Computational Research Methods for Science Education at Various Levels

Name: Mahak
Phd Scholar
(Department of Computer Science)KUK

Abstract
The role of Computational Science research methods teaching to science students at PG level is to enhance their research profile developing their abilities to investigate complex problems, analyze the resulting data and use adequately HPC environments and tools for computation and visualization. The paper analyses the current state and proposes a program that encompasses mathematical modelling, data science, advanced algorithms development, parallel programming and visualization tools. It also gives examples of specific scientific domains with explicitly taught and embedded Computational Science subjects.

Keywords: Computational Science Research Methods, Postgraduate Education, Science Subjects

➢ The Need

Current expectations from university education are to prepare Postgraduate (PG) level students for careers both in academia and industry. This became now even more relevant in the sciences where the needs of knowledge based economy with increasing number of smart jobs require science graduates to be “research ready” for a dynamically changing and computationally heavy project environment using High Performance Computing machines, environments and tools. The professional training courses currently available to computational scientists, or those established just few years ago, were targeting mainly computer science graduates, and although they are able to provide some practical skills which can resolve short term programming deficiencies or introduce some simulation techniques, they are not designed to substitute university education.

Large scale computing in science and industry has become an indispensable way to tackle societal and scientific Grand Challenges, and to address the needs of industry to innovate in terms of products and services (COM EC, 2012) and (Joseph, E. et al, 2010, September). The major societal drivers such as Energy, Climate Change, Urbanization, etc. (Luebkeman 2009), require advanced scientific methods and in particular Computational Science and HPC (High Performance Computing) are the key methods for helping address these challenges. The platforms allowing us to advance on such scale and magnitude are the Exascale Computing systems. Recently, Exascale Computing, an attempt to harness a thousand fold projected increase in computational power, has emerged as a Grand Challenge research area (Dongarra, J. et al, 2009). In addition, the development of novel mathematical methods and scalable algorithms for exascale in itself has been identified as a Grand Challenge (Dongarra et al, 2014).

Computational approaches to scientific Grand Challenge problems such as the detection and treatment of diseases like cancer, modelling of the human brain, and climate forecasting are beginning to bear fruit. Computational Science, an interdisciplinary field that melds basic sciences, mathematical modelling, quantitative analysis techniques, algorithms, parallel programming and tools and HPC techniques, is proving integral in addressing the big problems in industries ranging from manufacturing and aerospace, to drug design and risk management.

In the Energy area, the focus is on combustion, nuclear fusion, clean energy which includes solar energy, etc.. Al these require high-fidelity simulation of combustion phenomena, efficient particle and fluid simulation including fine-scale turbulence, for solar energy developing of models to model complex systems (specifically nano-systems with tens of thousands of atoms dynamically for long time (Simon et al 2007).

Biotechnology, bioinformatics and breakthroughs in healthcare technology are also a key focus areas both in Europe (Horizon 2020), USA (NECCEA, 2011) and Latin America (Alexandrov et al, 2014). The focus here, for example, is on advanced methods and technologies for preventive medicine as well as creating complete DNA sequencing for major diseases, (see, for example, Horizon 2020 and NECCEA, 2011).

In Socioeconomic modelling the focus is on integrated modeling of the social, economic and environmental systems coupling these elements and treating uncertainties and non-linearities (Simon et al 2007). For example, the challenges here are: how will different adaptation and mitigation strategies affect energy supply and demand, the overall economy, the environment , the public, public health etc.; how the demographic change, economic growth etc is connected to the above etc .(Simon et al 2007).

➢ Current State

The Recent SIAM (Society for Industrial and Applied Mathematics) working group report (SIAM Working Group on CSE Education, 2014) and the USA DoE Assessment of Workforce Development Needs in office of Science Research Disciplines (Chapman et al, 2014), as well as our own assessment have shown that despite increasing acceptance of Computational Science as interdisciplinary science there is still a lot of silos based approaches to university education and the compartmentalised culture towards curricula development and course design are preventing us from really truly bearing the fruit of implementing Computational Science research methods into variety of science degrees at time when we are at a critical junction when the research and industry are beginning to experience a stifling effect of lack of properly educated and trained specialists in Computational Science (Chapman et al, 2014) and (SIAM Working Group on CSE Education, 2014). We are at a crossroad where we need to decide what the way forward would be: poaching the few available ones from each other if we can, or re-design science degrees at PG level where the cycle is shorter and more flexible to update and thus provide the right calibre graduates.

Firstly, there are simply too few Master's and PhD programs in Computational Science and related areas (Scientific Computing, High Performance Computing, and Supercomputing).

.

Secondly, many of the existing Computational Science programs (sometime despite what their promotional literature says) tend to be highly focused on one part of the “Computational Science pipeline”. They produce highly skilled personnel in mathematical modelling or HPC computing techniques but they don’t expose them to the pipeline from basic science, through modelling and simulation, HPC application design, implementation and evaluation – all in combination with data-intensive computing. While technical depth is important, we believe the leadership in interdisciplinary areas like Computational Science, requires a breadth of experience and knowledge.

Thirdly, existing programs often expose students to a narrow range of application domains. For example, they may have great depth in matrix based and/or fluid-flow problems but do not expose students to applications based on string or geometric data (bioinformatics) or large scale graphs (web analytics). This narrowness not only limits an individual’s employment opportunities, it also impedes cross-fertilization between Computational Science application domains. This is also linked with the silos culture and compartmentalization dominant in many academic environments.

Fourthly, most existing programs in Computational Science don not have the opportunity to expose students to the practice of science in a multi-country/multi-cultural setting. This is a wasted opportunity. The practice of science, especially in the context of Grand Challenge problems, is becoming steadily more global. These problems are simply too big for single countries to tackle without global collaborations. Even industrial R&D teams now typically span multiple sites in multiple countries and time zones. We believe that learning to work effectively outside one’s own cultural home is an essential skill in being a productive team member or leader of large scale R&D efforts.

Finally with the emergence of Data Science focused on dealing with Big Data an additional dimension emerges that requires universities and research institutions to establish a common Data Science competencies profile and a common modularized (component-based) curriculum for education and training to the required job profiles .

Similar is the situation at doctoral level. The problem here is particularly acute since in many cases the PhD students are usually selected from the corresponding MScs and since these are clearly much underrepresented, here is even increasing shortage.

Analyzing overall situation it appears that the major problem still remaining is compartmentalization.

The **major drawbacks** of such programs have been identified, at least in USA, that they do not provide exposure to real-world applications, e.g. students are not able to grasp the complexities in the field. In general the universities are not adequately preparing students with the right skills to become tomorrow’s computational scientists and engineers, the study also stated that “the current programs do not teach students the skills essential to apply Computational Science and Engineering in modern scientific and technological enterprises.” It was also pointed out that there was “almost no university that have or are likely to develop curriculum focused on topics associated with petascale and exascale science.”

It is time to accept that a way forward could be an introduction of Computational Science into the core subject area of PG science degrees, not just computer science and/or mathematics graduates but as well those in life, earth and natural sciences are inevitably going to use HPC methods, tools and environments in their working life and a part of their education should prepare them to at least understand and implement, if not design, parallel algorithms and be familiar with the HPC machines and HPC tools and environments.

➢ Computational Science Research Methods

The graduates require set of skills to tackle both compute and data intensive applications independently if they will follow a career in academia or industry. In particular, the Grand Challenges outlined above require knowledge and expertise in mathematical modelling (multi-model, multi-scale, modeling continuous processes (PDEs, ODEs, etc), modelling discrete events), discretization techniques, advanced algorithms development (parallel algorithms, scalability, numerical and non-numerical algorithms), software implementations on variety of advanced architectures, program execution, tools for analysis, visualization etc, data analysis, visualization and validation of the results.

Data Scientists or data science teams on the other hand focus on solving complex data problems by employing deep expertise in one or more of these disciplines, as well as business strategy and domain knowledge. Personal skills in communication, presentation and inquisitiveness are also very important. In terms of skills a Data Scientist is seen as "a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle.

From point of view of computational scientists working in Supercomputing area, the body of knowledge the students need to acquire to be able to tackle the above Grand Challenges and to be able to successfully tackle project work is expected to cover:

1. **Mathematical modelling and algorithms**: requiring systematic approach to modelling (complex systems - introduction, overview, etc.), modelling continues events/systems (Ordinary Differential Equations (ODEs), Partial Differential Equations (PDEs) and systems of PDEs, multi-level, multi-scale methods) and discrete events modelling, numerical analysis (Linear Algebra, Optimization, etc.), discretization, stochastic numerical methods and stochastic modelling, advanced algorithms (numerical and non-numerical including parallel algorithms and parallelization techniques);
2. **Programming Environments and tools**: programming languages, including parallel programming approaches, advanced programing models and tools, variety of performance tools for parallel programs/computing including for HPC;
3 **Data Analysis**: stochastic and optimization methods for data analysis, data visualization techniques (scientific visualization, where appropriate Virtual Reality approaches, etc.

To be able to bridge the skills gap there is a need to change the current education programs at University PG level as well as the professional training. It requires paradigm shift in terms of introducing advanced mathematical modelling methods and scalable algorithms (stochastic, deterministic and hybrid ones) allowing change of thinking in terms of high levels of parallelism for exascale and beyond; advanced programming models and tools allowing advanced modelling and simulation (including simulation and modelling at scale); scientific visualization of data and Big Data processing.

The challenges are to integrate these key components in the curricula as separate subjects as well as embed components in domain specific subjects for maximum impact.

Example of above proposed approach is given below:

Scientific Domain: Environmental Sciences

- Mathematical modelling: Complex Systems; Systems approach to modelling; Multi-level multi-scale methods; PDEs and ODE.
- Programming Environments and tools: Parallel programming; MPI and OpenMP; Performance Analysis;
- Data Analysis: Stochastic and Optimisation Methods for data analysis; Visualisation;
- Embedded in the Dissertation Project Module: Introduction to Earth Science Simulation Environments, Performance tools .

Scientific Domain: Life Sciences

- Mathematical modelling: Complex Systems Modelling and Simulation; Stochastic
- Programming Environments: Parallel Programming, MPI, OpenMP, CUDA;
- Data Analysis: 3D Visualisation of Data, Bid Data Analysis;
- Embedded in the Dissertation Project Module: Simulation Environments for Life Sciences; Parallel and Distributed Programing Models, Virtual/ Augmented Reality.

Scientific Domain: Material Science

- Mathematical modelling: Complex Systems Modelling and Simulations; Stochastic
- Programming Environments: Parallel Programming, OpenMP;
- Data Analysis: 3D Visualisation of Data, Stochastic and Optimisation Methods of Data Analysis;
- Embedded in the Dissertation Project Module: Parallel and Distributed Programming Models, Virtual/Augmented Reality.

➢ Conclusion

Computational Science research skills gap is evident from variety of reports and studies, for example, (Chapman et al, 2014) and (SIAM Working Group on CSE Education, 2014). IDC has also identified the skills which are most difficult to find to tackle the inflection points, for example: Scientists with HPC capabilities ("Combined scientific background and HPC programming skills", "Computational scientists"), Parallel Programmers ("Experience in parallel software development", "Engineers and scientists that can program in HPC/parallel Fortran" , "Parallel code porting/optimization"), Algorithm Developers ("For computational science people who can help researchers develop and implement new algorithms"), System Administrators with high-end computing experience "Scientific computing system management experience", "System administrators with HPC expertise"), see (Joseph et al, 2010, July) and (Joseph, E. et al, 2010, September).

The graduates require set of skills to tackle both compute and data intensive applications independently if they will follow a career in academia or industry. In particular, the scientific Grand Challenges outlined, require knowledge and expertise in mathematical modelling, advanced algorithms development, software implementation on variety of advanced architectures, program execution, data analysis, visualization and validation of the results.

Apart from the Computational Science degrees that are taught in Computer Science and Mathematics departments, the graduates of MSc courses taught at domain science departments also must be addressed. It is necessary to increase the general understanding in Computational Science research methods. This can be achieved by introducing explicitly taught modules on mathematical modelling, HPC methods, tools and programming as well as provide opportunities to train graduates in practical skills as part of their dissertation project.

References

- Big Data White Paper (2013) Unleashing the potential of Big Data. A Whitepaper "World Summit on Big Data and Organization Design, 2013 [Online] http://www.essi.upc.edu/~aabello/publications/ 13.Whitepaper.pdf [Accessed: February 2015].
- Big Data Analytics (2014), Assessment of demand for Labour and Skills 2013-2020, [Online] https://www.thetechpartnership.com/globalassets/pdfs/bigdata_report_nov14.pdf [Accessed: February 2015].
- High Level Expert Group on Scientific Data, (2010, October). Riding the wave: How Europe can gain from the rising tide of scientific data. [Online]. Available from: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf [Accessed: February 2015].
- Horizon 2020, The EU Framework Programme for Research and Innovation, [Online] Available from: http://ec.europa.eu/programmes/horizon2020/ [Accessed: February 2015].
- Joseph, E., Conway, S., and Wu, J. (2010, July) *A Study of the Talent And Skill Set Issues Impacting HPC Data Centers Conducted On Behalf Of The US Department Of Energy*. IDC.
- Joseph, E. et al. (2010, September), *A Strategic Agenda for European Leadership in Supercomputing: HPC2020 – IDC Final Report of the HPC Study for the DG Information Society of the European Commission*. IDC. Hey,T., Tansley, S. and Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/ [Accessed: February 2015].
- Luebkeman, C. (2009) and the Foresight team at ARUP (Jennifer Greitschus eds.), *Drivers of Change*, Prestel Verlag and Arup, 2009, ISBN 978-3-7913-4224-5.
- Manyika, J. et. al (2011) Big data: The next frontier for innovation, competition and productivity, Report, McKinsey Global Institute, [Online] www.mckinsey.com/insights/business_technology/ big_data_the_next_frontier_for_innovation [Accessed: February 2015].
- NECCEA, (2011), National Economic Council, Council of Economic Advisers, and Office of Science and Technology Policy, *A Strategy for American Innovation: securing our economic growthand prosperity, February 2011*. [Online] Available from: http://www.whitehouse.gov/sites/default/files/uploads/InnovationStrategy.pdf [Accessed: February 2015].
- SIAM Working Group on CSE Education (2014),*SIAM: Graduate Education for Computational*
- *Science and Engineering* [Online] Available
- http://www.siam.org/students/resources/report.php,[Accessed: February 2015].
- Simon, H., Zcharia, T. and Stevens R. (2007) *Modelling and Simulation at the Exascale for Energy and the Environment, US Office of Science.* [Online] Available from: http://science.energy.gov/ ~/media/ascr/pdf/program-documents/docs/Townhall.pdf. [Accessed: February 2015].