CCP-EM Working Group 2

Computational Needs Workshop

Harwell, Thursday 29th May 2015

Minutes

Introduction

Alan Roseman introduced the session, and described the background to the Computational Needs Workshop. A preliminary meeting was held in Manchester on 30th April 2015, and minutes are available on the CCP-EM website. A survey of the needs of the community was carried out via SurveyMonkey, and Alan summarised the key results.

Discussion

There was a free-form discussion over several topics. Below, we summarise the main issues and points raised. Individual contributions are indicated by initials only.

Hardware:

- Many universities have purchasing frameworks, which limits the choice of suppliers for clusters. The Computing Insight UK event (formerly Machine Evaluation Workshop) is a good place to meet vendors, and exchange notes with other university purchasers.
- DB bought a 120-node cluster, with 256 RAM per node. There was an initial problem with MPI communication. They have 70 TB storage. The next investment is expected to be GPU cards.
- A suggested rule of thumb was 32 GB per core, at least with no frame processing.

Compute clusters:

- There are different usage scenarios. The LMB have a 7000 node cluster with around 70 users (structural biology and sequencing jobs), and is usually full (LP). The Leeds group use the central university cluster (NR). BB is the single user on a local machine.
- When using a shared central cluster, queueing priorities are an issue. Clearly this is not an issue for a local resource, but then there is likely to be less support for maintenance.
- There is a range of experiences of the quality of university IT support, from good to non-existent. There is often a language problem, and it would be useful to have a glossary of IT terms.

Cloud resources:

- These provide on-demand CPU resources. Spot instances are cheaper, see recent Elife paper.
- Data upload needs to be included in the cost estimate. This would be an issue if micrographs need to be uploaded, but is less of an issue for later processing with particle stacks. For large datasets, it is possible to post a hard drive to Amazon.

Software:

- Software: Based on a show of hands Relion, Spider, IMAGIC and EMAN are widely used (for single particle projects). There were also some users of Frealign and XMIPP.
- AR suggested we create a reference software installation, which others can clone.
- There is a some software that can make of GPUs, e.g. Motioncorr which does so via CUDA.

Relion:

- Popular software, but concern at the CPU usage.
- NR had found that Relion 1.4 is more efficient for frame processing.
- LC asked how one could tell a job was working, if it is taking more than a day. JH pointed out that on some cluster setups, the output is not visible until the job ends.

eBIC:

- Nick Rees spoke on behalf of IT at Diamond. The general plan is for data processing that is linked to data collection to be performed on Diamond/eBIC hardware, while downstream post-processing will take place on STFC hardware. This would be the general model for many science areas, although there is a major drive from the crystallographers.
- JH felt it was enough to do initial processing at eBIC. BB said that Bram at NeCEN wants to automate initial processing.
- Access to eBIC/OPIC requires proof of good initial data.
- Katie Cunnea supports TEM and SEM users at the Research Complex. They can support low end projects, but have no computing resources.
- There is a need for regional centres to help train users, and to help develop projects. How would these be funded? RF said that KCL could provide access to low end microscopes, for a small fee. More users, generating more data, would help justify expanding the compute cluster.

Training:

- There was a preference for hands-on courses, although morning lectures can be useful.
- Webinars would be useful for reaching a wider audience.
- MW wondered if a Python course would be useful, to help people with scripting (many cryoEM packages have a Python interface). The meeting felt that general introductions to Python could be obtained elsewhere, but a small amount could be included in a course on cryoEM software.

Actions

The following set of actions are intended as jobs for CCP-EM to help the community:

1. Find a way to help scientists to talk to local IT staff. This could be a glossary, a list of useful questions to ask, or 1-to-1 advice for particular issues.
2. Curate experiences and typical configurations for compute clusters. This could be done via a wiki, or to begin with a static web page. CCP-EM to solicit contributions from attendees of this workshop.

3. Curate a set of benchmarks for Relion. The length of jobs will depend on the stage of Relion, the size of the dataset, the compute hardware, and the number of MPI processes / threads being used. The benchmarks would therefore be for general guidance, rather than accurate predictions. CCP-EM to solicit contributions from attendees of this workshop.