Training and Education in High Performance Computing for eResarchers

Lev Lafayette
V3 Alliance, Melbourne, Australia, lev@v3.org

The need for high performance computing in eResearch is increasing, especially for conducting parallel processing on very large datasets. Datasets are growing in size due to the increasing scale of information-gathering devices and traditional tools for analysis; desktop unicore applications, relational databases etc, are simply not capable of providing research needs within a reasonable time and instead require parallel processing on high performance computing clusters and grids [1]. However the necessary skillset - the command line interface, job submission, scripting, parallel programming - is not common among researchers and training is not generally available. At least part of the cause of this is due to an assumed conflation of high performance computing with scientific computing. Whilst the critical issue of 'big data' makes it clear that this conflation is desired, it should not be assumed. The assumption, with its concentration on hardware and software capabilities, rather than the actual practise and input of the scientific and eResearch community has led to a situation where critics can argue that high performance computing is actually harmful [2].

As a result, there is a growing disparity between needs and competence which reduces research output. This disparity will become a critical issue for organisations claiming eResearch capability in the future with those that provide comptency training that match needs successfully adapting to the new environment. If the eResearch community do not have the opportunity to utilise parallel processing and high performance computing, they will continue to work with their preferred applications on desktop environments. As a result relative research output will decline, as empirically illustrated in studies of research productivity and investments in HPC [3]. This will prove to be a critical issue for such institutions, in the sense that the survival of the research capabilities of the organisation is at stake [4].

Two broad methods exist for achieving such a match; (i) modify the HPC environment to suit the existing skillset or (ii) develop the skillset to match the HPC environment. There has been significant develop in the former area, especially championed by software developers and management who want to simplify job submission tools. Well-known examples include xpbs, grid computing interfaces such as the former Grisu project, distributed computing installers such as folding@home, or even from the direction of applications developing parallel capacity, such as Matlab's DCS and parallel computing toolbox. It is arguable that however that many of these projects provide better grounding to the eresearchers or return on investment to a management perspective. Even the provision of user-friendly trivial submission tools remains challenging because parallelisation and high performance computing - like any other form of genuine computer literacy requires a degree of understanding of

the process. Without the grounded understanding the eResearcher will be learning (and relearning) applications. The alternative is to provide a graduated trainig that provides both the skillset for HPC utilisation but also implicit learning adaptable for future situations.

For past several years the Victorian Partnership for Advanced Computing (VPAC), and the successor organisation, V3 Alliance, have conducted a range of training courses designed to bring the capabilities of postgraduate researchers to a level of competence useful for their research. These courses have developed in this time, in part through providing a significantly wider range of content for varying skillsets, but more importantly by introducing some of the key insights from the discipline of adult and tertiary education in the context of the increasing trend towards lifelong learning. This includes an androgogical orientation, providing integrated structural knowledge, encouraging learner autonomy, self-efficacy, and self-determination, utilising appropriate learning styles for the discipline, utilising modelling and scaffolding for example problems (as a contemporary version of proximal learning), and following up with a connectivist mentoring and outreach program (including "hackathons") in the context of a culturally diverse audience.

The experience and development of these training programs serves as a case study of these principles. Starting from technically accurate but minimal and decontextualised training courses, the courses have developed over time for a significantly wider range of base skillsets and with a increasingly higher level of modularity, integration, mentoring and outreach, especially in the context of a highly multicultural research community. Current course content begins with basic CLI Linux, file transfers, environment variables, job submission and PBS scripts. An intermediate course covers more advanced CLI commands, regular expressions, scripting, and more elaborate job submissions (interactive, dependencies, arrays), basic compiling, with the advanced course providing a deeper theory of parallel computer architecture, and MPI programming fundamentals. The success of these courses correlated with improved cluster usage in terms of the quantity of job submissions and the number of users. With a larger and trained eResearch community future plans for an accredited graduate certificate in high performance computing becomes possible, thus further positioning the eResearch community as whole to better cope with future needs.

REFERENCES

[1] Adam Jacobs, The Pathologies of Big Data, Queue, Association for Computer Machinery, July 6, 2009
[2] Greg Wilson, High Performance Computing Considered Harmful, 22nd International Symposium on High Performance Computing
Systems and Applications, 2008
[3] Amy Apon, Standley Ahalt, Vijay Dantuluri, et. al., High Performance Computing Instrumentation and Research Productivity in U.S. Universities, Journal of Information Technology Impact, Vol 10, No 2, pp87-98, 2010

[4] For an elaboration of the concept of a critical issue, see Jurgen Habermas, Legitimation Crisis, Beacon Press 1975 (FP 1973)

ABOUT THE AUTHOR

Lev Lafayette is a HPC systems administrator, advanced computing educator, quality assurance coordinator, and project manager for V3 Alliance. Prior to V3, he worked as the ICT consultant for the Ministry of Foreign Affairs and Cooperation in Timor-Leste, in their first year of independence, and as a computer systems trainer for the Parliament of Victoria. He has an Honours degree in Politics, Philosophy, and Sociology, and a Graduate Certificate in Adult and Tertiary Education, both from Murdoch University, and a Masters of Business Administration (Technology Management) and a Graduate Certificate in Project Management from Chifley Business School. He is also an accredited PRINCE2 practitioner.