UPPNEX - A solution for Next Generation Sequencing data management and analysis

Samuel Lampa[1,2], Jonas Hagberg[1], and Ola Spjuth[1,3]

[1]Uppsala Multidisciplinary Center for Advanced Computational Science (SNIC-UPPMAX) , Uppsala, Sweden

[2]Science for Life Laboratory, Uppsala University,Uppsala, Sweden

[3]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

https://www.uppnex.uu.se/

We present a solution for Next Generation Sequencing (NGS) data management and analysis using a cluster-based approach with a shared parallel file system, together with a graphical client and a web-based knowledge base. The initiative is named UPPNEX, and has emerged as the leading platform for the vibrant NGS community in Sweden.

For analysis, 900 000 computing hours per month are available via a cluster of 2784 cores through the SLURM queuing system. For primary storage, more than 420TB of parallel storage are attached locally to the computing resources. For archiving, more than 1 PB of storage is available via the Swedish national long time storage system SweStore. To protect project data, UPPNEX is equipped with snapshots, disaster backup on tape, optional data encryption, and a tight security policy permitting only SSH connections.

To simplify for novice users of HPC systems, we have developed a graphical client for accessing UPPNEX resources based on the Bioclipse workbench1. Bioclipse leverages on the plugin-architecture of Eclipse, which allows for easy extensions and reuse of plugins from a large user community. A proxy component translates information from the local queuing system and exposes a transparent API, which is accessed via a persistent SSH connection provided by the Eclipse Parallel Tools Platform. Via Bioclipse, users can access their files via a graphical file browser, they are able to monitor jobs, inspect file and project quotas, and start new analyses. The Bioclipse-plugin takes advantage of the tool configuration files from the Galaxy platform to provide wizard-based configuration of cluster jobs for common bioinformatics tools, but users can also interact with UPPNEX via a regular terminal. A history view allows for inspecting the commands sent to UPPNEX and enables reuse and sharing of analysis scripts.

Apart from hardware and software, the UPPNEX project has several associated human resources ("system experts" and "application experts") serving the national NGS community with experience and know-how in both HPC and bioinformatics analysis via the UPPNEX Knowledgebase web portal3. The distinct focus on end-users has attracted over 130 projects in only 2 years at an increasing rate, and UPPNEX is currently serving over 400 TB of NGS data with the sequencing of 'Norwegian spruce' as one of its largest projects.

UPPNEX was originally funded by the Knut and Alice Wallenberg foundation and the Swedish National Infrastructure for Computing (SNIC) and is formally part of Uppsala Multidisciplinary Center for Advanced Computational Science4 (SNIC-UPPMAX).

References

1. Spjuth et al. *Bioclipse: an open source workbench for chemo- and bioinformatics*, BMC Bioinformatics 2007, 8:59.
2. Giardine et al. *Galaxy: a platform for interactive large-scale genome analysis*. Genome Res. 2005 Oct;15(10):1451-5.

Relevant Web sites

3. https://www.uppnex.uu.se/
4. http://www.uppmax.uu.se/