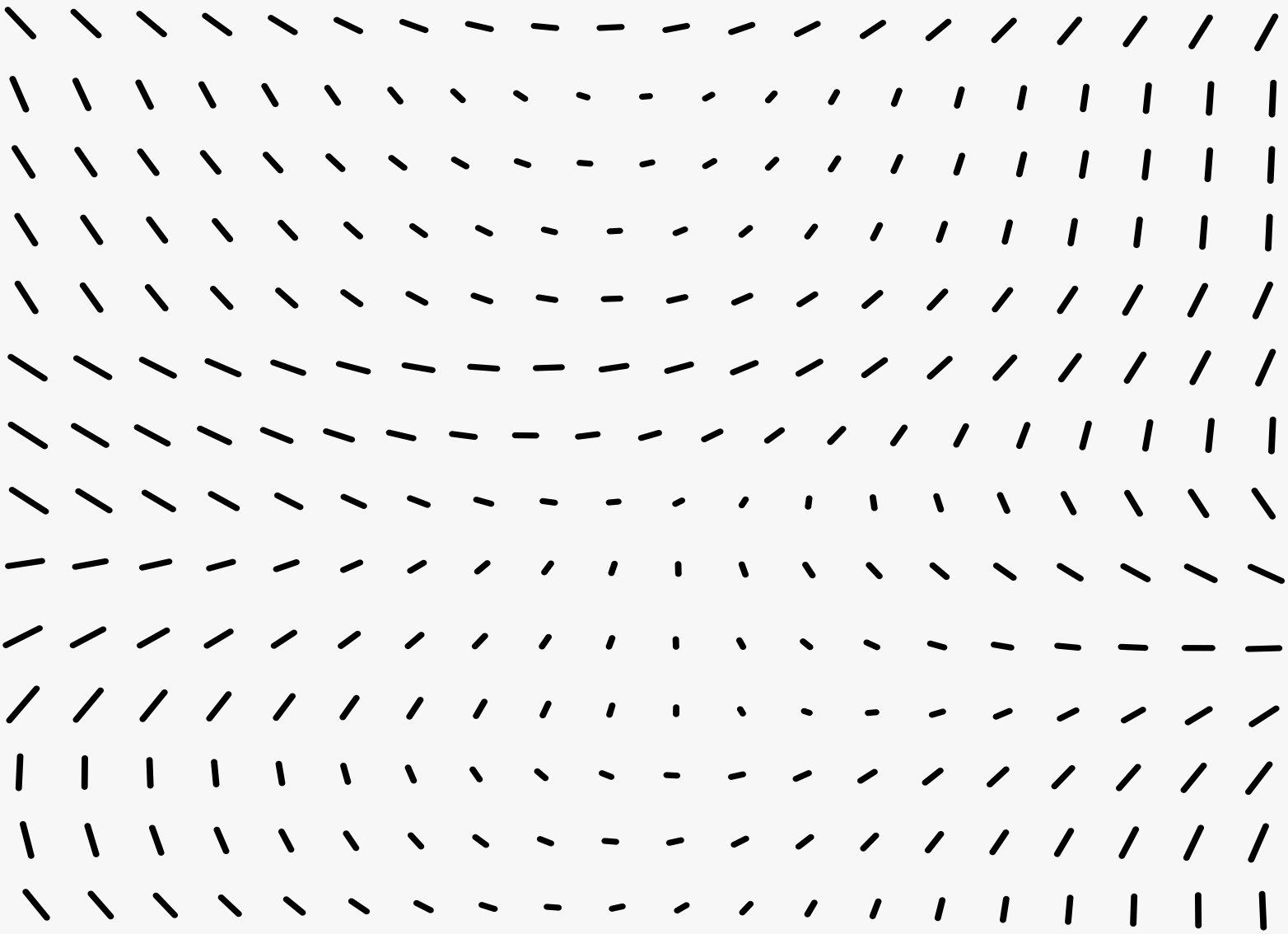
Practical!

Introduction to High Performance Computing (HPC) and the Triton Shared Computer Cluster (TSCC)

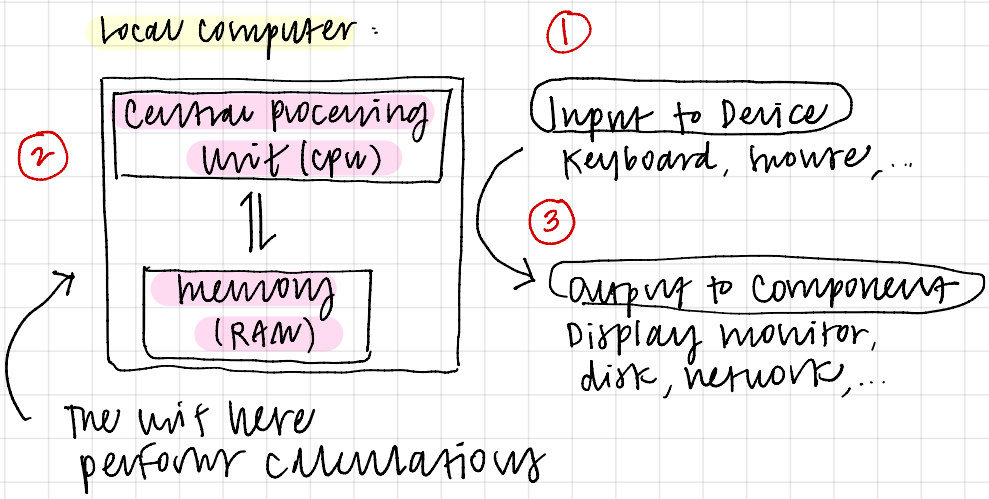
notes by Michelle Ragsac



Introduction to HPC (High Performance Computing)

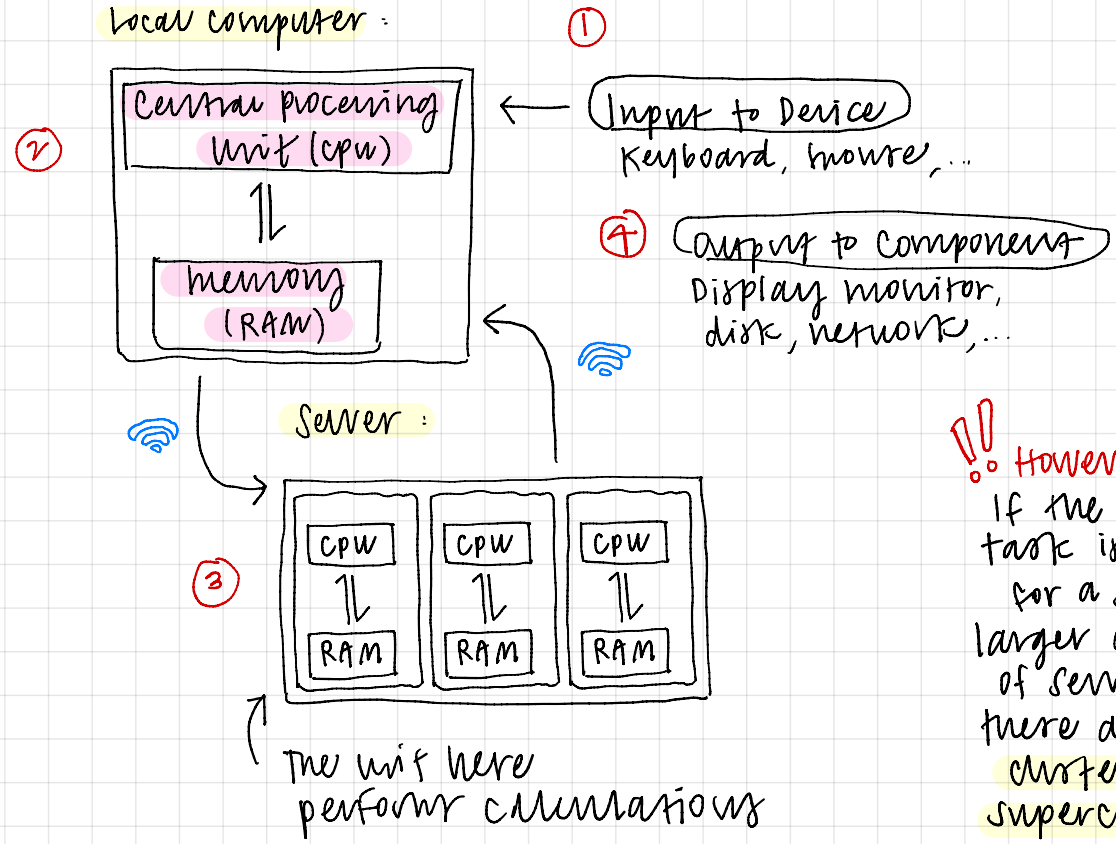
Motivations for using clusters

- There are research problems that outgrow the capabilities of a desktop or laptop computer so in these cases, access to more computers is needed



!! However ...
When the local computations become too intensive, the operations are outsourced from your local computer → elsewhere (e.g. server)

Servers are computers that reside in a data center — these centers don't need to be physically close to your local computer
→ They contain more memory, storage, and compute capacity than your local system
→ To access these servers, you use your local device to interact with this remote machine



!! However ...
If the computational task is too much for a single server, larger agglomerations of servers are used — these are known as clusters OR supercomputers

Introduction to HPC (High Performance Computing)

What is an HPC system?

Some keywords that are important for us to define early

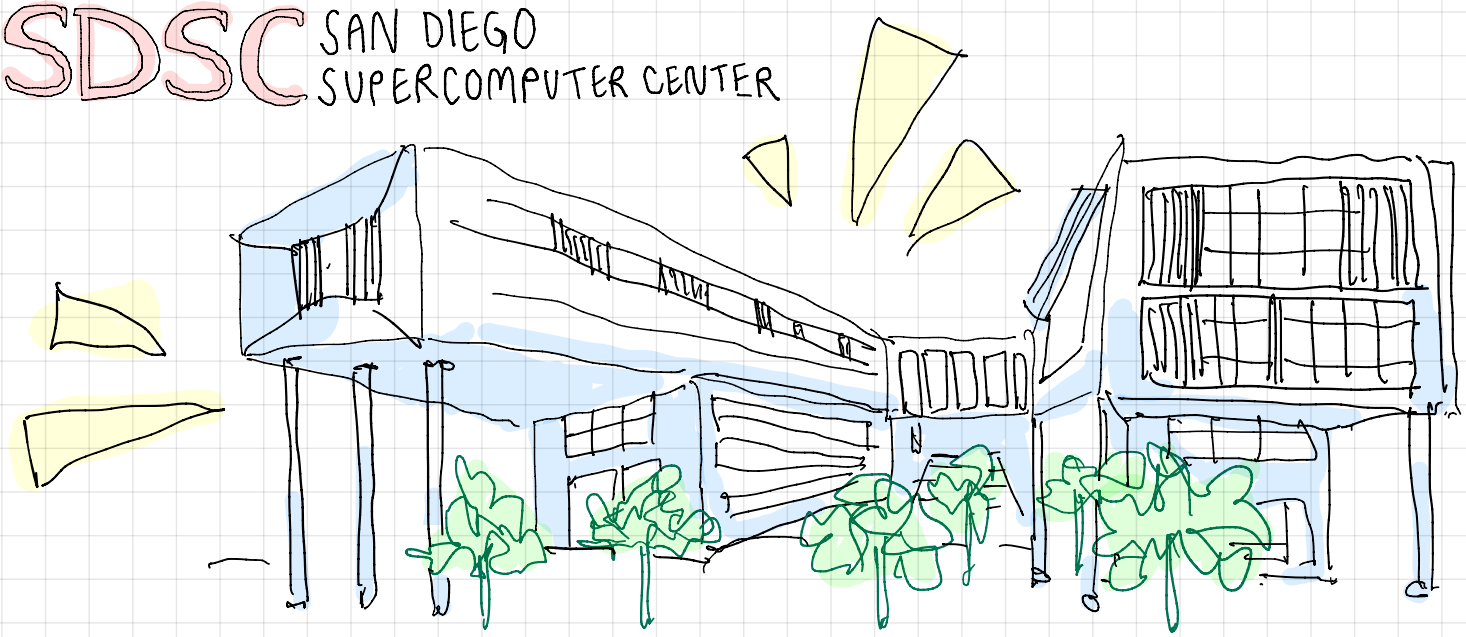
cloud — generic term to refer to computing resources as needed (or on demand) and represents real or virtual resources that may be physically located anywhere

HPC system — stand-alone resource for computationally intensive workloads — typically exist in a single, fixed location to accomodate for multiple computers connected together with networking cables

cluster — smaller-scale compute/HPC systems that are maintained in computing centers

- UCSD has its own computing center on-campus that is used by local researchers, biotech companies, as well as other groups across the country!

SDSC SAN DIEGO SUPERCOMPUTER CENTER



SDSC is based on the UCSD main campus — north campus next to RIMAC! It houses several clusters, including the ... NSF-funded COMET resource and the "crowd-funded" TSCC resource

Since there's a good amount of labs that use TSCC, and the skills that apply to one HPC resource can apply to another, we'll be doing some exercises using that system!

SIDE NOTE

researchers purchase and contribute computers to add to the TSCC resource, so the more people in the TSCC crew, the more powerful TSCC gets!

Introduction to HPC (High Performance Computing)

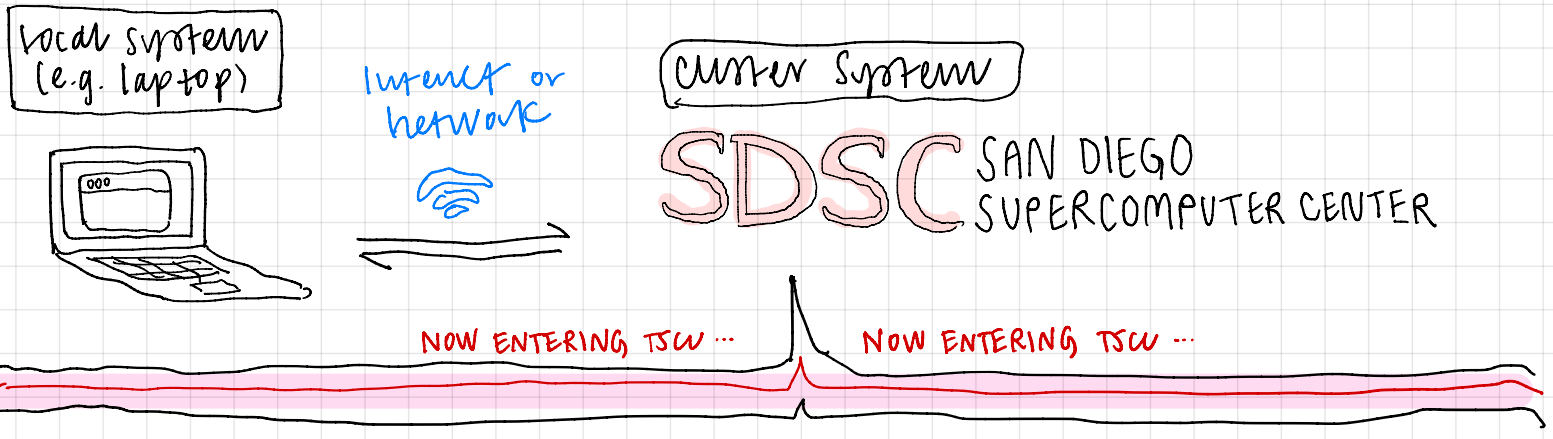
Interacting with HPC systems

NOTE: This portion of the notes will have more conceptual info instead of commands/code!

① Logging into the cluster (TSCC)

(local system)

The first step in using a cluster is to establish a connection from our laptop to the cluster using a secure shell client (SSH)



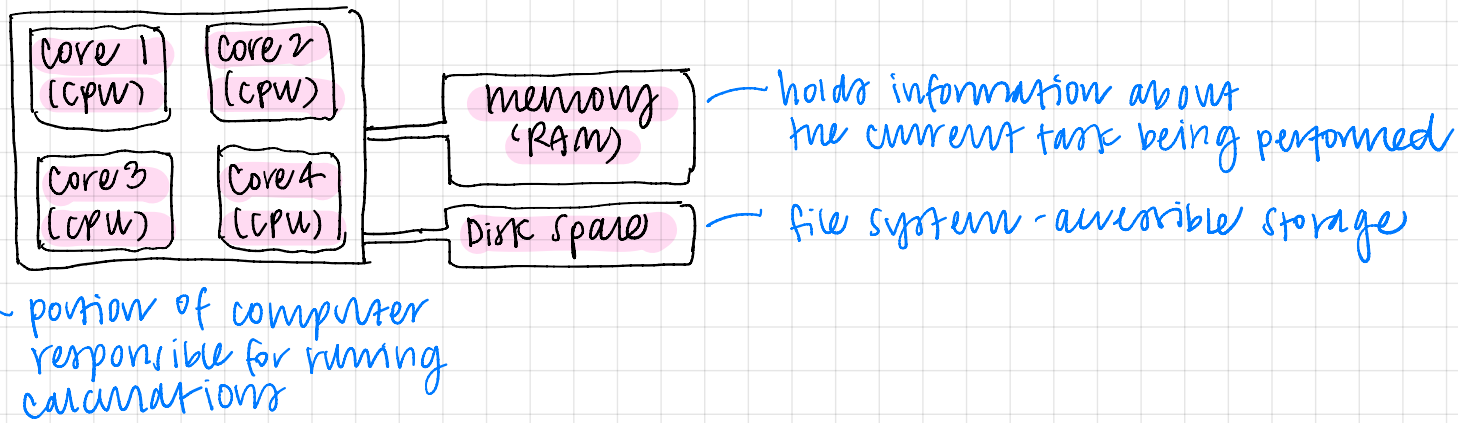
The organization of TSCC

When you first log into TSCC, you enter a space known as the login node — you can think of it as a computer with minimal amounts of compute power and memory

→ The purpose of the login node is to provide you with enough resources to access other portions of the cluster

BUT what are NODES? : Individual computers that compose a cluster are called nodes, and there are different types of nodes depending on the task

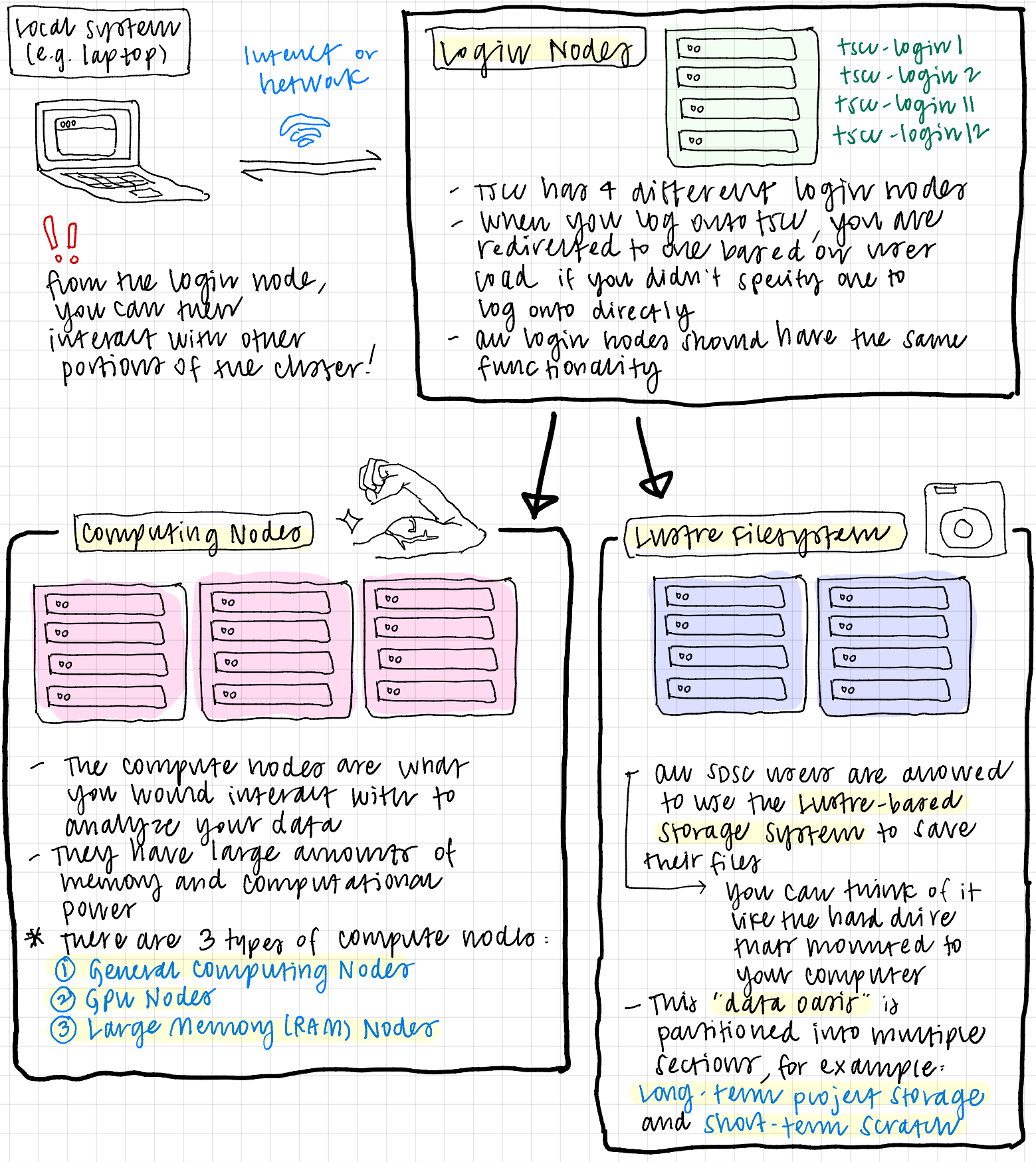
Node diagram:



Introduction to HPC (High Performance Computing)

The Organization of TSCC continued...

organizational diagram of TSCC:

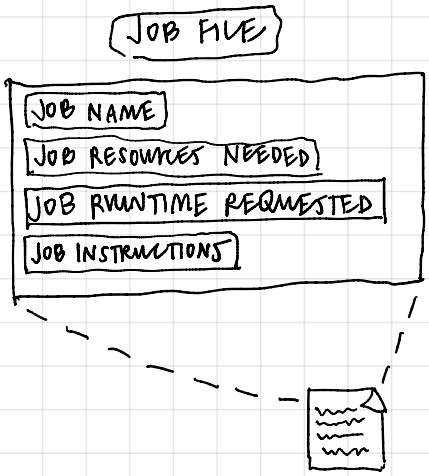


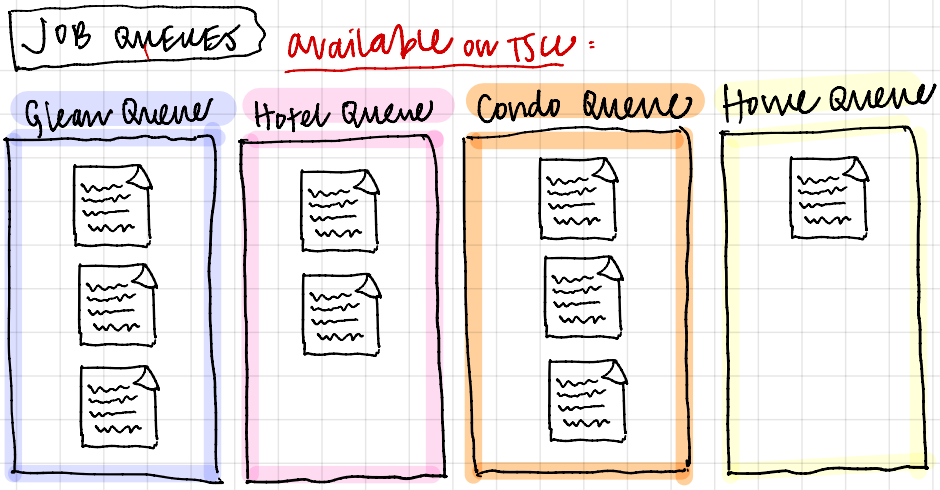
Introduction to HPC (High Performance Computing)

Job scheduling

- HPC systems have multiple nodes to share amongst many users —
  how do we decide who gets what resource and when?
  how do we ensure that tasks are run with the resources it needs?
- To answer these two questions, HPC systems use a special piece of software known as a **job scheduler**
- The job scheduler manages which compute jobs (e.g. sequence alignment, or RNA-seq quantification) run where and when

Job scheduler diagram:





① Within a job file, you specify the resources and time you need to complete your task

② After writing your job submission, you submit it to a job queue (a line) to be run!

Glean queue (Free samples): all jobs run on the glean queue are run at essentially NO cost

Hotel queue (pay as you go): Reserved for users that have compute resources, but do not have an account affiliated with a purchased TSCC node

Condo queue (Timeshare): Reserved for user that have an account affiliated with a purchased TSCC node and want to used the shared cluster resources

Home queue (Personal villa): Queue for users who want to use just the computational resources they've purchased