MBB KI

PSF Xray Newsletter 5

Quarter 3, 2014

Martin Moche

10/31/2014

The ambition of this newsletter is to create a simple summary of current actions and issues taking place at PSF Xray instead of sending out several emails.

Contents

Protein crystallography pilot project at NSC

Title: A pilot macromolecular 3D structure determination project - Year 2

Current allocation: SNAC Medium i.e. 20 000 core hours / month (Year 2)

Name: SNIC 2014/1-131

Pilot study PI: Martin Moche

Aim of pilot study

In May 2013 a pilot project started aiming at installing protein crystallography software in a HPC environment to investigate performance of such setup with respect to remote graphics, speed of calculation and potential to run supercomputer adapted software in a suitable environment.

Computing and people involved in pilot

Many PIs, postdocs and students have required and been granted access to the pilot setup (Table 1), however in terms of usage only a few have started using it on a regular basis today (Figure 1).

Name	Surname	Group	Dep.	Institute
Adnane	**Achour**	Structural Immunology	SciLifeLab	Karolinska Institutet
Anneli	**Wennman**	Biochemical Pharmacology	Farmbio	Uppsala Universitet
Damian	**Niegowski**	Chemistry II	MBB	Karolinska Institutet
Domnik	**Possner**	Molecular Structural Biology	MBB	Karolinska Institutet
Fatma	**Guettou**	Biophysics	MBB	Karolinska Institutet
Gunter	**Schneider**	Molecular Structural Biology	MBB	Karolinska Institutet
Herwig	**Schuler**	Biophysics	MBB	Karolinska Institutet
Johan	**Unge**	MaxIV and CMPS	BSB	Lunds Universitet
Joseph	**Brock**	Chemistry II	MBB	Karolinska Institutet
Lionel	**Tresaugues**	Biophysics	MBB	Karolinska Institutet
Luca	**Jovine**	Jovinelab	BioNut	Karolinska Institutet
Madhan	**Anandap.**	Structural Biology	IFM	Linköping University
Martin	**Moche**	Protein Science Facility	MBB	Karolinska Institutet
Martin	**Hällberg**	Hallberglab	CMB	Karolinska Institutet
Mathieu	**Coincon**	Drew group	DBB	Stockholm University
Michael	**Raba**	Biophysics	MBB	Karolinska Institutet
Pär	**Nordlund**	Biophysics	MBB	Karolinska Institutet
Rachel	**Lim**	Biophysics	MBB	Karolinska Institutet
Tim	**Schulte**	Structural Immunology	SciLifeLab	Karolinska Institutet
Tobias	**Karlberg**	Biophysics	MBB	Karolinska Institutet

Table 1. Researchers that have been granted access to the pilot study since May 2013.

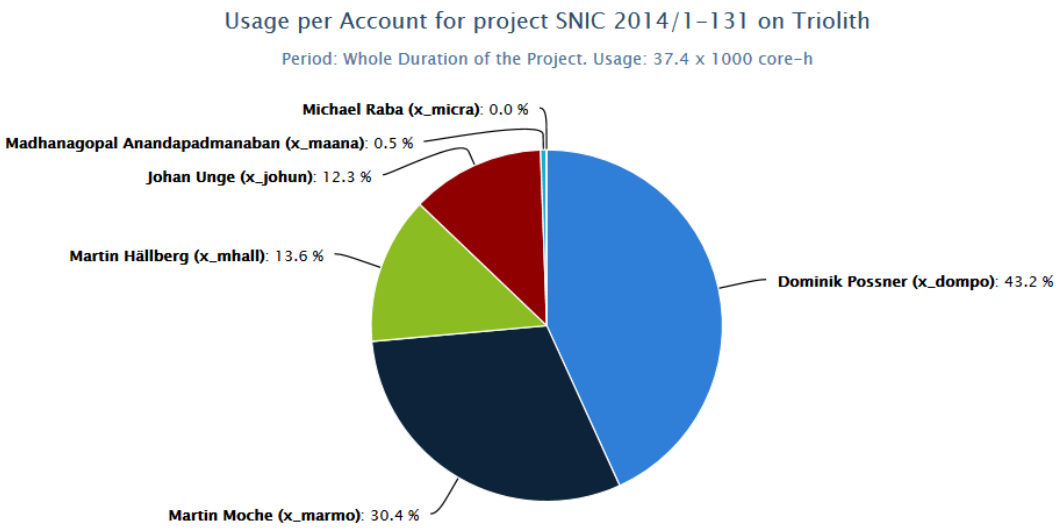

Figure 1. Computational time divided per pilot project user account.

Other benefits to the protein crystallography research group would be access to molecular dynamics and combination chemistry software already available at NSC. The people using most computational time within this pilot study during year 1 was Chemistry II from Karolinska Institutet now having their own NSC allocation outside of this pilot. Despite we are only a few users taking advantage of the current NSC setup we were close to running out of core hours in August 2014 as shown in figure 2.

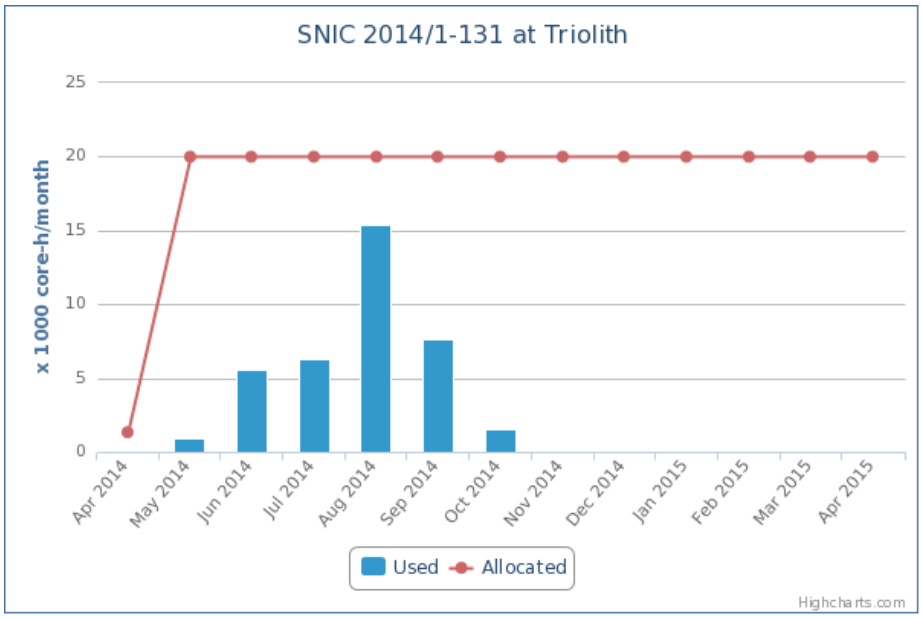

Figure 2. Computational time per month by the pilot study users today. During year 1 our allocation was 5000 core hours/month while todays allocation of 20 000 core-hours/month is sufficient for 2014.

Name	Home Page	PSF	NSC	Purpose	Comment
imosflm	http://www.mrc-lmb.cam.ac.uk/harry/imosflm/ver711/introduction.html	YES	YES	Data collection/processing	Data collection guidance, part of ccp4
xds	http://xds.mpimf-heidelberg.mpg.de/	YES	YES	Data Processing	parallel software
ccp4i	http://www.ccp4.ac.uk/index.php	YES	YES	Data processing/phasing/autobuild/refinement	Broadest software package available
phenix	http://www.phenix-online.org/	YES	YES	Data quality/phasing/autobuild/refinement	phenix GUI v.1.9 has slurm scheduler
shelx	http://shelx.uni-ac.gwdg.de/SHELX/	YES	YES	Data quality/phasing/autobuild/refinement	Find heavy atoms, phasing
cns	http://cns-online.org/v1.3/	YES	NO	Data quality/phasing/refinement	CNS torsion angle SA superior to phenix
xprep	http://shelx.uni-ac.gwdg.de/tutorial/english/xprep.htm	YES	NO	Data scaling and conversion	Prepare for shelx suite
GlobalPhasing (GlPh)	http://www.globalphasing.com/	YES	NO	Data scripting/phasing/autobuild/refinement	#1 phasing and refinement package
-SHARP (GlPh)	https://www.globalphasing.com/sharp/	YES	NO	Phasing by maximum likelihood	autoSHARP for finding HA and autobuild.
-BUSTER (GlPh)	https://www.globalphasing.com/buster/	YES	NO	Structure refinement	leading refinement package
chimera	http://www.cgl.ucsf.edu/chimera/	YES	NO	Free PyMol alternative	PyMol is more developed
ligplot	www.ebi.ac.uk/thornton-srv/software/LIGPLOT/manual/	YES	NO	Ligand picture making	Old program still used
coot	http://www2.mrc-lmb.cam.ac.uk/Personal/pemsley/coot/	YES	YES	Manual model building	Part of ccp4 6.4.0 package
xfit	http://www.duncanmcree.com/xtalview.html	YES	NO	Manual model building	Unsupported since several years
o	http://xray.bmc.uu.se/alwyn/TAJ/Home.html	YES	NO	Manual model building	
hkl2map	http://webapps.embl-hamburg.de/hkl2map/	YES	YES	Phasing GUI	shelx c/d/e GUI
vasco	http://genome.tugraz.at/VASCo/	YES	YES	Pymol plugin: Display surface properties	User friendly surface property displayer
apbs	http://www.pymolwiki.org/index.php/APBS	YES	YES	Pymol plugin: Electrostatic calculations	
usf	http://xray.bmc.uu.se/usf/	YES	YES	Tools for maps/pictures/pdb-edit/data/etc	moleman, rave, dataman, mapman etc.
pymol	http://www.pymol.org/	YES	YES	Visualize and make pictures	Not free software
adxv	http://www.scripps.edu/tainer/arvai/adxv.html	YES	YES	Visualize diffraction data at detector	Frequently used at beamlines
pymol plugins	http://www.pymolwiki.org/index.php/Category:Plugins	some	some	Visualize and make pictures	We should install more plugins!
xdsapp	http://www.helmholtz-berlin.de/forschung/funkma/soft-matter/forschung/bessy-mx/xdsapp/	NO	YES	Data Processing GUI	xds GUI being outdated at PSF
phenix mr rosetta	http://www.phenix-online.org/documentation/reference/mr_rosetta.html	NO	YES	Molecular replacement	runs with phenix 1.8.4 at present
arcimboldo	http://chango.ibmb.csic.es/ARCIMBOLDO/	NO	YES	Molecular replacement	supercomputer software
shake and bake	http://www.hwi.buffalo.edu/SnB/	NO	YES	Phasing with many sites	supercomputer software
Auto-Rickshaw	http://www.embl-hamburg.de/Auto-Rickshaw/	NO	NO	Automated crystal structure determination	EMBL server available
hkl3000	http://www.hkl-xray.com/	NO	NO	Data processing/phasing/autobuild/refinement	Not free software
dials	http://dials.sourceforge.net/	NO	NO	Diffraction data analysis software in development	MaxIV priority
epmr	http://www.epmr.info/	NO	NO	Molecular replacement	requested by NSC pilot user

Table 2. A list of protein crystallography software and their availability at PSF and NSC

Software installation

During the pilot study all protein crystallography software's have been installed by NSC staff within their modular system at their flagship system Triolith (https://www.nsc.liu.se/systems/triolith/) with the HPC scheduling environment provided by a Simple Linux Utility for Resource Management i.e. slurm. Unusual to HPC environments every NSC user have shell access making is possible to install their own pet program for personal use.

The role of the pilot study PI has been to assist NSC staff in making priorities between which software to install and to test and use the installed software. Another important role is to explain dependencies between software's for instance XDSAPP are using routines from phenix and ccp4 packages.

Johan Unge from MaxIV has been running some tests on the current NSC setup and was interested in having dials installed at NSC (http://dials.sourceforge.net/). The code for dials could be read by NSC staff however was not yet released not even in its alpha or beta state (August 2014) and there were no instructions on how to get it running so a decision was taken to wait for a planned beta release in late 2014.

Software testing

Triolith consist of 1600 nodes where each node has 16 processors and at least 32 GB of RAM per node. The majority of protein crystallography jobs are running on a single node where data processing using the parallel software xds is split between the 16 processors thereby running faster than at PSF having 4 processor machines. XDS running speed can be further enhanced provided images are stored at scratch disk of compute node.

Login to NSC is performed using thinlinc using any Windows/Linux/Mac computer and once login and opening a shell one is located at the login-node where graphics applications such as coot and PyMol can be conveniently run remotely. All computation of some magnitude is to be performed at compute nodes so since the majority of protein crystallography software is not supercomputer adapted one has to request a compute node and estimate computational time before computation with e.g. xdsapp/ccp4 can start. Exceptions to this are phenix where the GUI can be set to work with slurm scheduling and spread out a job across more than one node making phenix convenient to use at NSC. Phenix software´s today are however not yet parallel so when using multiple processors in phenix it is used as multiple runs.

Today's protein crystallography setup at NSC is complementary to PSF

Already in its infancy today the NSC protein crystallography setup is complementary and a great asset to PSF software users. The PSF Linux operating system from 2008 is outdated and therefore the PSF software setup is complemented by:

- Running the latest xdsapp version at NSC
- Running phenix MR Rosetta at NSC
- Running supercomputer adapted software Arcimboldo and Shake and Bake at NSC.

In the future PSF computers with large screens will be used as terminals accessing NSC.

Outlook for the future

Future protein crystallography setup at NSC

To develop an integrated HPC structural biology setup including protein crystallography an application called PReSTO have been submitted to Swedish National Infrastructure for Computing (SNIC) and to SciLifeLab. The PReSTO application has been submitted by the director of NSC Patrick Norman and Maria Sunnerhagen professor of Structural Biology in Linköping.

GlobalPhasing headed by Gerard Bricogne develops the leading phasing (sharp) and refinement (buster) software packages for protein crystallography today. GlobalPhasing also developed the grade web server for small molecule geometrical restraints and data processing (autoPROC), autobuild (autoSHARP) and automatic ligand fitting (pipedream) pipelines often dependent on modules from ccp4 and Uppsala software factory. Global Phasing has expressed an interest in adapt their code for supercomputer usage in collaboration with PReSTO.

Buster and sharp is frequently used at PSF so top priority from the PSF perspective is therefore to have sharp and buster available at NSC. Since the SUSHI sharp GUI is based on a http Apache server that is potentially an initial obstacle to NSC, not allowing web browsing, to be solved.

NSC user training for protein crystallographers

The current pilot setup for protein crystallography will improve when more people start to use it. To get more NSC users a training session need to be developed and first run with students and postdocs working at PSF and later with crystallographers at other universities. Using NSC is easy, however the small getting started issues is preventing people from get going analogous to how the PSF protein crystallization setup was mistreated prior to 2007 where non-trained

people were guessing on how to use expensive robots resulting in costly repairs and instrument downtime.

Training session content

To make it more fun users are encouraged to bring their own “active” datasets in need of re-processing.

A suitable number per event is 4-6 people with internet connection and ability to share computer desktops with a beam projector. Then issues that each student has with their own datasets can then be shared with the course participants.

- How to login at NSC

- How to transfer data from synchrotrons/home laboratory to NSC

- How to use login node and computational node

- How can data processing speed can be optimized for tricky data

- Running XDSAPP with users own data

- What protein crystallography software is available at NSC and how to get it run

- Dependent on MR or phasing for the various students/datasets

    - MR: Setup and run phenix and slurm scheduling MR job with user data
    - Phasing: hkl2map with user data

- How to use iRODS

After this feedback from course participants can be collected on what training should be performed further it required.