Physical architecture of a high-performance computing cluster

- All such clusters contain multiples compute nodes and at least one special node call head node, the head node role is “to rule them all”. There should be at least one head node, but in practice the situations when there are two head nodes is common

- The compute nodes are typically more powerful than the head nodes are. And they usually have a more powerful CPU or bigger RAM. This is happening because the compute nodes are meant to perform the heavy computational tasks, while the head nodes' role is rather a logistic one, an organizational one.
- TODO –Show the difference in resources between the head nodes and compute nodes on the big cluster

- In general all the compute nodes are identical, they have a homogenous architecture both from the hardware and software point of views. This similarity in hardware and software is pretty important since a task is usually run on more than one node in a cluster like environment. Each of the compute nodes should be able to run parts of the tasks and at the same time different supported files such as libraries should be available on each of the compute nodes.

- There are situations when the compute nodes have a slightly different hardware configuration. Example: The UBB cluster: some nodes have an additional Intel Phi coprocessor, while some other nodes are equipped with two NVidia Tesla K40 GPUs.
- New cluster’s extension: new nodes are physically located in a different building, but the most important aspect is the fact that the new nodes are of a different type, different machines, from different manufactures
- Even though the new nodes are of a different type, there is a certain hardware compatibility between the new nodes and the old one, all of them will be 64bit Intel compatible – and the software homogeneity will be easier to achieve anyway.

- A cluster’s compute nodes can be ordinary PCs such as desktop PCs or their form can be manufactured depended in a proprietary enclosure.
- TODO: picture of our first cluster ever built at UBB (reminder: some words about it)
- A modern, commercial cluster has its nodes mounted in a manufacturer proprietary enclosure. For examples a M5 machine in the big cluster looks like a drawer in a closet, in case of problem it can be easily pulled-out outside its enclosure and replaced or debugged.
- Important: Everything is modular!

Head nodes

- The head nodes are not required to have an identical configuration with the compute nodes, in UBB cluster there are two rackable dedicated machines in the IBM Flex enclosure. Typically, a head node is used to manage the whole cluster, to deploy operating system images to the compute nodes (this operation is called imaging the compute nodes), add or remove compute nodes, add or remove users, jobs management, queues management (will see that jobs are usually put and managed in a queue). At the same time, a head node can provide some specialized services such as routing or DHCPD services (that’s a sever that provides IP address for devices within a local network).

Compute nodes

- Compute nodes on the other hand are responsible for all the clusters heavy workload, their main responsibility being to run the jobs submitted via the head nodes. There are certain situation and scenarios when a compute node can directly receive highly intensive computational tasks directly, not submitted via a head node (we will discuss this scenarios in a future class).

- One notable difference between the headnode and the compute nodes is that usually compute nodes have no monitor, keyboard, mouse or any other special input device attached to them (except network). They do have usually some input connectors such as USB connectors or some proprietary input/output ports in the front of their case to support a facile debugging. On contrary, headnodes typically always have a keyboard, a mouse and a monitor attached to them, even though the cluster administration is not performed from the headnode console. Its administration is performed from the headnode but usually over the network.

- Considering that in a High-Performance Computing Cluster and in a Cloud Architecture there can be a lot of virtual machines with different responsibilities that offer different support services to the cluster functionality as a hole, it is worth mentioning that never the head nodes or the compute nodes are virtualized. In order to squeeze the most CPU power of these machines, the operating system on the compute nodes must be run on the physical hardware.
- (Exception: the cluster you'll have to install in a virtual environment for the first homework)

Mistakes in using a cluster’s nodes

- A common mistake among a cluster’s users is to run directly highly intensive computational tasks directly on the compute nodes. Such tasks should be submitted via a special software component of the High Performance Computing cluster called job scheduler. The job scheduler is responsible for finding the best compute nodes available for running a job (a job may consists of more than one task), deciding not only to where a job should be run, but also when and by who (I mean for which user). This is to properly sharing the cluster resources and for assuring fairness amount its users.

- Another common mistake among a cluster's users is to run very CPU demanding tasks directly to the headnodes (the tasks designated to run on the compute nodes sometimes are also run on the headnodes). This is a mistake considering that the head node is not so powerful as a compute node is.

- Exception: benchmarking your virtualized cluster. You'll also have to run some jobs on the head node itself – this is because you will not have so many cores available, you will only have the head node together with two or three compute nodes at most, each of them consuming one physical core of your physical computer (this in the most favorable situation in which you have a quad core processor).

Network infrastructure

- Each cluster has a network infrastructure to connect all nodes together. The simplest network that connects a headnode with its compute nodes is an Ethernet network, usually running at gigabyte speeds. The faster the network is, the better the communication process is between nodes, considering that head nodes and compute nodes typically exchange lot of data among them when running a job. The network should provide low latency, in fact the network itself and the inter node communication mechanism are considered sometimes the bottleneck of a computational process: this is because the timings needed to exchange some data in a shared memory context (that means between processes running on the same node or on the same computer) are far, far lower than the timings required to exchange the same data over the network between processes running on two different compute nodes.

- High performance commercial clusters have usually a dedicated fiber optic network. The UBB cluster has an Infiniband switch capable of speeds up to 56 gigabytes per second, with 1 to 1 subscription rate. The new nodes added to the cluster, although physically located in a different building (a new data center is hosting these compute nodes) are connected to the old cluster over optic fiber capable of speeds up to 40 Gigabytes per second. The UBB cluster has both an Ethernet network and a fiber optic network installed.

Additionally to the Ethernet network and to the fiber optic specialized network, there might be an additionally “support” network, usually running over Ethernet to. This support network connects the administration or administrative interfaces of the compute nodes, these administration interfaces are sometimes called embedded server management interfaces. Each of the big server providers in the industry (HP, IBM or Dell) has some sort of proprietary embedded server management technology:

- IBM calls it IMM from Integrated Management Module
- DELL calls it IDRAC, it is an abbreviation from Integrated Dell Remote Access Controller.
- HP calls it iLO, from integrated Lights-Out

- These interfaces look like a normal Ethernet interface, normally there is a dedicated Ethernet port for this interface, but they can also be configured to share the physical Ethernet data port (the port used for the Internet connection can also be used for the embedded management interface – for example in a situation when there are no ports available in the Ethernet switch). This embedded management interfaces are like a network capable BIOS, they allow accessing the machine even when the machine is turn off, allows also accessing its console (physical console) without the operating system installed or booted on the machine, or they can trigger a push for the power or reset buttons of the machines. These interface have usually a web based user interface that allows accessing the machine even when it is turn off, they also provide some sort of health reports, hardware status and so on. The single main condition to access this remote management interface is that the machine has to be connected to a power source or to an uninterruptible power supply. Such a management interface is like a mini independent computer with its own operating system (it's some sort of firmware) and it is always on if the machine is connected to a power source.

Internet connectivity

- In a cluster's network there is usually a router device that connects the cluster's network to the rest of the Internet. This router provides routing services, DHCPD services (IP addresses for the devices in the cluster's network – although compute nodes usually have a fixed dedicated IP, they usually need a dynamically allocated IP in the initial imaging process – when the operating system is installed on the compute node over the network). A router can also act like a firewall protecting the cluster or as a VPN server, being the endpoint of the VPN connections made by the clusters users.
- Example, the UBB cluster can only be access via VPN by a regular user or it can be access from on an enterprise allocated IP (I mean fixed IP), the cluster cannot be accessed by a home user from a dynamically allocated IP address.

- Sometimes the role of the router is taken by the head node itself. Example: your virtualized cluster running Rocks. In such a situation all the compute nodes will access the Internet via the head node, which in addition to the standard head node specific roles, will also provide routing services, firewall and VPN services, DHCPD services and will even act as a DNS server.
- In this situation, the head node of your cluster has to have two network interfaces, one network interface will provide connectivity to the Internet, and the other network interface will be connected to cluster network, where all the other compute nodes are also connected.

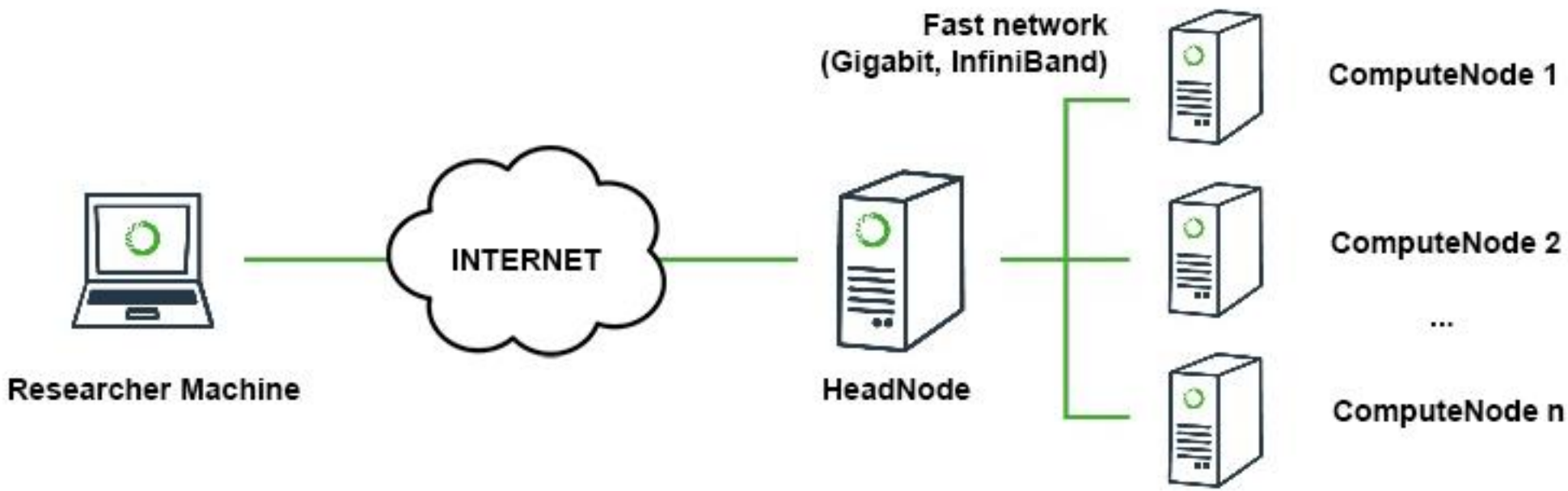
Another interesting fact about a cluster's router that is worth mentioning: it can be virtualized.

- Example1: UBB cluster's router is in fact a virtual machine that is running on the cloud component of the cluster.
- Example 2: The router of Computer Science Department's network, the network where our labs are in the FSEGA building, this router is also a virtualized one.

Storage

- Another important component of a cluster is the storage. Normally the storage goes in a specialized equipment called NAS (Network-attached storage) which is usually attached to the cluster via a high speed fiber optic network. The storage includes in general hot swap capable hard drives called spares (if a hard drive fails it can be changed without shutting down the whole system), the NAS also providing different levels of redundancy. (Reminder: Hotspares)
- In the UBB cluster the storage is shared between the high-performance component of the cluster and the cloud part. From the perspective of the high-performance computing component, the storage has to be shared among all nodes whatever they are: head nodes or compute nodes. And the storage must look the same on all the nodes: if a file is present or exists on the headnode, the same file must also be available to all the compute nodes. A cluster needs some sort of distributed file system, for example the UBB cluster uses a file system called GPFS (General Parallel File System, that’s an IBM proprietary file system), but other simpler file systems can be used such as NFS – Network File System which is an open standard and Linux specific.

- Another common scenario is the storage being attached to the head node, and from there it is exported through network file system to the compute nodes – this will be the case for your virtualized cluster running Rocks. The shared storage resides on the head node and from there it is shared to the compute nodes


Fast network
(Gigabit, InfiniBand)
ComputeNode 1
INTERNET
ComputeNode 2
...
Researcher Machine
HeadNode
ComputeNode n