

Remote Web Services for Model Building

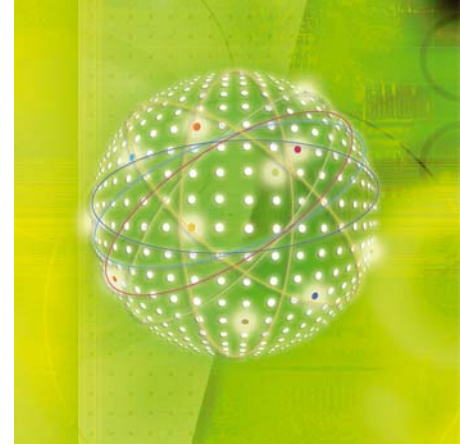


Venkataraman Parthasarathy SPINE/BIOXHIT

Gerrit Langer NIH

Frank Schmitz EMBL

GRID computing



http://en.wikipedia.org/wiki/GRID_computing

Grid computing is an emerging computing model that provides the ability to perform higher throughput computing by taking advantage of many networked computers to model a virtual computer architecture that is able to distribute process execution across a parallel infrastructure.

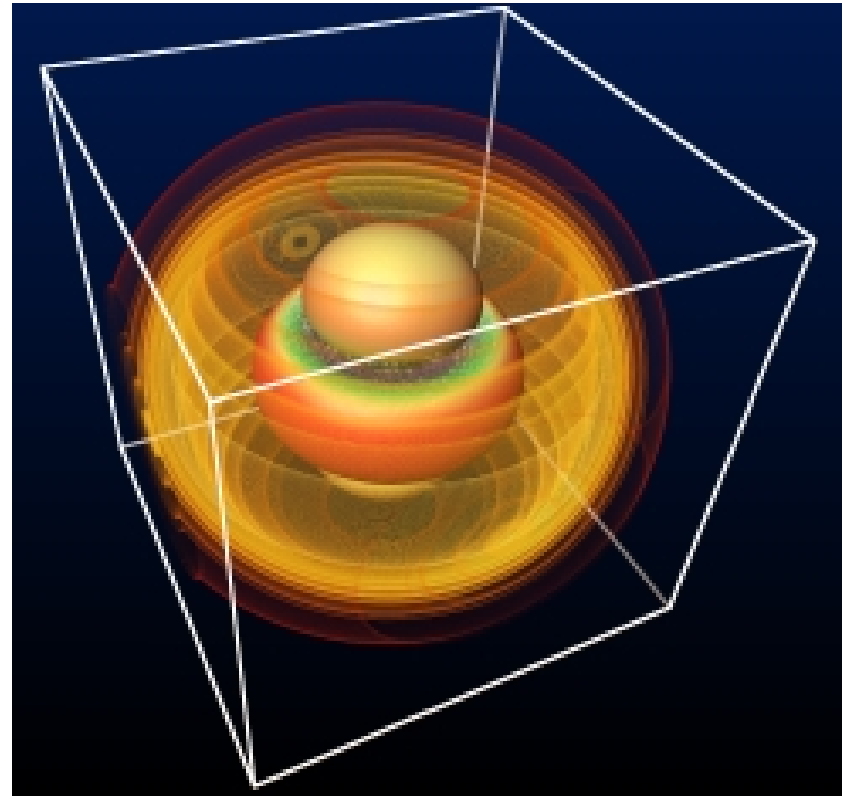
Visualisation of gravitational waves, computed in accordance with Einstein's theory of General Relativity

San Diego Supercomputer Center (SDSC)

National Center for Supercomputing Applications (NCSA)

1,500 processors in total

Cactus Computational Toolkit



Wild dreams for a CPU demand



19ID 400 PDB structures per annum => 2 structures per day
30 min to collect a dataset => 2 datasets per hour **x2 = 2**

One short (helices) AutoRickshaw job takes 30 min
Assume 10 AutoRickshaw jobs per data
In 8 space groups => 5 hours **x5 = 10**

Assume one needs 5 hours to run ARP/wARP to a user's satisfaction.
Assume one wants to run it in 8 space groups
=> 40 hours **x40 = 400**

Assume one wants in case of molecular replacement (2/3rds of the cases)
to test 20 models => 10 hours **x10 = 4,000**

This is 4,000 CPU hours per hour of 1 beamline

Central computer cluster for remote execution of CCP4 and ARP/wARP jobs

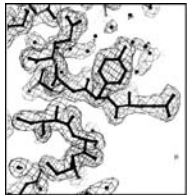


The main aims

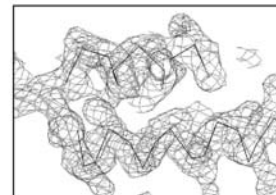
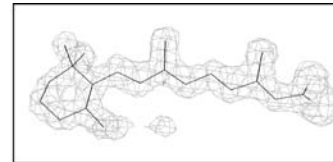
- to allow the users to go far beyond routine data acquisition
- to obtain data for software development and project tracking
- to provide the users with the latest (working !) executables
- to provide the users with the CPU power

Architecture

- 16 processors Linux-i686 2.4 MHz Xeon
- Job queuing and logging system
- Easy job submission



Chain tracing
Ligand building
Helix search
Solvent structure



ARP/wARP and remote submission



CCP4 Program Suite 5.0.2 CCP4Interface 1.3.20 running on lamzin-mac.embl-hamburg.de Project: MIRBUILD

Model Building

- FFFear - Fragment Searching
- FFJoin - Merge fragments
- XtalView/xfit
- ARP/wARP
- ARP/wARP ReBuild
- ARP/wARP LigandBuild
- ARP/wARP HelixBuild

ARP/wARP Version 6.1: Model Building and Density Improvement Initial parameters from /Users/

Job title

Run ARP/wARP for

Dock the autotraced chains to sequence after autobuilding cycles.

MTZ in

Fobs

PHIB

Sequence in

Required parameters

There are total residues in the AU, which belong to molecule(s).

Do cycles of autobuilding (total cycles).

Use for REFMAC5 the protocol and do the Free R flag

ARP/wARP flow parameters

Refmac parameters

Crystal parameters

Submit a remote job at the Hamburg Cluster

Submit the job for remote execution at the Hamburg cluster

Your Email address

Job data

Task viewer



```
Found 100 (152 requested) and removed 0 (152 requested) atoms.  
Cycle 17: After refmac, R = 0.202 (Rfree = 0.000).  
Found 118 (155 requested) and removed 9 (77 requested) atoms.  
Cycle 18: After refmac, R = 0.188 (Rfree = 0.000).  
Found 68 (133 requested) and removed 10 (66 requested) atoms.  
Cycle 19: After refmac, R = 0.179 (Rfree = 0.000).  
Found 57 (121 requested) and removed 35 (60 requested) atoms.
```

Writing map files covering molecule ...

Normal termination of warpNtrace

[Download Results](#)

After you download all the required files click below to indicate that you are done

[DONE](#)

If the job fails or you have any other comments click below to send comments

[COMMENT](#)

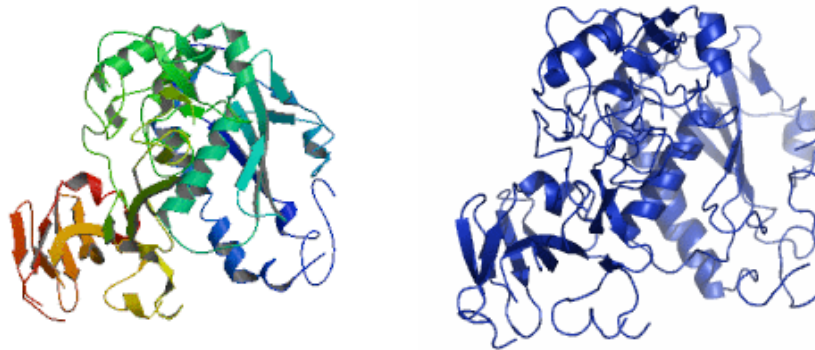


Image created using molscript Image created using PYMOL

Submission via http



Welcome to the web server for remote execution of ARP/wARP model building tasks

The use of the ARP/wARP software requires you to consent to the ARP/wARP and the CCP4 license agreements



[ARP/wARP Web Services](#)

[View the ARP/wARP license](#)

[View CCP4 licensing conditions for academic and commercial use](#)

[I agree to the conditions and wish to proceed with the remote services](#)

Submission via http



Step one

[ARPwARP Web Services](#)

Run ARPwARP for automated model building starting from:

Your email address:

MTZ file:

Submission via http



Step two

ARP/wARP Web Services

Total number of residues in Asymmetric Unit:	<input type="text"/>
No of chemically identical molecules in the Asymmetric Unit:	<input type="text"/>
File with sequence for 1 molecule(optional):	<input type="text"/> Browse...
No of autobuilding cycles:	<input type="text" value="10"/>
Label for FP:	<input type="text" value="FP"/>
Label for SIGFP:	<input type="text" value="SIGFP"/>
Label for PHI:	<input type="text" value="PHIDM"/>
Label for FOM:	<input type="text" value="FOMDM"/>
Dissemination level for data:	<input type="text" value="World (Data can be archived and made available to any software developer who asks for it)"/>

You are required to cite:

Perrakis, A., Morris, R.M. & Lamzin, V.S. (1999) Automated protein model building combined with iterative structure refinement. Nature Struct. Biol. 6, 458-463

CCP4 (1994) Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D50, 760-763.

Murshudov, G.N., Vagin, A.A. & Dodson, E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D53, 240-255.

The service also uses the FFT program from the CCP4 see [ccp4 refernces](#) for references.

Also see [arp-warp site](#) for additional ARP/wARP references.

ARP/wARP 6.1 model building: Status after 20 months

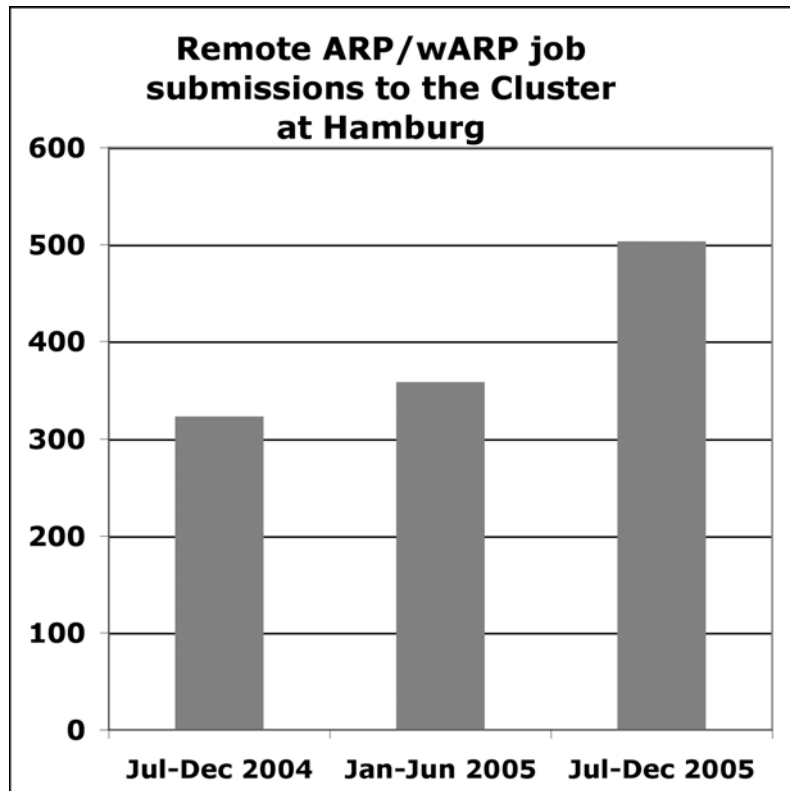


Users Domains

Home downloads 1845 923

Remote jobs 204 120

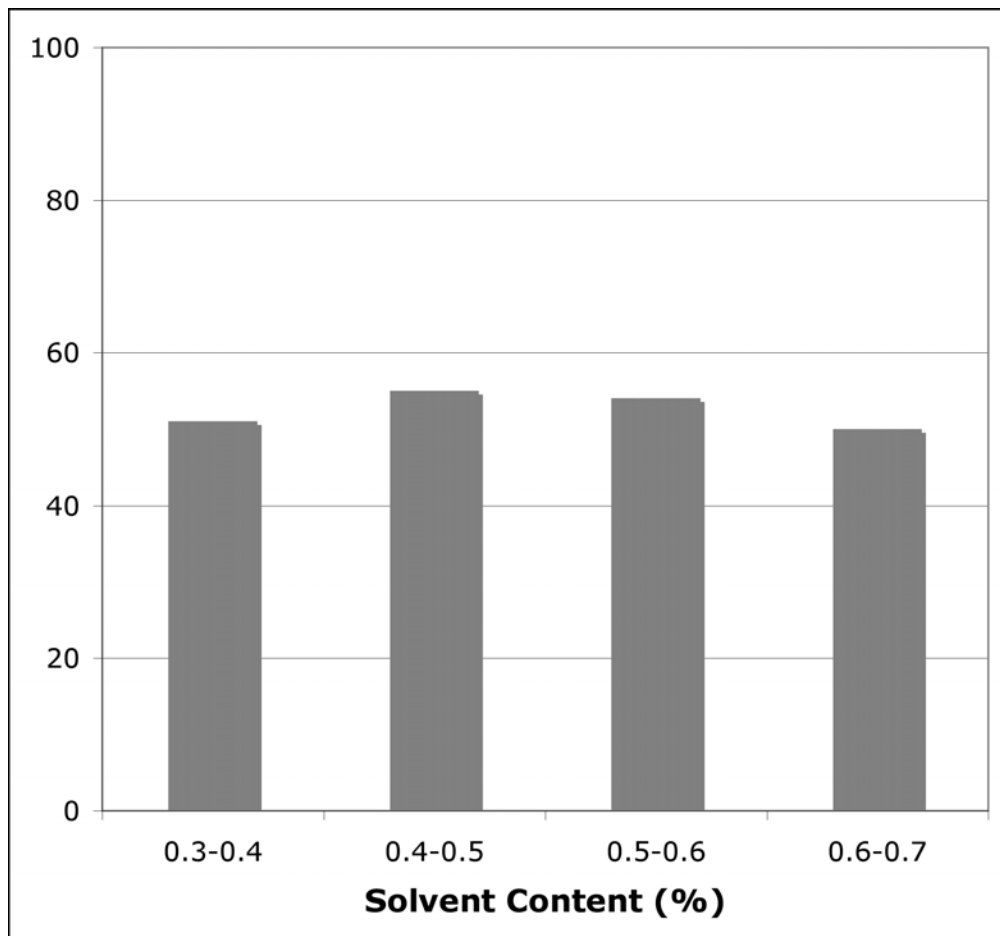
percent 11% 13%



Lessons for core software development



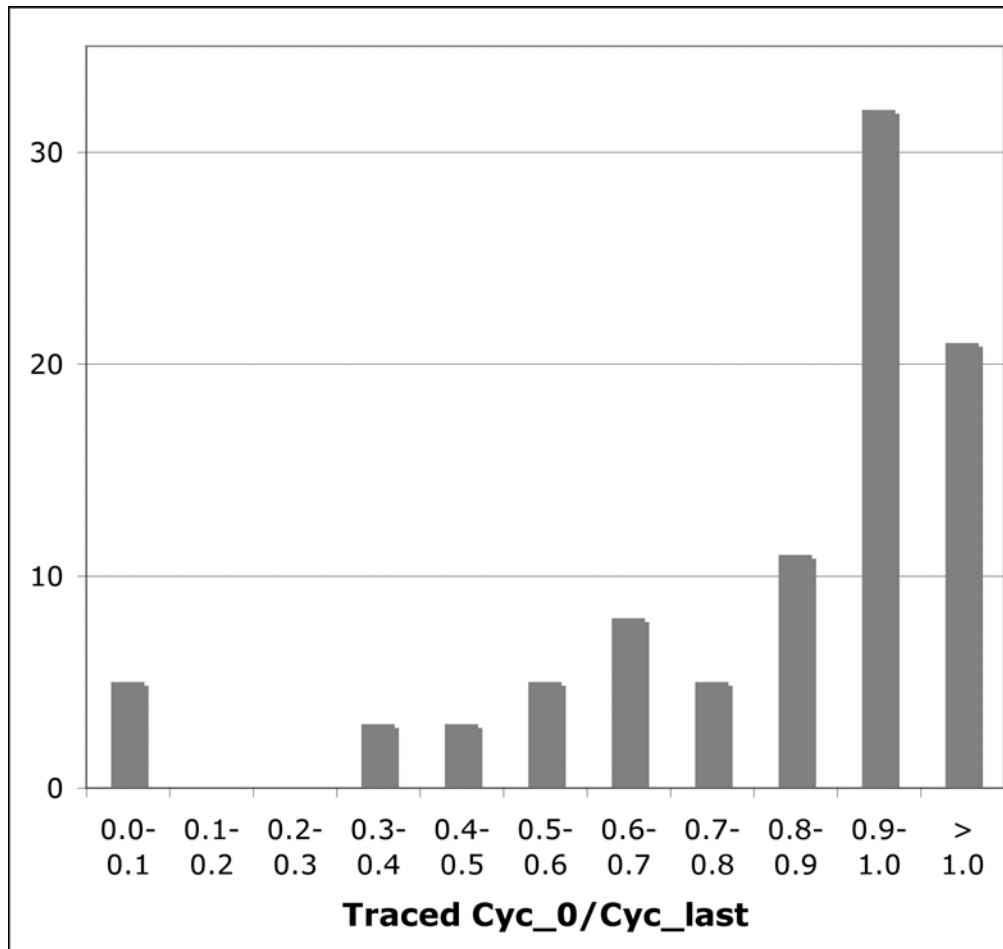
Tracing performance and solvent content - 242 non-redundant jobs



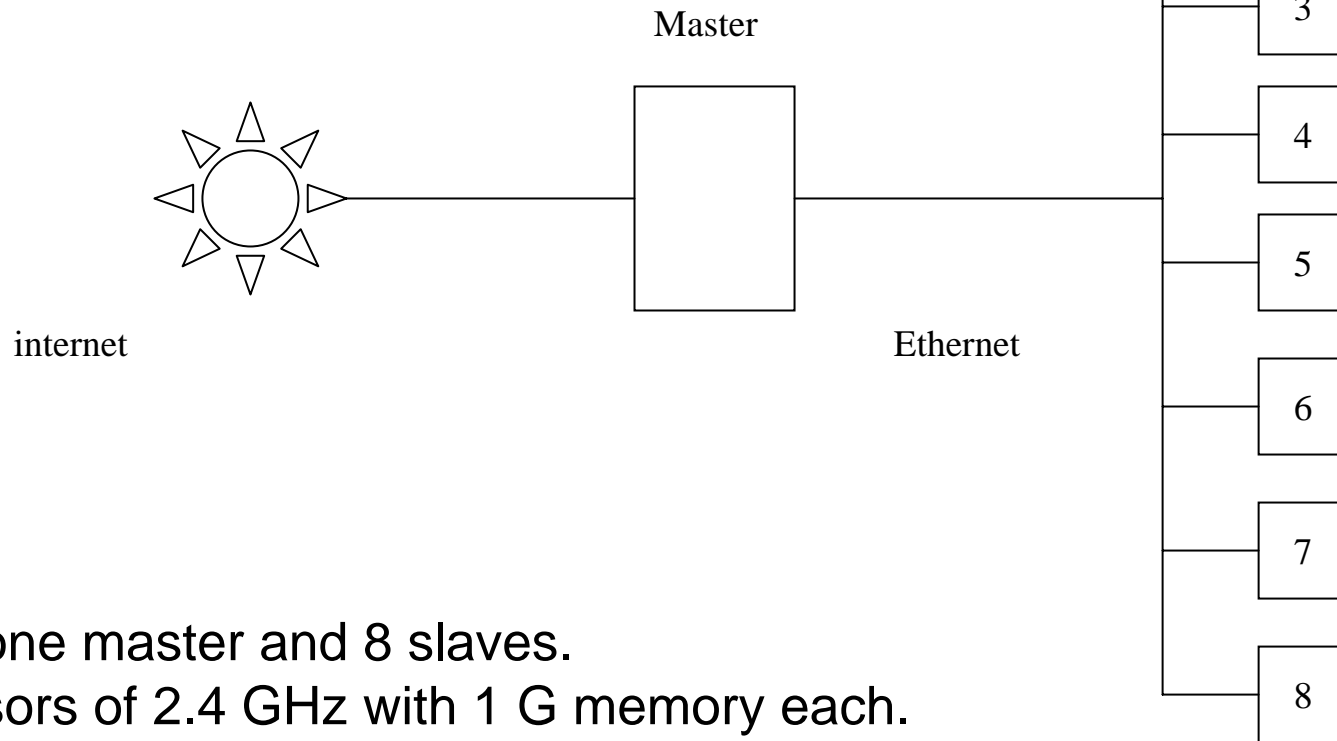
Lessons for protocol development



Initial and final tracing - 93 non-redundant jobs starting from a map



The Cluster at EMBL-Hamburg



Cluster has one master and 8 slaves.
Dual processors of 2.4 GHz with 1 G memory each.

Runs under Red Hat 9.0.
PBSPro5.3 is used for job submission and queuing.
Obtained in May 2003 (NIH funding)

Cluster Architecture for Remote Services



User starts ARPwARP from ccp4 GUI. There is an option for submitting the job remotely.

When that is done a tcl script creates a par file from the input to GUI, then it creates a tar file containing:

- the par file
- the mtz data file
- structure pdb file (optional)
- the sequence pir or seq file (optional)

The name given to the tar file contains the user's computer name and time of submission.

Cluster Architecture for Remote Services



This tar file is then uploaded to the cluster at EMBI-Hamburg using cURL (The connection via cURL to the Cluster is checked when the user installs ARPwARP).

An empty log file with the same name as the tar file is also uploaded to the cluster.

A cronjob is running at the cluster every minute which looks for uploaded files.

When a file is uploaded a random name-number directory is created and the tar file along some shell scripts is copied to the newly created directory, and untars the tar file.

The shell scripts prepares the job for submission to the queue. We use PBSPro for job scheduling.



Cluster Architecture for Remote Services

If the user is local i.e. from EMBL-Hamburg, then the job is launched automatically and an email notification is sent to the user with instructions.

If the user is from outside, an email is sent to the user with instructions:

- URL
- userid
- password

In the CCP4 GUI the user is asked for “keepdata” status. Based on his input, a week after the job was launched the following is done:

- ‘confidential’ - everything is removed except the arp_warp.log, arp_wilson.log and .par
- other than ‘confidential’ - the data is stored in a different directory.

Lessons learned from the use of the Linux cluster



PBSPPro is good for scheduling jobs, have had no problems on scheduling. In most cases it has indicated errors correctly. One case where it has failed to indicate an error was when a few jobs had stopped.

The current setup can be reinstalled on this or any other cluster within a couple of days. All needed files are backed up. There is also a periodic backup of all data (9 DVDs).

We are considering to go for a multi-processor Apple cluster in the near future. It will help in handling increased load when people can use ARPwARP directly via web and AutoRickshaw is opened to the users outside EMBL-Hamburg beamlines. It will help with the security issues.



Lessons learned from the use of the Linux cluster



The following shows type and frequency of stops for ARP/wARP jobs

Pointer to the failure in arp_warp.sh Number of formal stops or crashes Percent

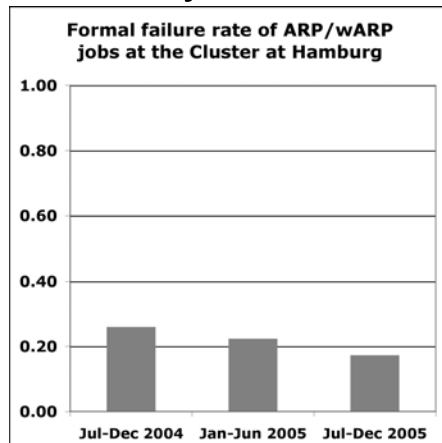
Pointer to the failure in arp_warp.sh	Number of formal stops or crashes	Percent
REFMAC5	110	0.37
SNOW	47	0.16
MAIN_SCRIPT	34	0.11
ARP_MODE_WILSON	29	0.10
FFT	24	0.08
ARP_WARP_BUILD	12	0.04
MAPMASK	10	0.03
ARP_WARP_MODE_UPDATE_BUILD	9	0.03
PEPT_HMAIN	8	0.03
CAD	4	0.01
ARP_WARP_MODE_UPDATE	3	0.01
RESHUFFLE	3	0.01
ARP_WARP_MODE_MIRBUILD	1	0.00
CUBES	1	0.00
MAIN_ARP_WARP_SCRIPT	1	0.00

Lessons learned from the use of the Linux cluster



The following list some atypical things

- Case of par file attributes
- Case when user tries to submit jobs in quick succession
- Email id not entered
- A new queue started thereby killing the first, have changed script to to accommodate that
- Link to a file was uploaded instead of the file
- Job just dies (no error clues found)
- Sequence file format
- History-based empirical formula to estimate CPU required



Status after 20 months



No of unique users	204
No of jobs completed	706
No of jobs with cluster problem	19
No of jobs from AutoRickshaw	55 (8%)
No of jobs with keepdata status WORLD	378
No of jobs with keepdata status EUFW6	8
No of jobs with keepdata status WARPTEAM	228
No of jobs with keepdata status CONFIDENTIAL	433 (40%)

Status after 20 months



Resolution range	0.9 to 3.8 Å
Range of molecular weight	up to 680 kDa
CPU hours provided	3,800
Average CPU per structure	5 hours (equivalent to 800 residues for the 5x10 standard protocol)
Average CPU load	2-3%
Peak CPU load	45% (single user - several jobs)

Along the pipeline... - the AutoRickshaw talk

