# HPC Carpentry: Recent Progress and Incubation Toward an Official Carpentries Lesson Program

Andrew Reid
National Institute of Standards and Technology
andrew.reid@nist.gov

Trevor Keller
National Institute of Standards and Technology
trevor.keller@nist.gov

Alan O'Cais
University of Barcelona
alan.ocais@cecam.org

Annajiat Alim Rasel
BRAC University
annajiat@bracu.ac.bd

Wirawan Purwanto
Old Dominion University
wpurwant@odu.edu

Jane Herriman
Lawrence Livermore National Laboratory
herriman1@llnl.gov

Benson Muite
Kichakato Kizito
benson_muite@emailplus.org

Marc-André Hermanns
RWTH Aachen University
hermanns@itc.rwth-aachen.de

## ABSTRACT

The HPC Carpentry project aims to develop highly interactive workshop training materials to empower novices to effectively leverage HPC to solve scientific and technical problems in their domains. Modeled after The Carpentries training programs, the project's goal is to develop foundational HPC skills and a sense of empowerment, rather than expertise. The workshop setting provides learners with hands-on experience that elicits confidence working with HPC systems and provides sufficient vocabulary to make subsequent self-study more effective.

In a major milestone, the steering committee is leading HPC Carpentry through the formal incubation process to become an official Carpentries lesson program alongside the existing Software, Data, and Library Carpentry programs. This achievement is the product of significant work over the past several years, incorporating valuable materials from many contributors. Our most recent focus has been developing materials for a user workshop. We begin with an introduction to the command-line shell (using Software Carpentry's Unix Shell lesson), followed by our Introduction to HPC lesson, covering remote access and resource management. We end with a newly developed lesson on HPC workflow management, which walks learners through the execution of a scaling study on an HPC system, emphasizing both the benefits and limitations of the system for domain applications. This workshop program was recently run in full at the Lawrence Livermore National Laboratory.

Future plans include building a developer workshop, reconnecting with disparate contributors, and engaging with the broader community through regular open conference calls and outreach.

## KEYWORDS

Cyberinfrastructure, Training, Pedagogy, HPC, Parallel Computing, Big Data, Non-degree Training, Hands-on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright ©JOCSE, a supported publication of the Shodor Education Foundation Inc.

© 2025 Journal of Computational Science Education
https://doi.org/10.22369/issn.2153-4136/16/1/7

## 1 BACKGROUND

HPC Carpentry [12] is an informal training project with a mission to provide a set of lessons aimed at introducing the basic "know-how" of running applications on high-performance computing (HPC) resources to new audiences, including investigators from fields which are not traditional users of HPC systems, as well as novice users from fields in which HPC is commonly used. Eventually, the project's goal is to empower HPC novices to effectively leverage HPC to solve scientific and technical problems in their respective domains. The project paves the way for the potential users from non-traditional HPC disciplines to tap into HPC resources for their data analysis, modeling, and simulation needs while remaining relevant for beginners from the traditional HPC disciplines.

The project is patterned after the broader Carpentries [11] project, and adopts many of that organization's priorities, including an emphasis on accessibility and open-source materials. The current project is the product of significant work over the past several years, incorporating valuable materials from many contributors.

### 1.1 History

Efforts we have drawn on include materials from Compute Canada (now the Digital Research Alliance of Canada), and Peter Steinbach's "HPC In A Day" lessons. The community had numerous important discussions at a birds-of-a-feather event at the SC17 conference, at CarpentryCon in 2018, and another birds-of-a-feather session at the SC18 conference.

The current leadership team coalesced around the content at the end of 2019, when competing demands absorbed the attention of many previous contributors. During the height of the COVID-19 pandemic, the new team had useful discussions with the Carpentries curriculum development team at CarpentryCon 2020@Home,

where the idea of seeking to become a new lesson program alongside Software Carpentry, Data Carpentry and Library Carpentry began to be seriously considered. Several of our lessons benefited from learner feedback following workshops [15], especially a semester-long HPC workshop piloted at BRAC University [7]. This proof-of-concept began in 2020 to provide the students of Parallel, Distributed, & High-Performance Computing (HPC) and Distributed Computing Systems their first exposure to HPC through these open-source lessons.

In 2021, a number of HPC Carpentry lessons were moved into the Carpentries Incubator, and the strategic plan for workshop development began to come into focus. To gather community perspectives, a bird of a feature session was held at SC'21 [17]. The project had a significant presence at CarpentryCon 2022, including a sprint that began serious development on the workflow lesson

Since the SC23 conference session on HPC Carpentry – A Scalable, Peer-Reviewed Training Program to Democratize HPC Access [18], we have become much better connected with the broader HPC education community, which has already led to new contributors and workshop opportunities.

## 2 LESSON DEVELOPMENT EFFORTS

The recent focus of HPC Carpentry has been the development of a complete workshop program for new HPC users.

### 2.1 HPC User Workshop

The HPC User workshop consists of three lessons and targets HPC novices. Though we see variety in learner profiles, an example learner might come to an HPC user workshop with some background in a domain science and a research problem that requires the use of HPC or cluster resources; that user may or may not have former programming experience in a laptop or desktop environment. The workshop aims to cultivate basic HPC skills and a sense of self-efficacy for post-workshop learning; additionally, key themes of the workshop are automation, reproducibility, and reusability.

We begin with an introduction to the command-line interface using Software Carpentry's Unix Shell lesson [16]. This lesson gives learners a basic familiarity with command-line operations in the shell, the structure of Unix commands, and the use of arguments to control the behavior of the command, as well as basic operations on hierarchical file systems.

This lesson is followed by our Introduction to High Performance Computing lesson [13], covering remote access and resource management. Starting where the Unix Shell lesson leaves off, it begins by introducing clusters and how to connect to them over SSH. Learners then explore remote resources and differences from traditional desktop machines, *e.g.*, distinguishing between "login" and "compute nodes". They learn to launch and investigate jobs on compute nodes using Slurm, the resource manager used throughout this lesson, and to customize their software environment through the module system. This lesson also introduces how to execute parallel tasks over distributed resources using an MPI-parallelized mini-application called 'amdahl' that mimics the scaling behavior of a parallel program retaining serial sections.

We end with a newly developed lesson on HPC workflow management [14], which walks learners through the specification and execution of a scaling study on an HPC system, emphasizing both the benefits and limitations of HPC systems for domain applications. Instead of writing a series of Slurm job scripts that run the `amdahl` binary with different parameters, learners automate a workflow that sweeps their scaling study's parameter space using a workflow tool (either Snakemake or Maestro).

This set of three lessons is sufficient to offer a two-day hands-on workshop in a format similar to that of the Software Carpentry or Data Carpentry workshops.

### 2.2 HPC Developer Workshop

The project plans to develop more advanced training resources for HPC developers. Whereas target learners for the HPC User Workshop may only need to run pre-built software for their work and research, target learners for the HPC Developer Workshop are expected to have software development tasks and experience programming outside an HPC environment. In this workshop, instructors will have the option to substitute the workflow lesson with a coding exercise in a parallel framework such as MPI, for example.

The content for the HPC Developer Workshop's lesson on parallel programming will likely grow from existing sources. For example, several relevant lessons, including an introduction to parallel programming using the Chapel programming language, already exist within the Carpentries Incubator. A lesson using Programming Big Data with R (pbdR) is also in development [9]. Furthermore, we have received engagement from HPC community members at large who explore the potential of merging their in-house lessons into HPC Carpentry's lesson portfolio. This is still an ongoing effort and engagement with the community.

### 2.3 Customizing Workshop Materials

Lesson material customizations may be desired to better reflect the configuration of a workshop host's HPC resources or to better address the needs of a particular learner audience. There are two principle ways that the lessons can be adapted to local resources.

Firstly, the HPC Intro lesson has a "snippet library", which allows the rendered lesson to be customized to use the correct resource manager for a site (Slurm, GridEngine, or Torque) and to present the correct machine names and site-specific command arguments. Making these customizations reduces cognitive load for the Instructor during the workshop and provides correct commands for learners revisiting the content afterwards.

Secondly, these lessons are modular. Individual modules can be dropped if they are inappropriate for a particular site. Because these lessons are primarily meant to be presented in real time to learners, instructors can and should "read the room", or consult with helpers, to decide when to slow down or dive deeper into particular parts of the material.

In keeping with Carpentries best practice, all of this material is open-source and freely licensed (CC BY-4.0). We share the Carpentries' commitment to continuous improvement of the lessons

through learner and instructor feedback. This mechanism also allows others from the community to adapt and re-use the material as they see fit.

### 2.4 Building Community

Additionally, we plan to boost our engagement both with the Carpentries and the HPC community through specific interactions at Carpentries events, such as the CarpentriesConnect'24 in Heidelberg, Germany. Beyond the core curriculum—proposed in the Lesson Program Incubator—we want to discuss with the community, which other lessons exist, are in development, or even planned, that are relevant to the HPC Carpentry, and how to integrate them within HPC Carpentry. To enable such discussion, lesson development is done in the open on GitHub and there are bi-monthly online calls. This will be highly valuable for enabling a broader topic range for the proposed HPC Developer Workshops and ensures in-depth engagement of additional instructors and lesson developers as well as keeping the available topics relevant.

## 3 RECENT WORKSHOPS

The complete HPC Carpentry workshop program for new users was recently offered online for the BioNT project in February of 2024 [10] and at Lawrence Livermore National Laboratory in June of 2024. The HPC Intro lesson was offered as a stand-alone unit at the National Institute of Standards and Technology in July of 2024. Previous workshops were held at University College Dublin, BRAC University, Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS), University of Mauritius, Florida International University, Delft University of Technology, National Institute of Standards and Technology, and EPFL CECAM.

Feedback from these workshops has been crucial in improving the lesson material.

### 3.1 Gathering Attendee Feedback

In keeping with the Carpentries' methods, HPC Carpentry workshop feedback is solicited in three primary ways. First, prior to the workshop, a pre-workshop survey gathers information about the attendees' skill levels and familiarity with programming concepts. The instructors can then attempt to present workshop content at the level best suited to each unique audience. Second, throughout the workshop, users are asked to use colored sticky notes to indicate whether they are keeping up with the instructor or if they require assistance catching up from one of the non-lecturing instructors circulating the room. At the end of each in-person session, instructors will also sometimes prompt their learners to leave written comments on sticky notes (typically, one good thing, one bad thing). These methods help to capture learner feedback with little overhead and while it is still fresh, presenting the instructors with realtime feedback on the pacing and clarity of their instruction as well as the quality of the content. Third, a post-workshop survey gathers information about the attendees' workshop experience and their level of comfort with the concepts treated in the workshop. These post-workshop responses allow us to modify our material and approach for future workshops.

### 3.2 Lessons Learned

From the run-up to the BioNT workshop, we discovered some important version-specific updates in the workflow tool we chose to use (Snakemake), which required significant modifications to the draft lesson to accommodate.

From the Lawrence Livermore workshop, we heard from learners that the process of building up a workflow configuration file is sensitive to a loss of context: if a learner misses a step, it's hard to recover, because the next version of the file depends on a consistent prior version. Having a shared virtual notepad with "checkpoint" versions of the relevant files can help learners recover the needed context and not lose the thread of the lesson. The instructor team used such a notepad to catch learners up in real time during the latter part of the LLNL workshop, as well as using the `give` tool to transfer files in their entirety [6].

The NIST lesson was run for members of that institution's Summer Undergraduate Research Fellowship program. Learners self-selected based on a published invitation, and the instructor team found it beneficial to teach the lesson more slowly and with deeper attention to the details of the commands than had been the case at preceding Livermore workshop. This provided a very striking example of adapting the lesson to the room, which is a significant benefit of live workshops.

### 3.3 HPC Infrastructure

One important issue we are facing in offering HPC Carpentry workshops is the need for HPC infrastructure for learners to use during the workshop. While some HPC site operators have their own HPC systems to conduct their own workshop, others, particularly from under-resourced institutions, do not have their own HPC resources. More recently, we have acquired support from Jetstream2 through ACCESS to set up a "standard" reference HPC Carpentry cluster in a virtual-machine-based environment. The cluster set-up has been prepared in an automated fashion using the Magic Castle tool [4], built on the Terraform cloud provisioning tool [5]. This effort could pave the way to allow instructors to configure their own clusters, irrespective of the existence of a local HPC cluster in their own institution. This provisioning scheme is itself open-source, and significantly lowers the barrier to HPC operations for anyone in the community.

A cluster alone is not enough, one also needs a software stack. Magic Castle provides the user with a complete HPC cluster software environment including a Slurm scheduler, a Globus Endpoint, JupyterHub, LDAP, DNS, and over 3000 research software applications compiled by experts. With respect to a scientific software stack, Magic Castle provides two primary options: the Digital Research Alliance of Canada (the Alliance) software stack [2], and the European Environment for Scientific Software installations (EESSI) [3]. Both of these are distributed via CernVM-FS [1] and can be easily utilised by the participants after the workshop.

## 4 CARPENTRIES LESSON PROGRAM INCUBATION

In a major milestone, the steering committee is leading the project through the formal incubation process towards becoming an official Carpentries lesson program alongside the existing Software, Data,

and Library Carpentry programs. This process is expected to last for about 18 months, with expected completion in December 2025. We hope to establish a strong community, a useful set of lessons and a good governance process by that time.

## ACKNOWLEGMENTS

We acknowledge the infrastructure support hosted on Jetstream2 cloud environment. Allocation to Jetstream2 was provided through ACCESS, which is funded by the US National Science Foundation.

Alan O'Cais is co-funded by the European High-Performance Computing Joint Undertaking (JU) and countries participating in the project under grant agreement No 101093169 (MultiXscale), and the European Union under grant agreement number 101100604 (BioNT).

The semester long trainings for the students of Parallel, Distributed, & High-Performance Computing (HPC) at BRAC University [7] through HPC Carpentry Workshops were made possible since 2020 using the HPC resources sponsored by the European initiative LearnHPC [8] and the recent JetStream2 arranged by Alan O'Cais, Intel Dev Cloud, Amazon Cloud and other resources arranged by Benson Muite, Oracle Cloud arranged by Annajiat Alim Rasel.

## REFERENCES

[1] Jakob Blomer, Carlos Aguado-Sánchez, Predrag Buncic, and Artem Harutyunyan. 2011. Distributing LHC application software and conditions databases using the CernVM file system. In *Journal of Physics: Conference Series*, Vol. 331. IOP Publishing, Bristol, United Kingdom, 042003.

[2] Maxime Boissonneault, Bart E. Oldeman, and Ryan P. Taylor. 2019. Providing a Unified Software Environment for Canada's National Advanced Computing Centers. In *Practice and Experience in Advanced Research Computing 2019: Rise of the Machines (Learning) (PEARC '19)*. Association for Computing Machinery, New York, NY, USA, Article 51, 6 pages. https://doi.org/10.1145/3332186.3332210

[3] Bob Dröge, Victor Holanda Rusu, Kenneth Hoste, Caspar van Leeuwen, Alan O'Cais, and Thomas Röblitz. 2023. EESSI: A cross-platform ready-to-use optimised scientific software stack. *Software: Practice and Experience* 53, 1 (2023), 176–210. https://doi.org/10.1002/spe.3075

[4] Félix-Antoine Fortin, Etienne, Frédéric Fortier-Chouinard, Maxime Boissonneault, Charles Coulombe, Erik Sundell, Simon Guilbault, Alan ÓCais, Étienne Parent, R. P. Taylor, jboschee, Kenneth Hoste, Adam Huffman, Chris Want, Darren Boss, Justin Lagüe, Marie-Helene Burle, Quoding, and Pier-Luc St-Onge. 2024. Magic Castle. https://doi.org/10.5281/zenodo.10953488

[5] Mitchell Hashimoto. 2014. Terraform. https://www.terraform.io

[6] Shawn Instenes. 2017. give/take. https://github.com/hpc/give

[7] B. Muite, A. A. Rasel, M. I. Hafiz, and M. A. Rahman. 2020. Two Decades Making an Impact, Brac University Annual Report 2020. , 34 pages. https://www.bracu.ac.bd/sites/default/files/resources/ar/Brac%20University%20Annual%20Report%202020-compressed.pdf

[8] A. O'Cais. 2021. LearnHPC - Scalable HPC Training. https://www.learnhpc.eu/

[9] G. Ostrouchov, W.-C. Chen, D. Schmidt, and P. Patel. 2012. Programming with Big Data in R. http://r-pbd.org/

[10] The BioNT Project. 2024. An introduction to High Performance Computing. https://www.cecam.org/workshop-details/an-introduction-to-high-performance-computing-1270

[11] The Carpentries Project. 2024. The Carpentries. https://carpentries.org

[12] The HPC Carpentry Project. 2024. HPC Carpentry: Teaching Basic Skills for High-Performance Computing. https://www.hpc-carpentry.org/

[13] The HPC Carpentry Project. 2024. hpc-intro: Introduction to High-Performance Computing. https://carpentries-incubator.github.io/hpc-intro/

[14] The HPC Carpentry Project. 2024. hpc-workflow: HPC Workflow Management with Snakemake. https://carpentries-incubator.github.io/hpc-workflows/

[15] The HPC Carpentry Project. 2024. Past Workshops related to HPC Carpentry. https://www.hpc-carpentry.org/past-workshops/

[16] The Software Carpentry Project. 2023. Software Carpentry: the UNIX shell. https://doi.org/10.5281/zenodo.7886938

[17] A. Reid, T. Keller, A. O'Cais, and A. A. Rasel. 2021. HPC Carpentry: Introducing New Users to HPC. https://sc21.supercomputing.org/presentation/index-id=bof161&sess=sess403.html

[18] A. Reid, A. O'Cais, T. Keller, W. Purwanto, and A. A. Rasel. 2023. HPC Carpentry – A Scalable, Peer-Reviewed Training Program to Democratize HPC Access. https://sc23.supercomputing.org/proceedings/workshops/workshop_pages/ws_bphpcte106.htmlandhttps://jocse.org/downloads/jocse-15-1-6.pdf