**Course: Advance Bio Informatics**

**Module Title: High Performance Computing**

**Module No: 138**

High performance computing can be deployed using two major mechanisms, i.e. supercomputers and clusters. This section provides basic details and differences of both of these paradigms.

**Supercomputers**

They are constructed by a major vendor (IBM, HP, ...). They Use custom components (processor, network). They are based on custom (Unix-like) operating systems. They are expensive to buy. They have High scalability/availability. They enjoy idea of pooled resources: shared with many users (remotely accessed)

**Virtualization:** high utilization of hardware resources (no idling)

**Elasticity:** dynamic scaling without capital expenditure and time delay

**Automation:** build, deploy, configure, provision, and move without manual intervention

**Metered billing:** pay-as-you-go, only for what you use

**Clusters**

Clusters are assembled by vendor or users. Commodity-of-the-shelf components (COTS). It uses Linux operating system. Factor 10 cheaper. It has Low availability/ scalability

In a cluster, each machine is largely independent of the others in terms of memory, disk, etc. They are interconnected using some variation on normal networking. The cluster exists mostly in the mind of the programmer and how s/he chooses to distribute the work.

In a Massively Parallel Processor (supercomputer), there really is only *one* machine with thousands of CPUs tightly interconnected. MPPs have exotic memory architectures to allow extremely high speed exchange of intermediate results with neighboring processors.

The major variants are SIMD (Single Instruction, Multiple Data) and MIMD (Multiple Instruction, Multiple Data). In a SIMD system, every processor is executing the same instruction at the same time, only on different bits of memory. Essentially, there is only one Program Counter. In a MIMD machine, each CPU has its own PC.

MPPs can be a bitch to program and are of use only on algorithms that are *embarrassingly parallel*(that's actually what they call it). However, if you have such a problem, then an MPP can be shockingly fast. They are also incredibly expensive.

A cluster is a bunch of machines, normally usually Ethernet interconnect (read: network), each running it's own and separate copy of an OS which happen to serve a single purpose.

An MPP supercomputer usually implies a faster propitiatory very fast interconnect (e.g. SGI NUMALink) that supports either Distributed Shared Memory (run processes on different MPP nodes that use shared memory over the fast interconnect to share data as if they were running on a single computer) or even a Single System Image (a single instance of an operating system, mostly Linux, running on all the nodes at the same time as if on a single machine - e.g. "ps aux" on any node will show you all the processes running on the MPP).

As you can see the definition is quite fluid, it's more a question of scale rather than clear cut differences.

Traditionally Supercomputers have only been built by a selected number of vendors: a company or organization that required the performance of such a machine had to have a huge budget available for its Supercomputer. Lots of universities could not afford the costs of a Supercomputer by themselves; therefore other alternatives were being researched by them. The concept of a cluster was born when people first tried to spread different jobs over more computers and then gather back the data those jobs produced. With cheaper and more common hardware available to everybody, results similar to real Supercomputers were only to be dreamed of during the first years, but as the PC platform developed further, the performance gap between a Supercomputer and a cluster of multiple personal computers became smaller.

**Grid Computing**

It is a new type of Supercomputing.

**Idea:** power of grid.

**Concept:** cluster, Parallel computers.

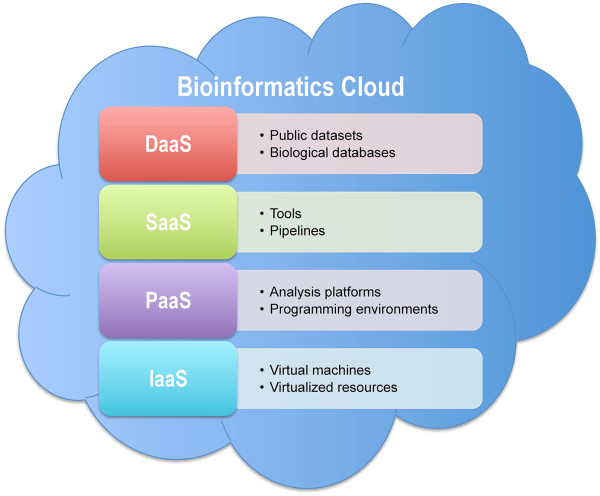
**Problem:** programming, management, security.

**HPC in Bio Informatics**

* 1. Drug Discovery
  2. Gene finding
  3. Multiple alignment
  4. Homology Search
  5. Protein finding
  6. Genomic Sequence analysis
  7. Phylogenetic Inference

**Cloud Based Resource in Bioinformatics**

* + DaaS (Data As A service)
  + Saas (Software As A Service)
  + PaaS (Platform As A Service)
  + Iaas (Infrastructure As A Service)



* + Molecular Modeling for drug design.
  + Protein Folding Design and Docking.
  + Molecular Sequence Analysis.
  + Biomedical Image Simulation.
  + Analysis Of Spectra Data
  + Micro-Array Management

**Future Orientation**

Software/tools must be available for public use. Data compression for data transfer problems. It helps in large amount of data to be accessible over cloud.