HPC CARPENTRY

Andrew Reid, NIST

Nov. 18, 2024

SIGHPC Education BPHTE24

CO-AUTHORS

- Trevor Keller, NIST
- Alan O'Cais, University of Barcelona
- Annajiat Alim Rasel, BRAC University
- Wirawan Purwanto, Old Dominion University
- Jane Herriman, LLNL
- Benson Muite, Kichikato Kizito
- Marc-André Hermanns, RWTH Aachen

OUTLINE

- The Carpentries as a solution to education challenges in a research setting
- HPC Carpentry, its intended audience and relation to The Carpentries
- How we got here – a history of HPC Carpentry
- Where we're going – our existing content and strategic plan
- Challenges – some of what we want to do is hard
- Communications – how to get involved

THE CARPENTRIES – THE PROBLEM IT SOLVES

The Carpentries is fundamentally motivated by the problem of increasing availability of, and reliance on, computational methods to address research problems, in the absence of instruction on best practices in utilizing computational resources.

In an academic setting, this results in transient contributors (students and post-docs) rolling their own idiosyncratic solutions, solving the domain problem, and leaving behind inadequate or irreproducible tools.

THE CARPENTRIES – THE METHOD

The initial Carpentries effort was "Software Carpentry", founded by Greg Wilson in the 1990s.

The central idea is that the skills gap can be closed and better practices taught in a workshop setting.

- Hands-on with the actual research computers
- Work in real time alongside an instructor
- Build muscle memory and vocabulary to facilitate later self-study

THE CARPENTRIES

THE CARPENTRIES – THE METHOD

The Carpentries method is not arbitrary, it's grounded in scientific literature about how people learn things.

Carpentries lessons themselves are open-source, and are continuously developed in response to learner and instructor feedback.

The Carpentries organization has formal instructor training, which familiarizes instructors with the relevant educational literature, and strongly emphasizes empathy with learners.

THE CARPENTRIES

The initial "Software Carpentry" effort focused on an introduction to the command-line shell and scripting and automation for workflows, and then on coding in commonly-used programming languages, and version control.

New lesson programs, Data Carpentry and Library Carpentry, arose afterwards and in 2018, all of these programs were united under the umbrella of "The Carpentries".

The common thread is hands-on workshop instruction for research-relevant technical problems, with a strong emphasis on empathy for learners.

THE CARPENTRIES

HPC CARPENTRY – RESEARCHERS

In the HPC world, we see that small clusters are becoming more common and less expensive to acquire, and cloud providers give push-button access to interconnected computers.

The lower barrier to availability has not made it easier to make effective use of these systems.

Researchers face technical barriers to tapping in to the power of available resources, and often do not have time to engage with formal training, because of domain-specific educational demands, or mid-career time constraints.

HPC CARPENTRY – OPERATORS

HPC facility operators also find themselves facing a more diverse set of users.

Opening up the system to a "long tail" of novice users lets operators benefit a larger audience, and up-skilling this audience can save on support hours.

This is especially important for small research clusters, where the "cluster operations staff" may be a fraction of one person's time. In this regime, having low-barrier access to training materials can be a huge benefit to cluster operations.

HPC CARPENTRY – OUR SOLUTION

The parallels with The Carpentries are compelling.

- Hands-on workshop material uses researcher time efficiently
- Researchers benefit from better use of HPC resources
- Lessons support researchers from many domains

We share the Carpentries' fundamental conceit – open-source, pedagogically-grounded, hands-on training materials open the HPC world to researchers from a wide range of research domains.

HPC CARPENTRY – "EARLY" HISTORY

The current HPC Carpentry effort is the result of many threads coming together.

- Contributions from ComputeCanada (now the Digital Research Alliance of Canada)
- Peter Steinbach's "HPC In A Day"
- The SC17 BoF event – Andy Turner, Christina Koch, Tracy Teal, Bob Freeman, Chris Bording
- CarpentryCon 2018 conversations
- SC18 BoF – Andy Turner, Christina Koch, Peter Steinbach, Alan O'Cais, Jeffrey Stafford, John Simpson, Daniel Smith, Bob Freeman
- Informal SC19 session – Andrew Reid, Trevor Keller, others

HPC CARPENTRY – HISTORY

In early 2020, Peter Steinbach, who owned the relevant repos, stepped back from his role. A new governance model was built, and content development continued.

- CarpentryCon 2020@Home – Trevor Keller, Christina Koch, and others
- SC21 BoF – feedback from many HPC site operators
- Content moved to the Carpentries Incubator, winter 2021
- Lesson program incubation conversations start with The Carpentries
- Multiple sessions at CarpentryCon 2022
  - Lightning talk on HPC Intro materials
  - Breakout session
  - Sprint on what would become the Workflow lesson
- Formal acceptance into Lesson Program Incubation, June 2023
- SC23 activites, connection with SIGHPC-education community
- Carpentries lesson program incubation goals finalized, fall 2024

HPC CARPENTRY – LESSONS OF HISTORY

The idea of combining Carpentries methods and HPC technical content is not an entirely novel one, and the organization has in fact survived a change of leadership.

It's only in the past couple of years that enough of the right people have had sufficient bandwidth for the effort to gain more traction, to make connections with related efforts (SIGHPC education and Code Refinery, for instance), and to get serious about the lesson-program incubation process.

HPC CARPENTRY – MATERIALS

The HPC Carpentry project is the custodian of a wide variety of materials at varying levels of development, and also links to materials owned and developed by other organizations. A key organizing tool in comprehending this breadth of material is a focus on specific workshop tracks. These tracks are our highest development priority.

The first track is a **User** track, for which we now have a complete set of materials.

The second track is a **Developer** track, for which we lack a parallel development lesson.

HPC CARPENTRY – SHELL INTRO LESSON

Historically, the HPC Carpentry project had an **HPC Shell** lesson, which mirrored the Carpentries **Unix Shell Intro** lesson in many respects. It starts from the same starting point, that of an introduction to the shell. Whereas the Carpentries lesson is purely local, the HPC Shell lesson includes the SSH connection tool, its related environment variables and behavioral quirks.

This lesson did not benefit from the feedback and updates that the more widely-used Carpentries lesson did, so we made the decision to retire **HPC Shell**, and move the SSH material to the beginning of our **HPC Intro** lesson. Since **HPC Intro** is in both our strategic workshops, this does not introduce any duplication of material or effort.

HPC CARPENTRY – HPC INTRO LESSON

The **HPC Intro** lesson is the most mature part of our portfolio. It takes novice HPC users from basic shell commands through to running a parallel job with varying degrees of parallelism on an HPC resource.

We use a "stunt executable", a project-specific program called *amdahl*, which does a particular amount of "work" (actually timed "sleep" commands), some of which is serial and some of which is parallelizable. In this lesson, learners are introduced to both parallelization and the concept of diminishing returns from parallelization.

HPC CARPENTRY – WORKFLOW LESSON

Our most recent addition is a workflow lesson, which uses the *snakemake* workflow tool and the *amdahl* executable from the **HPC Intro** lesson to walk the learner through a scaling study, where they systematically collect data on the run-times of the *amdahl* runs at varying degrees of parallelization, and produce a plot reflecting Amdahl's Law.

A variant of this lesson, developed for a specific workshop at LLNL, uses the *maestro* workflow tool in place of *snakemake*. We found, as hoped, that the lesson content has sufficient focus on concepts that swapping out the tool was a relatively straightforward lesson development task.

HPC CARPENTRY – RELATED MATERIALS

We have a number of other lessons that get varying degrees of attention, but all of which have value.

- The **HPC Chapel** lesson, which demonstrates the use of the Chapel parallel-programming language. This lesson recently got some developer attention.
- The **HPC Parallel Novice** lesson, which implements a parallel Monte Carlo integrator in Python to estimate pi. This material may form the basis for a future HPC developer lesson.
- The **HPC Python** lesson, from which much of our workflow *snakemake* material was taken, and which also will likely inform a future development track.

This list is by no mean comprehensive.

HPC CARPENTRY – STRATEGIC PLAN

Our current development plan is designed to focus our attention on building out two two-day workshops, as mentioned in the foregoing material.

- The **HPC User** workshop, consisting of Unix Shell Intro, HPC Intro, and HPC Workflow lessons, taking learners from the command line to a scaling study. All of this material currently exists, and the workshop has been taught in its entirety several times. Edits and improvements are on-going.
- The **HPC Developer** workshop, consisting of Unix Shell Intro, HPC Intro, and a future HPC Programming lesson, using a parallel framework, such as MPI. The last lesson in this set does not yet exist, but there is good material to draw on.

HPC CARPENTRY – LLNL WORKSHOP

The recent (June 2024) workshop at Lawrence Livermore National Laboratory was a good proving ground for the **HPC User** workshop, and was the first time all the material was taught together in its entirety.

We learned a number of things in real time from user feedback.

- The lesson set is coherent and works
- Having a shared notepad with checkpoints of relevant files is useful
- Pre-arranging HPC access is high-value

HPC CARPENTRY – COMMUNITY CHALLENGES

Probably the most important challenge facing us is identifying learners who will be well-served by our materials and methods. Many computing centers have good training materials of their own; our strengths lie in the "long tail" of novice users, and users of small or transient clusters outside of large centers. We can also serve another "long tail" of users not served by center-specific materials, but it's hard to know where those people are.

Outreach efforts like this are part of the solution!

HPC CARPENTRY – COMMUNITY CHALLENGES

Our history includes many disparate forks of closely related materials. It is obviously desirable to get post-divergence feedback from other people teaching the lesson, to deliver on the promise of continuously self-correcting, pedagogically-sound open-source materials.

Reaching back to fork owners is an on-going effort.

HPC CARPENTRY – DEVELOPMENT CHALLENGES

We are currently a small community of enthusiasts and developers. In the glorious future, we hope to be able to deliver quality workshops more or less on request anywhere in the community.

This is in tension with the hands-on character of the method.

Building up a network of qualified instructors is the key to this. Qualified instructors are "corner cases", people with learner empathy from Carpentries instructor training, and sufficient HPC operational expertise to manage the unexpected in a lesson session.

HPC CARPENTRY – DEVELOPMENT CHALLENGES

Having HPC resources to teach on is an important requirement for our workshops.

Our solutions involve using the "Magic Castle" tool on top of the "Terraform" cloud provisioning infrastructure to stand up a cluster configured for our workshops. This requires workshops to incur a cost, either a monetary one to obtain the relevant cloud resources, or a logistical one to arrange for access to HPC resources, e.g. the US NSF's "Jetstream2" HPC cloud.

HPC CARPENTRY – DEVELOPMENT CHALLENGES

Closely related to HPC resource access is site-specific lesson configuration tools. The legacy Jekyll-based lessons used top-level configuration switches and "snippet" files for site localization, which was already beginning to scale poorly several years ago. A likely path forward is to have a smaller set of higher-level configuration switches that do "good enough" site localization, striking a better balance between cognitive burden on learners (from the mismatch between lesson materials and their actual environment) and logistical burden on workshop hosts.

HPC CARPENTRY – COMMS

Our main website:
https://hpc-carpentry.org
Github project page:
github.com/hpc-carpentry
Slack:
#hpc-carpentry on **carpentries.slack.com**
(Invitation: **slack-invite.carpentries.org**)
Topicbox:
carpentries.topicbox.com/groups/discuss-hpc

hpc-carpentry.org