Demystifying the Dark Arts of HPC - Introducing Biomedical Researchers to Supercomputers

1st Andrea Townsend-Nicholson
Division of Biosciences
Research Department of Structural and Molecular Biology
University College London
London, UK
a.townsend-nicholson@ucl.ac.uk

2nd David Gregory
Division of Biosciences
Department of Computer Sciences
University College London
London, UK
d.gregory@ucl.ac.uk

3rd Art Hoti
Division of Biosciences
Research Department of Structural and Molecular Biology
University College London
London, UK
art.hoti.17@ucl.ac.uk

4th Cristin Merritt
Alces Flight Limited
Bicester, UK
cristin.merritt@alces-flight.com

5th Stu Franks
Alces Flight Limited
Bicester, UK
stu.franks@alces-flight.com

***Abstract*—The ability to bring personalised medicine from science fiction to science fact and into everyday clinical practice is the focus of the CompBioMed Centre of Excellence - a group of private, public and research organisations led by University College London. To reach this goal, CompBioMed's Education, Training and Sustainability team teaches university students studying a broad range of disciplines to run biomedical and bioscience applications on High Performance Computing (HPC). For this, students acquire skills in core computing methodologies, learning the 'dark arts' of optimisation to produce efficient, effective workloads. CompBioMed has created a scalable training model for remotely-delivered 'dark arts' teaching, starting with optimisation of the QIIME2 application on a cloud cluster they have named nUCLeus. This talk covers the collaborative efforts that CompBioMed has used to develop a repeatable scalable training model and how students around the world can successfully learn to use HPC without breaking the bank.**

***Index Terms*—HPC in the Cloud, HPC Training and Education Strategies**

I. Introduction

The CompBioMed Centre of Excellence focuses on turning the concept of personalised medicine into a reality. With computational methods in biomedical research starting to make the leap from the desktop into HPC the CompBioMed team readily engages in training that intertwines the computational science of biomolecules and HPC. But with social distancing restrictions now inhibiting traditional models of on-site education and a continuing requirement to deliver training (due to the need to improve digital skills globally, together with ensuring a state of preparedness for the incoming exascale technology), the CompBioMed team were looking for innovative new ways to keep the education program moving forward. The idea of cloud HPC entered the discussion and with this, concepts around repeatable, scalable training models began to evolve.

Project Funded by the Alces Flight HPC Community Outreach Project. CompBioMed (675451) and CompBioMed2 Centres of Excellence (823712) are funded by the European Commission for Horizon 2020.

Starting with a rigorous assessment of what could be safely taught in pandemic conditions, and bringing in the right skills through collaboration, the process of building up a scalable education in biomedicine has begun under nUCLeus. This persistent, scalable cloud cluster, built through CompbioMed's Associate Partnership with Alces Flight, is the foundation for sustainable training initiatives - beginning with QIIME2 workflows. This open-source bioinformatics pipeline software application was an ideal candidate for remote, cloud HPC education due to its non-compute-intensive nature and the recent proliferation of microbiome studies, which lie within the field of computational biomedicine. Over the course of three months of foundation work, followed by a live training session, the team pulled together a centralized knowledge base of what to do, the scale at which it could be achieved and the level of investment required. The result is an affordable, project-driven cloud HPC model that can move beyond QIIME2 workflows into continuing education models for all those interested in HPC training.

II. Dark Art vs Black Box – The Biomedical Sciences and HPC

The idea of personalised medicine has been immortalized within the science fiction realm. From a quick scan the correct patient treatment is assessed and undertaken almost simultaneously, returning the protagonist to the story in an effortless manner. While such stories may currently be fantasy, researchers are actively investigating methods to achieve diagnosis and personalised treatment in an almost identical manner. Thanks to recent advances in the capabilities and availability of HPC and Artificial Intelligence (AI), concepts that once sat on a desktop are now being ported to and used on HPC clusters. QIIME, and its successor QIIME2, developed as desktop applications, are just two examples of what can be used to build large-scale, automated workflows that leverage

HPC. A key research topic is how the processing of these workflows can be automated and optimised as students move from running them on smaller workstation-class machines to much larger, more capable HPC clusters – something that is currently considered a 'dark art' from the perspective of a researcher with little familiarity with HPC. HPC has the potential to be as much a tool of the trade as the pipettes and test tubes that biomedical researchers use in their day to day practice, if researchers are appropriately introduced to its use and understand how to optimise application parameters to best meet the needs of their scientific question. Biomedical researchers engaging successfully with computational biology must understand how to efficiently utilise HPC in order to utilise the technology and its software to progress their field.

III. Project-Driven HPC: Building nUCLeus

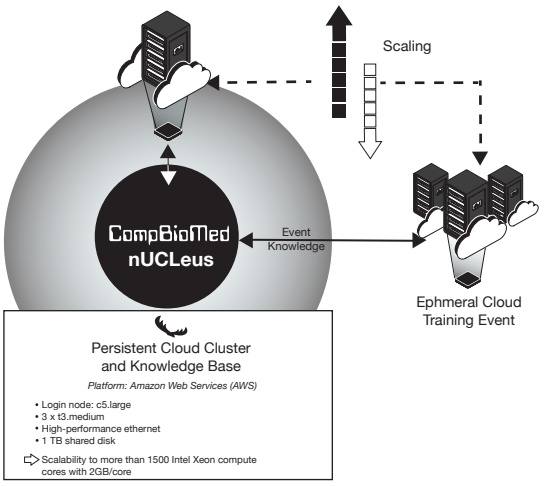

Fig. 1. Basic Design and Function of nUCLeus

The concept of the nUCLeus research environment (or "cloud cluster") was borne out of the need to scale training both within and outside the membership remit of the CompBioMed team. With forthcoming exascale initiatives requiring a larger, more collaborative approach the team came to the conclusion that:

- Tethering training and education initiatives to a specific, fixed site would not be optimal for growth and sustainability
- Tethering work to a specific cloud platform would also have the potential to hinder growth and sustainability

As a result, they chose to collaborate with Associate Partner Alces Flight to build a cloud HPC cluster that functions in a 'platform agnostic manner' - allowing platform requirements to sit in a secondary function under the primary function of educational problem solving.

To save costs when the nUCLeus research environment is not in use, the cluster's persistent idle functionality is negligible. To complement this, the materials and educational IP developed by CompBioMed will reside in a central, web-accessible location. The researchers selected Alces Flight Center, a single portal for managing HPC resources, to form a core of knowledge around how to undertake scalable training and education initiatives and, once instructed, to move those initiatives into active clusters (either cloud or site-based). Federated with Flight Center, nUCLeus initially functions as a testing site for each educational build, allowing the team to create points where the students are welcome to test their capabilities - taming the 'dark arts' to tune QIIME2 and its workloads to address the scientific question under study. Once the parameters for training are established, nUCLeus scales out into a production cluster that students engage with for the duration of their project. When the project is complete, information on the training event is stored, improvements prioritised and nUCLeus returns to an idle state, waiting for the next event.

IV. Growth and Sustainability

By building up nUCLeus and creating a knowledge hub through Alces Flight Center the CompBioMed team is pulling together a system built for growth and sustainability. By holding their knowledge next to a flexible, scalable, persistent cloud HPC cluster the team are able to build out training that not only meets technical requirements, but also creates predictable cost patterns for training – a feature that will be of benefit for incorporation into budget planning for future education programs.

A. Running Cost Projection - CompBioMed Training Initiative

Item	Persistent Cost	Project Cost
Alces Flight Center - Core Knowledge Base Subscription - Persistent nUCLeus Cluster - Managed Service	£1,300 / month	
Cloud Platform Credits: Use as needed (Assumption of 6-8 events per calendar year)		£6,000 - £8,000 (per project)
TOTAL PER YEAR (6-8 events)	∼£16,000	∼£50,000
ONE OFF EVENT	∼£4,000	∼£6,000

Cost range for 6-8 events over a 12-month period: £52,000 to £80,000
Estimated cost per QIIME2 event: ∼£10,000

V. Lessons Learned and Conclusion

While the impact of this project is still being assessed our top three take-aways have so far been:

1) Ensure you have a clear understanding of the problem you wish your students to solve before engaging with a platform.
2) Clearly identify and assign job roles within your project to complement, not cause conflict, in the creation and execution of your training event.
3) Create an accessible central repository of knowledge in order to allow assets to be reusable.

The CompBioMed team aspires to build an education model for HPC. Through collaboration and focus on flexible learning, it is our aim to demystify HPC for clinicians and biomedical researchers in order to move closer to the goal of achieving accessible personalised medicine for all.