UNIVERSITY OF CAMBRIDGE

Introduction to High Performance Computing (HPC)

Graduate practical class

November 2016

Jenny Barna

Aims

- Very brief overview
- Awareness of resources
- How to find out more
- Possible use for projects

What we will look at today

- What is HPC?
- What aspects are useful?
- When would we use it?
- What HPC can we use in Cambridge?
- Using a shared HPC facility
- A few exercises

What is HPC?

- HPC involves some sort of ‘big computer’ and may use
  - one or a few large computers (servers)
  - a cluster of many computers
  - a large amount of memory (RAM)
  - large/fast storage (disk space)
  - GPUs (Graphics Processing Units) as well as CPUs (Central Processing Units)

One HPC server...and hundreds

One server or a cluster?

- Depends on software type and data size
- Different ways of splitting up a problem over several processors
- Using just one server, sharing memory (eg OpenMP)
- Using a cluster, communicating between servers (eg MPI)
- Using independent servers (or cluster) for high throughput
- Need to understand what the software/data require
- Depending on project but likely to have a mixture

Large memory (RAM)

- Large shared memory is required for some analyses
- Needs to be all in the same computer (server)
- RAM is still relatively expensive
- Tend to have a few large RAM servers for specific needs
- In biology an example use is some genome assembly software
- Usually 256GB-512GB RAM; occasionally more than 1TB
- Servers can also perform smaller tasks (ie several at a time)

Large/fast storage (disk space)

- Much research now involves ‘Big Data’
- Disk space has become a common limiting factor
- Also need fast access to the data
- Space may be shared across all the computers in use (cluster)
- Various high performance file systems eg Lustre
- Developments eg solid state disks
- ‘Omics, imaging, modelling…..
- Used in projects involving many GB or 1TB+ data

CPU or GPU?

- All modern computers have multi-core CPUs
- Speed-up using all CPU cores in one server is limited
- Using large CPU cluster may still be slow
- Some software re-written (partly) for GPUs
- If successful may obtain large speed-up
- Examples include image processing
- Most projects use CPUs but GPUs are available if required

What HPC can we use?

- Some groups have HPC servers or small clusters
- University has an HPC service (HPCS)
- Includes a CPU cluster: Darwin
- GPU cluster for suitable software: Wilkes
- Linux is the operating system
- Darwin servers (nodes) have 16 cores and 64GB RAM
- Some nodes have more cores and RAM

BioCloud servers

- New servers within the HPCS
- ‘Normal’ have 24 cores and 256GB RAM
- ‘Himem’ have 64 cores and 1TB RAM
- Available for biologists
- Extra login nodes
- More readily available resources
- Will use in exercises today

Using a shared HPC facility

- Shared facilities usually have a scheduler
- It works with the operating system (Linux)
- The scheduler shares out resources
- There may be different priorities and needs
- It can be set up to fit lots of work in efficiently
- One needs to learn a few ‘words’ that are not Linux
- Tasks are submitted to the scheduler

Submitting tasks on an HPC facility

- Users have their own logins
- Data are transferred to the facility's storage
- Software is installed if required
- A script is prepared and submitted to the scheduler
- In the exercises some of the steps have already been done
- The Cambridge HPCS scheduler is called SLURM:

  Simple Linux Utility for Resource Management

Exercises

- Notes are online at the URL below
- Work at your own pace
- http://people.ds.cam.ac.uk/jcjb/ucamonly/gradhpc
- Make sure you have login details first
- Links to more information are in the notes
- A more detailed HPCS course/handout is available