02 Intro to HPC

YouTube: https://youtu.be/o6Cp59KlM7c

Hi, and welcome back to ARCC's training video series.

In this video we will introduce some more basic concepts on High Performance Computing (HPC), some high level architectural concepts, and related terms to understand.

The video does assume that you have already watched the 'Intro to ARCC Services' video.

High Performance Computing generally refers to the **practice of aggregating computing power** and resources in a way that delivers much higher performance than you would typically get from a desktop computer.

One thing to take note of and remember as you work through these series of videos is that:

HPC ≠ Desktop

There are some similarities, but there is a process of migrating something that maybe runs on a desktop/laptop to work and utilize the benefits that a HPC system provides.

This aggregated computing power is provided by what we call a **cluster**.

In its simplest form, a cluster is a collection of compute nodes (computers) connected together via a fast interconnect (type of network), all with access to a shared parallel file system (storage). As a user you access the cluster via a login node.

Future videos will demonstrate how to access and use this login node.

You can consider the basic components of each compute node in a similar way to your desktop. They have a number of processors (or cores), they have memory, some might have a GPU (Graphical Processing Unit), they have local storage (hard drive), as well as access to shared/external storage.

The main difference is the amount of memory, and the number of CPUs, and the number of GPUs.

Across our clusters we have compute nodes that vary across these components. Our low range compute nodes have a minimum of 64G of memory and 16 cores. Our standard nodes typically have 128G and 32 cores. Then we have more specialized nodes with memory increasing to 512G and even a couple with up to 4T.

We also have a selection of nodes that come with a variety of NVidia GPUs.
With this potential variety of compute nodes across a cluster there are generally two type of HPC systems:

There are **Homogenous** systems where all nodes share exactly the same architecture. i.e. each node has the same number of CPUs and type, the same amount of memory.

Heterogenous systems on the other hand have nodes that can vary architecturally. What we do on these systems is to group similar compute nodes into named **partitions**, which if a user specifically requests then they know the type of compute nodes they are getting.

Teton, the name of one of our core clusters, which has over 600 connected compute nodes, is a heterogeneous system.

The final term that we will introduce is the idea of the "**condominium model**".

ARCC as a service is available to all researchers, and as such all researchers can use our core cluster called "teton".

To facilitate this, we use the "**condo model**". This allows researchers to **invest into the cluster**, purchasing additional compute nodes that they get **priority to use**, but when not in use these compute nodes are available to the rest of the community. If an investor wants to use their compute nodes, but someone outside of their project is currently using them, we '**preempt**' this existing job (basically stop and reschedule it) and allow the investor to start their jobs immediately.

This is managed by defining **investor partitions** that specifically define the compute nodes purchased by that researcher.

In future videos we'll detail how to:

- Use the login nodes to access the cluster.

- What your working environment looks like and how it is organized.
- Set up your environment for your specific needs, using LMOD which is our module system.
- How to interactively work, and/or schedule jobs on the clusters using Slurm our Workload Manager.
- Request specific partitions or capabilities on compute nodes.
- As well as talking in more depth about the types of serial and parallel jobs that can be run across the cluster.

To summarize, in this video we have introduced:

- A high-level architectural overview of what a HPC cluster looks like.
- How clusters can be divided into partitions that can be defined by compute node type and/or investor purchased nodes.
- And the idea of the “condo model” that allows a more open and usable cluster.

Check out our wiki and search for terms such as ‘teton’ to find out more about our clusters and how to access and start using them.