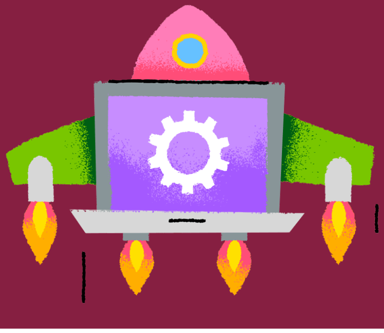
TECHNOLOGY WHITEPAPERS

High Performance Computing

Chris Gahn
Academic Computing Consultant for the Sciences



VASSAR cis

March 2019

What is high performance computing (or HPC)?

In a basic sense, the term "high performance computing" refers to a field of computer science that is concerned with using very powerful computers to accomplish tasks that a less powerful computer couldn't do. You could think of these HPC systems as "servers" (i.e. computers that wait patiently for specific requests and diligently work on those jobs, sending you the output when it is finished). The field of HPC encompasses far more than just the computing hardware. Computational research is becoming an academic field unto itself, and many PhD researchers have left their chosen field of study to work on facilitating the research of others via HPC. The field of HPC includes researchers, hardware vendors, software vendors, open-source developers, systems administrators, institutional administrators, facilitators, and students. It is one of the fastest growing fields, and serves to bridge the old gaps between disciplines and can create a whole new hub of collaboration.

Big questions require big solutions to answer. I'm sure you've heard the term "big data" before. I like to think of "big data" as any data set that cannot be crunched through and analyzed within a few minutes on your laptop or desktop computer. Researchers in every field deal with tremendous amounts of data that come in all shapes and sizes, and the data available seem to grow exponentially as time goes on. HPC allows us to leverage very powerful computers to accomplish big-data tasks that are impossible on even the most powerful computer you can buy from Apple or Dell. Here are just a few examples of how HPC systems are being used today:

- Statistical analysis of linguistics in Shakespeare's plays
- Sequencing the full genome of the Wooly Mammoth
- Atmospheric and Environmental Modeling for the EPA
- Modeling complex economics

Perhaps you've heard of "cloud computing?" Well, HPC operates in the cloud! This is just a fancy way of saying that the task you're performing is being done on a computer that is somewhere else. Since HPC systems always live in data centers, you never actually get to sit down in front of a HPC system! You'll always be connecting from your own computer.

What is the value of high-performance computing?

1. **It's faster.** HPC can drastically speed up workflows for researchers. For example, let's say you have 3 consecutive steps to your procedure and each one takes about 5 hours to run on your desktop computer. That might seem like a reasonable timeframe and not worth the hassle of figuring out how to run it on a system that might take 10 minutes. However, many workflows have multiple steps in the computing process that depend on the previous step. Let's say there was something wrong with your code in step 2, but you didn't find out until you wasted three hours of processing time on your desktop before the program decided to crash. If you had utilized a HPC system, you could begin the debugging process much sooner and waste far less time in the long run. So, it's not just huge computational jobs that can benefit from HPC!

2. **It's cheaper.** Utilizing HPC can reduce costs associated with many projects by batching jobs together, and reducing the need for additional research assistants in some fields. This cost saving has the secondary effect of allowing researchers with less funding the opportunity to compete on a more level playing field with researchers at big universities with more funding.
3. **It's more robust.** Have you ever left your computer running overnight, perhaps downloading a large file, or rendering a 3D design? Have you ever had the computer crash on you or turn off due to a power outage in the middle of an important job? Ever lost important data when a hard drive in your computer died or the file system became corrupt? Utilizing HPC systems can mitigate these risks as well, since almost all of these systems have redundant backup power and storage drives.
4. **It's greener.** HPC systems are almost always more efficient than desktop workstations, when it comes to energy consumption. If you're concerned about your carbon footprint, using HPC will reduce wattage considerably as compared to a desktop or laptop computer. This is assuming, of course, that your job is big enough to use HPC.
5. **It's more convenient.** A great thing about using HPC to do your heavy lifting is that you can be anywhere in the world and work on your data analysis or modeling project. Location is irrelevant when you are connecting to a server. You can even use a computer at a hotel, café, or library to connect! With proper planning, this also eliminates the need to carry around portable hard drives to store your data while traveling.
6. **It's more efficient.** If there are multiple steps to your data workflow, they can be built into a single script called a "pipeline," so that you don't have to monitor the current task and start the next one manually when it is finished. This saves a lot of time and frees you up to think about bigger questions.

How is HPC used at Vassar?

There are many researchers here at Vassar who are already taking advantage of HPC resources. Departments represented include Biology, Chemistry, Cognitive Science, Economics, Mathematics & Statistics, Physics, and Psychology. Project subjects include: molecular modeling for antitumor compounds, computational chemistry, bioinformatics in microbiology, consumer economics, and psychological analysis of historical Reddit posts.

Researchers at Vassar have used several different HPC resources for their projects. Here on campus, we have a computer cluster, currently named "Junior," which will be renamed shortly by the Asprey Center for Collaborative Approaches to Science (ACCAS). For information about Junior or to obtain an account, get in touch with Chris Gahn or Jerry Bailie (chgahn@vassar.edu, jebailie@vassar.edu) or visit http://pages.vassar.edu/accas/high-performance-computing-with-junior/. Other HPC resources used by Vassar researchers include Amazon Web Services (AWS), a cloud computing platform by Amazon, and the NSF-funded "Extreme Science and Engineering Discovery Environment" or XSEDE, which provides access to multiple supercomputing clusters at R1 institutions across the country, for free!

How can I get access to high-performance computing resources?

The most common method for interacting with a HPC resource is via the terminal, or command line. Nearly every HPC system in existence runs some form of Linux, so obtaining a basic familiarity with Linux terminal commands is highly suggested. There are plenty of resources for learning to use the Linux command line on LinkedIn Learning. A good place to start is "Learning Linux Command Line." by Scott Simpson. With a basic knowledge of the command line, whole computational worlds open up to you and your project.

The most common command-line program you will use is called SSH (or Secure Shell), which is a program that allows you to connect remotely to another computer via the command line. This is the gateway into the realm of HPC, so some knowledge of how to use SSH will be necessary. Here is a course to learn all about SSH: https://www.linkedin.com/learning/learning-ssh/what-is-ssh?u=35187788.

After you have a working knowledge of basic Linux commands and SSH, the next step is to identify what software or programming language will be most useful to you in your project. For many researchers, there are field-specific software programs or languages that are widely used and popular. These include programs such as C/C++, Discovery Studio, Java, Javascript, JMP, MATLAB, Mathematica, Python, QIIME, R, SAS, SPSS, etc. If that list seems daunting, and you are unsure of your specific needs, you should try reaching out to colleagues in your field, and looking through methods in the literature to find out what your peers are using for their work. Once you've zeroed in on a particular programming language or software package, LinkedIn Learning may have a detailed course that can help you on your way.

After choosing an approach (a programming language or specific software package), it's time to start conceptualizing your project. Think of this like writing down a cooking recipe. List all of the ingredients (inputs) for your project, the final product (outputs), and then begin to fill in the steps to get from your inputs to your desired outputs. No programming experience is necessary for this part of the planning.

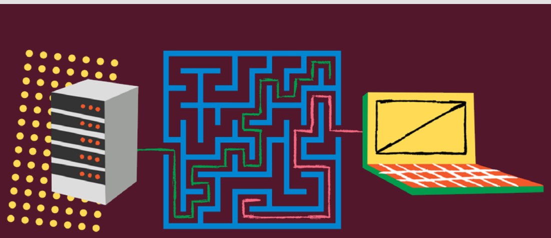
Once you have come up with your recipe of steps needed to complete your project, it's time to translate that recipe into some code that the computer can understand. This part of the process can be time consuming and frustrating, but when properly designed, will result in all of the benefits listed above! There are plenty of resources available to help you through this part of the project, including Chris Gahn (chgahn@vassar.edu), LinkedIn Learning (https://linkedin.com/learning), your colleagues, and a multitude of other resources on the web that are only a search term away!

What are the downsides?

1. **Learning Curve.** As with any method, there are downsides to HPC. One of the biggest is the steep learning curve which can serve as a substantial barrier to entry. For many people, computer programming and the command line terminal can be very daunting. We live in a digital world in which the GUI (graphical user interface) reigns supreme. When you take away the windows and icons, it can be very difficult to visualize what the computer is doing behind the scenes.
2. **Access.** As in many fields, inequities can exist in HPC which may affect accessibility for some people. Great strides have been made in leveling the playing field in HPC, and the availability of training materials and tutorials on the internet has made knowledge in this area much more widespread. The publicly funded XSEDE project can reduce or completely eliminate costs to access HPC resources, but certainly, inequity exists.
3. **Security.** Data security can be a concern when working on remote systems. If the nature of your research demands that your data be protected, extra precautions will need to be implemented to ensure safety from a cybersecurity perspective. Any time you upload data to a remote server that you don't own and operate, there are risks involved. You can add security measures to mitigate or eliminate the risks, but this requires time and planning, which can slow down the process.

How do I get started?

Set up a consultation with Chris Gahn (chgahn@vassar.edu), and take a look on LinkedIn Learning for courses in the software and programming languages mentioned above!



Illustrations by Ouch