- Accesses files and collections from the Replica Catalogue
- Simple node and data brokering
    - Process on “proximity” to data
- File transfer is handled by scheduler

Lyle Winton

Replaced now by scheduler from Gridbus Project

GQSched: An example job script

- A parametric job description file:

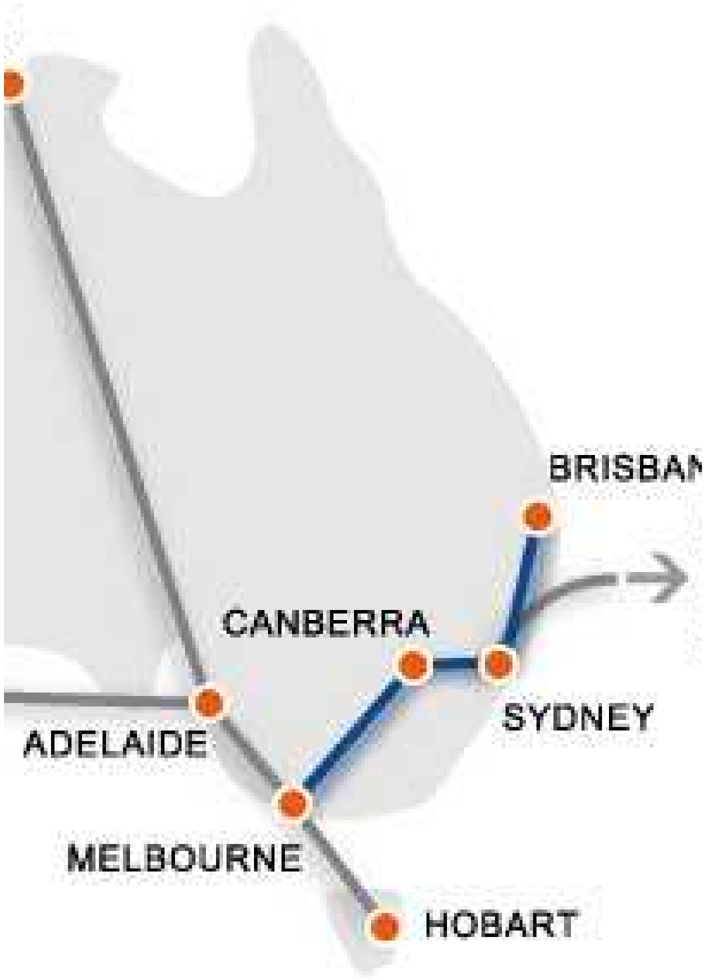
```
#!/bin/csh -f
#:Param FILE GridFile lfn:/users/winton/test.belle/*.mdst
#:Param EVTSKIP Numeric 0 to 9000 step 1000

#:StageIn recon.conf ; event.conf
#:StageIn particleTest.conf particle.conf
#:StageIn libanalyser.so ; user_ana.so ...

echo Processing Job $JOBID on $FILE eventskip $EVTSKIP host ‘hostname‘
setenv FPDA_IO_PACKAGE fpdagrid.so
basfexec -v b20020424_1007 << EOF
path create main
module register user_ana
path add_module main user_ana
initialize
histogram define somehisto.hbook
process_event  $FILE  1000  $EVTSKIP
terminate
EOF
echo Finished JobID $JOBID .

#:StageOut  somehisto.hbook  myoutput.${JOBID}.hbook
```

Testbed Facilities: small distributed nodes

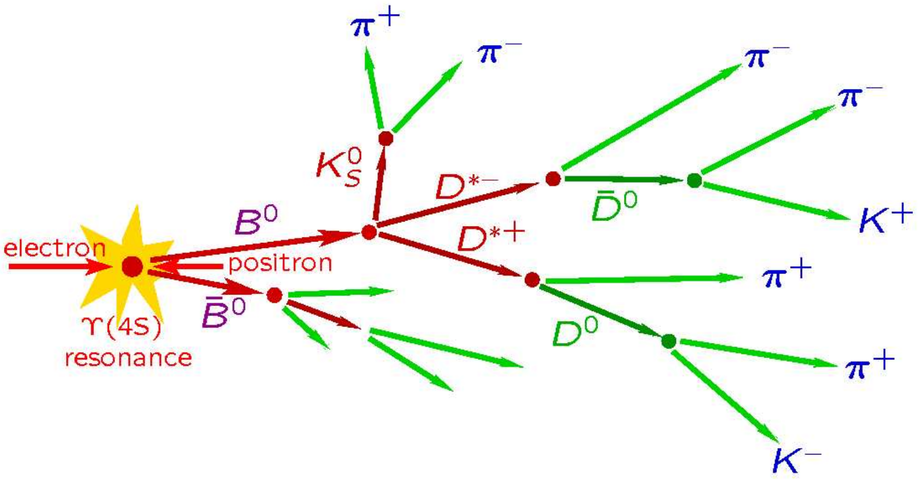


- Uni.Adelaide CS group
  - 4 CPU 2.6GHz (IBM); 70 GB disk
- APAC/GrangeNet (at Canberra)
  - 4 CPU 2.6GHz (IBM); 70 GB disk
- Uni.Melbourne EPP group
  - 1 CPU Intel 1.7GHz ; 70 GB disk
- Uni.Melbourne Computer Science
  - 4 CPU 2.6GHz (IBM); 70 GB disk
- Uni.Sydney HEP group
  - 4 CPU 2.6GHz (IBM); 70 GB disk

Centralised Replica Catalog for management of data

Demonstration at PRAGMA

- Live demonstration at PRAGMA4
  Pacific Rim Applications and Grid Middleware Assembly, June 2003
- Testbed construction began 9 days before!
- Generation of Belle data
- Centralised Replica Catalog
- Discovery of data via global Replica Catalog
- Analysis of all available data



After PRAGMA4

- We have a collaboration with:
  - *Rajkummar Buyya's Gridbus group, CS, University of Melbourne*
  - *Adelaide University CS*
  - *IBM Singapore*
- To deploy and extend the *GridBus* scheduler
  - *Economy based scheduler*
    - Deadline or budget scheduling
  - Designed for computation grids
  - Works with *globus*, *condor*, *legion*, . . .
- *Being extended for data grids*
  - Talk to Replica Catalogue
  - True network and storage "costs"

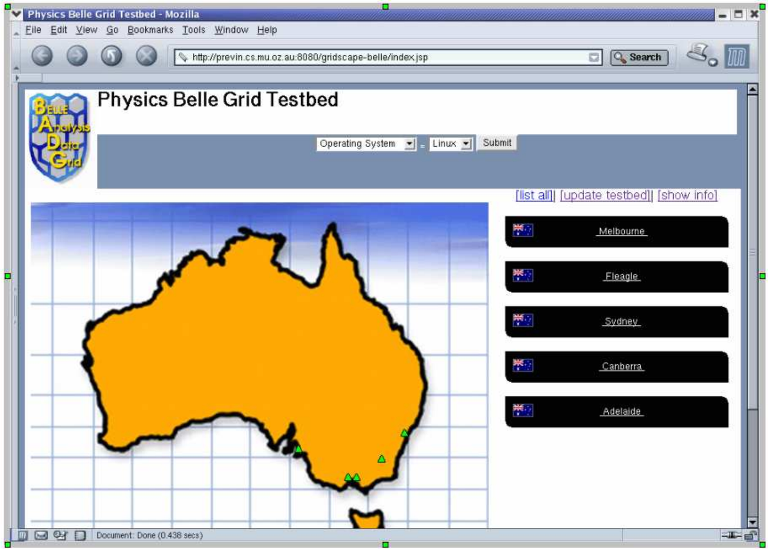
  We also use their:

  - *Gmonitor*: Grid Job Monitoring tool
  - *GridSim*: Grid Simulation Toolkit

After PRAGMA4

We have also migrated to web interfaces:

- Job/Grid Monitoring Services
  - Control and monitor execution of jobs



Physics Belle Grid Testbed - Mozilla
File Edit View Go Bookmarks Tools Window Help
Search
Belle Analysis Data Grid
Operating System
Linux
Submit
[list all] [update testbed] [show info]
Melbourne
Fleagle
Sydney
Canberra
Adelaide

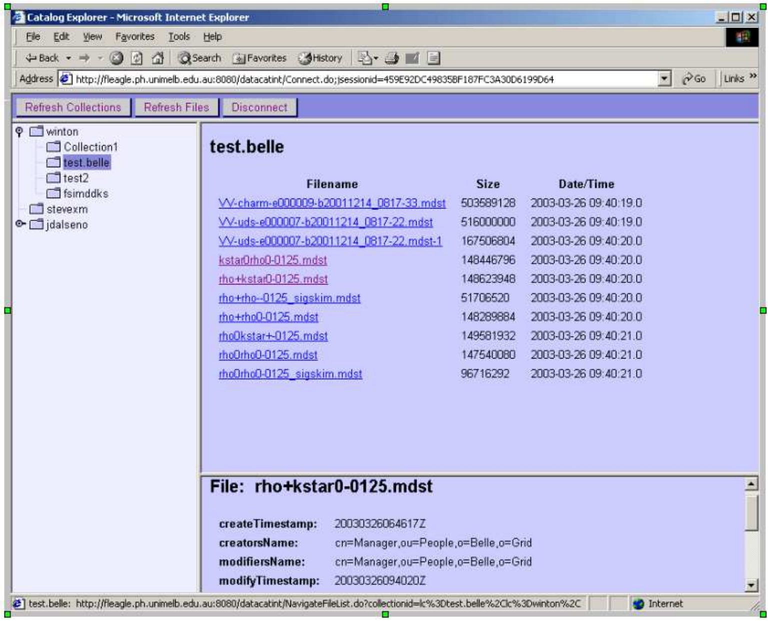
http://previn.cs.mu.oz.au:8080/gridscape-belle/index.jsp

Physics Belle Grid Testbed

Operating System = Linux

Document: Done (0.438 secs)

- Web application/Portal Interface
  - Single point of entry
  - Familiar browser interface
  - Open-source tools, easily portable
  - Shields high-level interactions, and user from lower-level Middleware (Globus)



Address: http://fleagle.ph.unimelb.edu.au:8080/datacatint/Connect.do;jsessionid=459E92DC498358F187FC3A30D6199D64

- winton
  - Collection1
  - test.belle
  - test2
  - fsimddks
- stevexm
- jdalseno

test.belle

Filename	Size	Date/Time
VV-charm-e000009-b20011214_0817-33.mdst	503589128	2003-03-26 09:40:19.0
VV-uds-e000007-b20011214_0817-22.mdst	516000000	2003-03-26 09:40:19.0
VV-uds-e000007-b20011214_0817-22.mdst-1	167506804	2003-03-26 09:40:20.0
kstar0rho0-0125.mdst	148446796	2003-03-26 09:40:20.0
rho+kstar0-0125.mdst	148623948	2003-03-26 09:40:20.0
rho+rho--0125_sigskim.mdst	51706520	2003-03-26 09:40:20.0
rho+rho0-0125.mdst	148289884	2003-03-26 09:40:20.0
rho0kstar+-0125.mdst	149581932	2003-03-26 09:40:21.0
rho0rho0-0125.mdst	147540080	2003-03-26 09:40:21.0
rho0rho0-0125_sigskim.mdst	96716292	2003-03-26 09:40:21.0

File: rho+kstar0-0125.mdst

createTimestamp:	20030326064617Z
creatorsName:	cn=Manager,ou=People,o=Belle,o=Grid
modifiersName:	cn=Manager,ou=People,o=Belle,o=Grid
modifyTimestamp:	20030326094020Z

test.belle: http://fleagle.ph.unimelb.edu.au:8080/datacatint/NavigateFileList.do?collectionid=lc%3Dtest.belle%2Clc%3Dwinton%2C

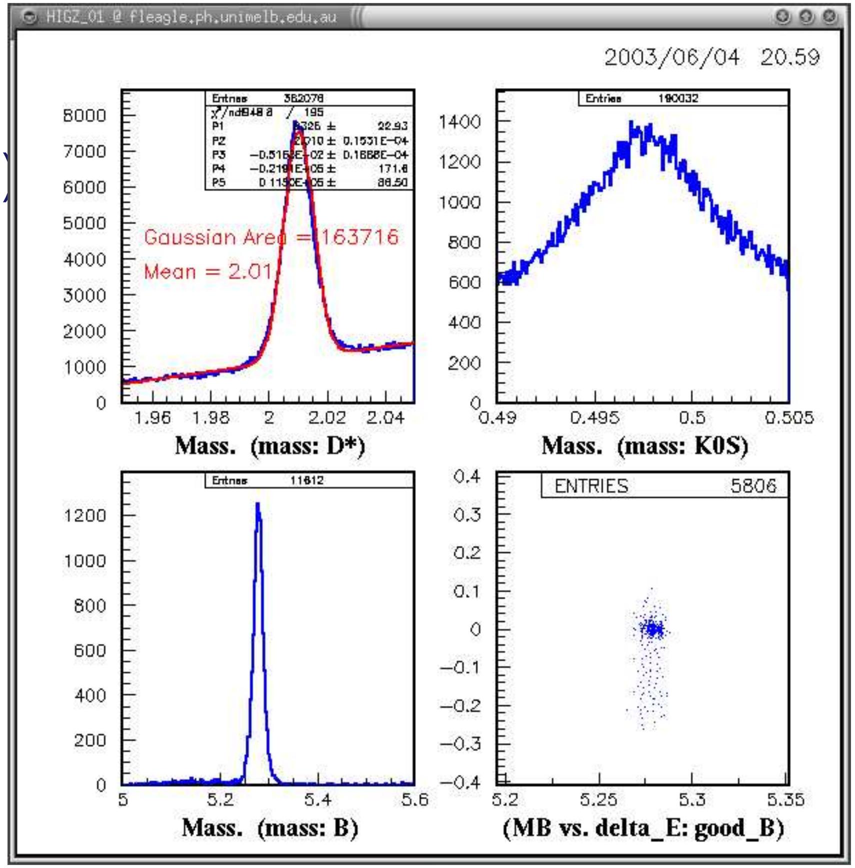
Belle data analysis demonstration at SC2003

The Global Data-Intensive Grid Collaboration

`http://gridbus.cs.mu.oz.au/sc2003/`

- 1,000,000 events analysed using Grid-enabled BASF
- Gridbus broker discovered the catalogued data (lfn:/users/winton/fsimddks/*.mdst) and:
  - decomposed into 100 Grid jobs
  - nodes in Australia and Japan.
- Optimised job assignment to minimise:
  - data transmission time *and*
  - computation time.

Completed in 20 minutes.



We are working on:

- *Robustness*
  - Problems in interface between Globus and PBS
    - Some jobs go missing
- Globus 3?
  - ...with IBM Singapore
- *Metadata specification for Belle data*
  - Reconstructed data
  - Skim files
  - Monte Carlo simulated data
- *Collating results from user analysis jobs*
  - Merge ntuples and histograms

Strategy for the future

Take advantage of new grid computing resources in Australia:

Australian Partnership for Advanced Computing (APAC):

- Coming:
  - 147 node PC cluster (3GHz Xeon)
- Currently:
  - MDSS PetaStore - Direct connect to GrangeNet
  - 150 node PC cluster (2.66GHz Pentium 4)
- Globus 2.4

Victorian Partnership for Advanced Computing (VPAC):

- 97 node, 194 CPU PC Cluster (2.8GHz Xeon)
- Globus 2.4

University of Melbourne

- 48 node, 96 CPU PC cluster (2.4GHz Xeon)
- Globus 2.4

Strategy for the future

Continue development of basic frame work:

- Improve robustness
- Remove vulnerable points of failure

Utilise third party computing resources for Belle:

- Monte Carlo Simulation
- Data analysis

Incorporate new tools as available:

- EDG/LCG tools, SRB, . . .

Work with KEK Computing Research Centre:

- Support broader deployment of a Grid for Belle data analysis