HOWTO Do an Intake Interview

ACI-REF VR Workshop 2019 (intro. level)
David Chin (Drexel University)
dwc62@drexel.edu

Why do an intake interview?

- Get to know your researchers and their work
  - That's (possibly) why you're in academic/research HPC. For me, it is interesting to know what kinds of research are going on.
  - Blurb for your website e.g. https://drexel.edu/research/urcf/research/groups/
- Gauge fit: how well can researcher's work be served by your system?
  - Explain clearly the use cases your system is suited for; also use cases your system is NOT suited for (e.g. processing backend for a gene sequencer machine)
- Gauge start-up time: how soon can researcher get going?
  - Make a plan for getting the group up and running
- Gauge throughput: how much work can researcher expect to do per unit time?
  - NB this will necessarily be fuzzy until they actually get going, BUT they may have an idea if they have done work on other HPC systems, or various PCs/workstations

What do you need to know?

- Do they have requirements that your cluster cannot meet?
  - Specific hardware: FPGAs? GPUs? Intel MIC?
  - RAM requirements? (Bio and machine learning workflows may need multiple TB RAM)
  - Data at rest (i.e. amount of storage)? Public datasets? Protected datasets?
  - Data in motion/transit? (e.g. field sensor data like UCSD WIFIRE project on wildfires)
  - I/O-bound workload? To shared or local storage? Parallel filesystem?
  - Refer to other resources: XSEDE, etc -- see Thu 3:30pm session
- Do you already have what they need?
  - Other researchers relying on same requirements, already installed
- If not, how hard is it to get?
  - Software: usually doable
  - Hardware: more complicated

What does researcher want to know?

- How fast can they get started?
- How do they get their students/postdocs/staff accounts?
  - Can they have outside collaborators?
- How fast can their computation run?
- How much storage do they get?
- Is the data backed up?
- How much will it cost?
- Grant proposal support - loop in research support office, library
  - Facility description boilerplate
  - Cost estimates
  - Data management plans

Getting to know you

- Ask about their project(s)
  - Probably won't understand much if any of what is described due to specialized terminology
- Are they a new hire, or are they bringing an established group along?
- How did they hear about you?
  - Info to tailor your future outreach efforts
- They may jump ahead and talk about specific software versions and other requirements
  - That's OK as long as you cover all you need to cover
  - If they have very specific hardware requirements, figure out if they are requirements or just something they are accustomed to that may not be hard requirements

Hardware and infrastructure

- You may have to help them figure out these requirements - make plans
- How have they done, and how do they currently do their research computing?
  - PC or workstation?
  - HPC cluster elsewhere? XSEDE? Commercial cloud services?
- Co-processors
  - GPUs? TPUs? FPGAs? Intel MIC (Xeon Phi)? Quantum computers?
- RAM
  - Bio workflows may need multiple TB RAM; machine learning may have large RAM reqs.
- Storage
  - Amount of space? Parallel filesystem (for multinode parallel i/o-bound work)? Local scratch?
- Network fabric
  - InfiniBand? Other?

Hardware investment

- If you have a "condo" model where PIs may buy hardware to expand cluster
- Does researcher have funds (startup? grant?) to invest in cluster?
  - Complex topic and very specific to your institution

OS and Software

- What OS have they used for their research computing?
- What software?
  - Do you already have it? Point to your web page listing software available.
  - What version? Research software can change drastically between versions, whether in functionality or build process.
  - If not
    - How do you get it? Commercial (talk to your IT software group, other academic units re license), binary distribution (e.g. Anaconda), source distribution.
    - Can they build it? Probably not.
    - Can you build it? Probably, but this may take more time than you think.
- Software/hardware combo
  - E.g. GPU-intensive 3D analysis and/or modeling

Data: public and protected

- There are many open datasets available in various fields: bioinformatics, economics, ecology, finance, etc.
  - Do they need any?
    - Shared, read-only access to avoid storage charge and data duplication
- Do they have protected health information (PHI), or other restricted data?
  - Are you set up to handle it?
  - Talk to your research support office and IT compliance staff for details: you do not want to get this wrong

Make a plan

- Assures researcher that there is a path forward
- Accounts for students/postdocs/staff
- Basic training
- Software installations
- Benchmarking
- Talk to appropriate 3rd parties: research office, central IT, library

Thank you!

Practicum breakout