DR. BABASAHEB AMBEDKAR MARATHWADA UNIVERSITY, AURANGABAD

Department of Computer Science & Information Technology

(2022-2023)

SEMINAR ON

"Cluster Computing"

Master of Science (Information Technology)

(SEM 4)

GUIDANCE BY

Dr. C. Namrata Mahender

SUBMITTED BY

Miss. Manasi Manohar Sapkale

Department of Computer Science and Information Technology,

Aurangabad

CERTIFICATE

This is to certify that the seminar report entitled **"Cluster Computing"** has been submitted by **Miss. Manasi Manohar Sapkale** student of M.SC Information Technology (CSI475). Department of Computer Science and Information Technology, Aurangabad

In the partial fulfillment for the requirement of award Master of computer science degree of Dr. Babasaheb Ambedkar MarathwadaUniversity, Aurangabad in the academic year 2022-2023 is a record of student own study carried under my supervision and guidance.

This report has not been submitted to any other university or institution for the award of the any degree.

Paper Code: CSI475

Seat Number: _____

Dr. C. Namrata Mahender **Pro. Sachin N. Deshmukh**

(Assistant Professor) **(HOD)**

External Examiner

ABSTRACT

The High-Performance Computing allows scientists and engineers to deal with very complex problems using fast computer hardware and specialized software. Since often these problems require hundreds or even thousands of processor hours to complete, an approach, based on the use of supercomputers, has been traditionally adopted. Recent tremendous increase in a speed of PC-type computers opens relatively cheap and scalable solution for High-Performance Computing, using cluster technologies.

The conventional Massively Parallel Processing supercomputers are oriented on the very high-end of performance. As a result, they are relatively expensive and require special and also expensive maintenance support. Better understanding of applications and algorithms as well as a significant improvement in the communication network technologies and processors speed led to emerging of new class of systems, called clusters of networks of workstations, which are able to compete in performance with Massively Parallel Processing supercomputers and have excellent price/performance ratios for special applications types.

A cluster is a group of independent computers working together as a single system to ensure that mission-critical applications and resources are as highly available as possible. The group is managed as a single system, shares a common namespace, and is specifically designed to tolerate component failures, and to support the addition or removal of components in a way that's transparent to users. In this paper we will introduce the basics of cluster technology.

INDEX

1. Introduction

Figure 1: Cluster Computing

Introduction Cluster Computing

Cluster computing defines several computers linked on a network and implemented like an individual entity. Each computer that is linked to the network is known as a node.

Cluster computing provides solutions to solve difficult problems by providing faster computational speed, and enhanced data integrity. The connected computers implement operations all together thus generating the impression like a single system (virtual device). This procedure is defined as the transparency of the system.

A computer cluster is a set of computers that work together so that they can be viewed as a single system. Unlike grid computers, computer clusters have each node set to perform the same task, controlled and scheduled by software.

The components of a cluster are usually connected to each other through fast local area networks, with each node (computer used as a server) running its own instance of an operating system. In most circumstances, all of the nodes use the same hardware and the same operating system, although in some setups (e.g. using Open Source Cluster Application Resources, different operating systems can be used on each computer, or different hardware.

Cluster computing refers that many of the computers connected on a network and they perform like a single entity. Each computer that is connected to the network is called a node. Cluster computing offers solutions to solve complicated problems by providing faster computational speed, and enhanced data integrity. The connected computers execute operations all together thus creating the impression like a single system (virtual machine). This process is termed as transparency of the system. Based on the principle of distributed systems, this networking technology performs its operations. And here, LAN is the connection unit. This process is defined as the transparency of the system.

Cluster computing goes with the features of:

- All the connected computers are the same kind of machines
- They are tightly connected through dedicated network connections
- All the computers share a common home directory.

Clusters hardware configuration differs based on the selected networking technologies. Cluster is categorized as Open and Close clusters wherein Open Clusters all the nodes need IP's and those are accessed only through the internet or web. This type of clustering causes enhanced security concerns. And in Closed Clustering, the nodes are concealed behind the gateway node and they offer increased protection.

Key components of cluster computing include:

- Cluster nodes
- Cluster Operating System
- Switch or node interconnect
- Network switching hardware

Cluster computing is the process of performing a computational task across multiple computers which are connected generally in the local area network (LAN) and perform like a single entity. Each computer in the network is represented as a node. The nodes execute the task in tandem making it look like one large system responding to the user requests. This is termed as transparency of the system. Cluster computing plays a major role in high traffic applications which have the requirement to extend the processing capability and with zero downtime.

2. History of Cluster Computing

The history of computer clusters is best captured by a footnote in Greg Pfister's in Search of Clusters: “Virtually every press release from mentioning clusters says, who invented clusters’ IBM did not invent them either. Customers invented clusters, as soon as they could not fit all their work on one computer, or needed a backup. The date of the first is unknown, but it would be surprising if it was not in the 1960s, or even late 1950s.”

The formal engineering basis of cluster computing as a means of doing parallel work of any sort was arguably invented by Gene Amdahl of IBM, who in 1967 published what has come to be regarded as the seminal paper on parallel processing: Amdahl's Law. Amdahl's Law describes mathematically the speedup one can expect from parallelizing any given otherwise serially performed task on a parallel architecture. This article defined the engineering basis for both multiprocessor computing and cluster computing, where the primary differentiator is whether or not the interposes communications are supported "inside" the computer (on for example a customized internal communications bus or network) or "outside" the computer on a commodity network.

Consequently, the history of early computer clusters is more or less directly tied into the history of early networks, as one of the primary motivations for the development of a network was to link computing resources, creating a de facto computer cluster. Packet switching networks were conceptually invented by the RAND(research and development) corporation in 1962.

Using the concept of a packet switched network, the ARPANET(Advanced Research Projects Agency Network**)** project succeeded in creating in 1969 what was arguably the world's first commodity-network based computer cluster by linking four different computer centers (each of which was something of a "cluster" in its own right, but probably not a commodity cluster).

The ARPANET project grew into the Internet—which can be thought of as "the mother of all computer clusters" (as the union of nearly all of the compute resources, including clusters, that happen to be connected). It also established the paradigm in use by all computer clusters in the world today—the use of packet-switched networks to perform interposes communications between processor (sets) located in otherwise disconnected frames.

The development of customer-built and research clusters proceeded hand in hand with that of both networks and the Unix operating system from the early 1970s, as both TCP/IP and the project

created and formalized protocols for network-based communications. The Hydra operating system was built for a cluster of minicomputers called C.mmp at Carnegie Mellon University in 1971. However, it was not until circa 1983 that the protocols and tools for easily doing remote job distribution and file sharing were defined (largely within the context of Unix, as implemented by Sun Microsystems) and hence became generally available commercially, along with a shared filesystem.

The first commercial clustering product was net, developed by Datapoint in 1977. net was not a commercial success and clustering per se did not really take off until Digital Equipment Corporation released their VA cluster product in 1984 for the operating system.

The net and cluster products not only supported parallel computing, but also shared file systems and peripheral devices. The idea was to provide the advantages of parallel processing, while maintaining data reliability and uniqueness. cluster, now McCluster, is still available on OpenVMS running on Alpha, Itanium and x86-64 systems.[2]

Two other noteworthy early commercial clusters were the Tandem Himalaya (a circa 1994 high-availability product) and the IBM S/390 Parallel Simplex (also circa 1994, primarily for business use).

No history of commodity computer clusters would be complete without noting the pivotal role played by the development of Parallel Virtual Machine (PVM) software in 1989. This open source software based on TCP/IP communications enabled the instant creation of a virtual supercomputer—a high performance compute cluster—made out of any TCP/IP connected systems. Free-form heterogeneous clusters built on top of this model rapidly achieved total throughput in that greatly exceeded that available even with the most expensive "big iron" supercomputers. PVM and the advent of inexpensive networked PCs led, in 1993, to a NASA project to build supercomputers out of commodity clusters.

In 1995 the Beowulf cluster—a cluster built on top of a commodity network for the specific purpose of "being a supercomputer" capable of performing tightly coupled parallel computations—was invented,[3] which spurred the independent development of grid computing as a named entity, although Grid-style clustering had been around at least as long as the Unix operating system and the Arpanet, whether or not it, or the clusters that used it, were named.

3. Types of Cluster Computing

3.1. High Availability (HA) and Failover Clusters

These cluster models generate the availability of services and resources in an uninterrupted technique using the system's implicit redundancy.

The basic term of Cluster is that if a node declines, then applications and services can be made available to different nodes. These methods of clusters deliver as the element for critical missions, mails, documents, and application servers.

In the fail-over configuration, services run in one computing node while the other waits to take over during outages. It is mainly used to add failure resiliency. If any main node service fails, the manager node moves the virtual IP to its backup node.

Also, the failed node loses data access to its standby. That helps avoid the risk of multiple writes to duplicate files. So, when the backup node takes over, it will take the necessary steps to re-establish the services. For instance, checking the data integrity, reapplying uncommitted journal entries, and so on. The transactions that were in process during the outage will fail.

So, we'll see some downtime while the cluster is reconfigured. Then, the design of such a cluster must access the maximum acceptable downtime. This configuration is appropriate when the software systems do not support concurrent service instances consistently.

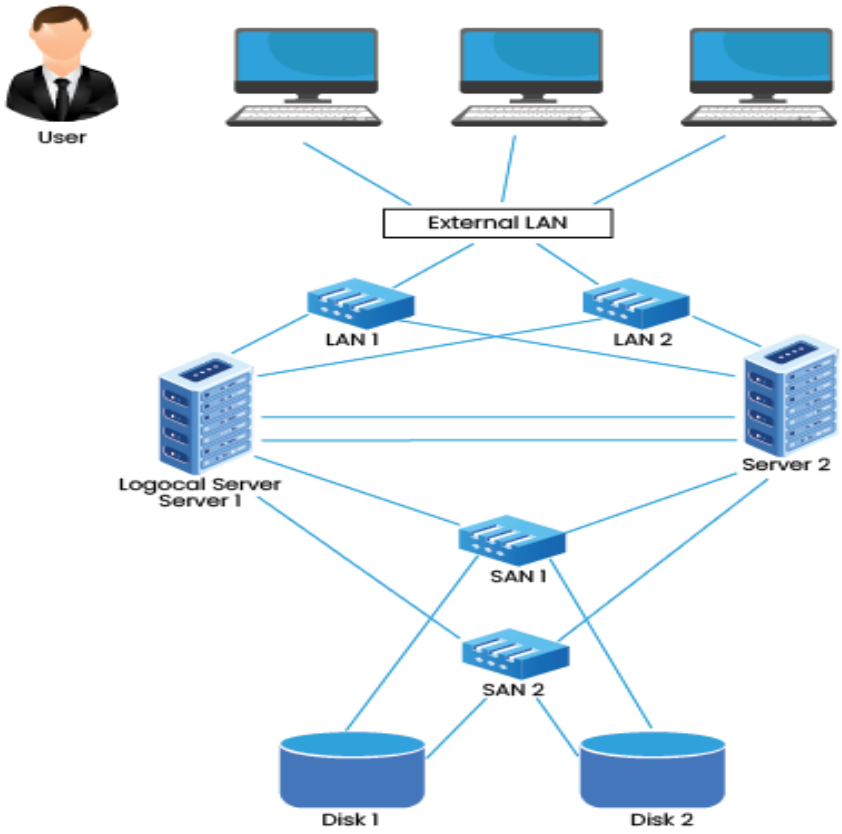

Figure 2: High Availability (HA) and Failover Clusters

The main benefit of fail-over clusters is that they don't require modifications to existing software. For Linux, the more known open-source implementation is the Linux HA. However, a handful of commercial software implementations from Legato, Veritas, Oracle, IBM, and others exist.

3.2. Load Balancing Clusters

This cluster allocates all the incoming traffic/requests for resources from nodes that run the equal programs and machines.

In this cluster model, some nodes are answerable for tracking orders, and if a node declines, therefore the requests are distributed amongst all the nodes available. Such a solution is generally used on web server farms.

In the load balancing cluster, the load is distributed among the available computing nodes. The techniques to distribute the load varies, round-robin the user requests or connections is the more common. The more transactions are independent of each other, the best. That means we want the computing nodes as independent as possible from each other. One transaction from one node should not need to wait for another transaction in another host. That is the concept of parallelism, we have a good tutorial on parallel processing that shows how it works. Load balancing is cost-friendlier than fail-over. As the computing nodes share the load, the overall transaction throughput improves. In node failure events, the load balancer redistributes the requests to the remaining online nodes.

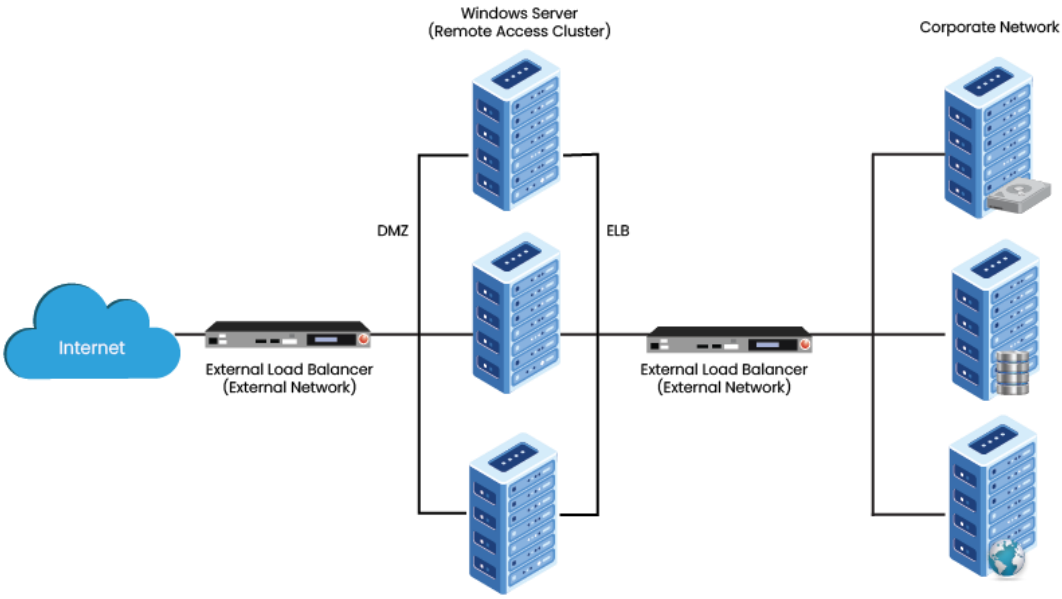

Figure 3: Load Balancing Clusters

3.3. Distributed & Parallel Processing Cluster

This cluster model boosts availability and implementation for applications that have huge computational tasks.

A large computational task has been divided into smaller tasks and distributed across the stations. Such clusters are generally used for numerical computing or financial analysis that needs high processing power.

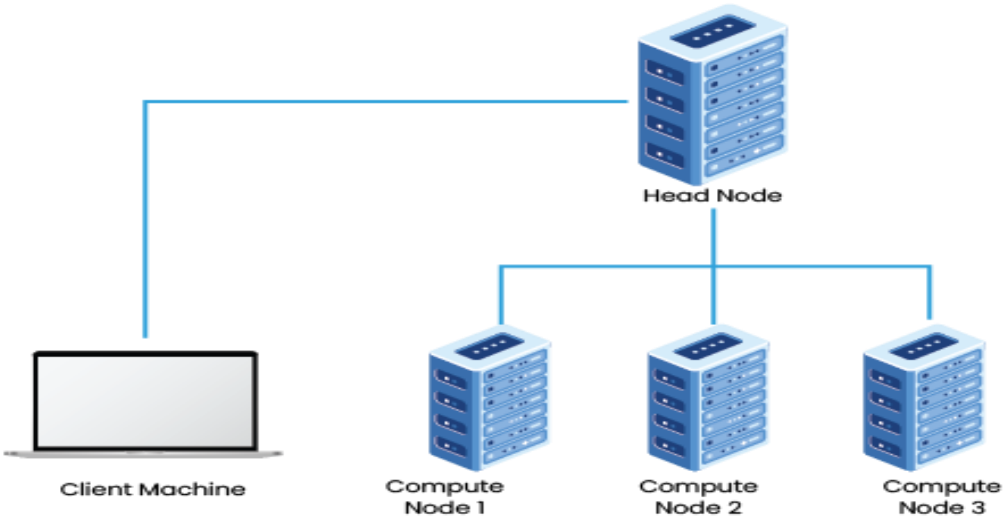

Figure 4: Distributed & Parallel Processing Clusters

4. Benefits of Cluster Computing

Cluster computing offers a wide array of benefits. Some of these include the following-

Cost-Effectiveness – Compared with the mainframe systems, cluster computing is considered to be much more cost-effective. These computing systems offer enhanced performance with respect to the mainframe computer devices. Cluster computing is considered to be much more cost effective. These computing systems provide boosted implementation concerning the mainframe computer devices.

Processing Speed – The processing speed of cluster computing is justified with that of the mainframe systems and other supercomputers present in the world. The processing speed of cluster computing is validated with that of the mainframe systems and other supercomputers demonstrate around the globe.

Expandability – Scalability and expandability are another set of advantages that cluster computing offers. Cluster computing represents an opportunity for adding any number of additional resources and systems to the existing computing network.

Increased Resource Availability – Availability plays a vital role in cluster computing systems. Failure of any connected active node can be easily passed on to other active nodes on the server, ensuring high availability.

Improved Flexibility – In cluster computing, superior specifications can be upgraded and extended by adding newer nodes to the existing server.

Benefits of Cluster computing include:

Cluster computing is relatively inexpensive when compared to large server machines. Not just for scalability, you can also use cluster computing to increase the available time, load balancing, processing speed, etc.

The nodes that are connected in cluster computing are identical machines. They can be either tightly or loosely coupled through the dedicated network. They have the same home directory.

Hardware configuration differs based on the networking technologies. It can be either Open or Closed, where open clusters are the ones that make the node available over the internet and can be accessed through IPs. As they are publicly available, they need to have enhanced security features. Whereas, in the case of closed clustering, the nodes behind a gateway node increase their protection.

5. Need of Cluster Computing

Why is Cluster Computing important?

Cluster computing gives a relatively inexpensive, unconventional to the large server or mainframe computer solutions.

It resolves the demand for content criticality and process services in a faster way.

Many organizations and IT companies are implementing cluster computing to augment their scalability, availability, processing speed and resource management at economic prices.

It ensures that computational power is always available.

It provides a single general strategy for the implementation and application of parallel high-performance systems independent of certain hardware vendors and their product decisions.

Everyone might have faced the situation of low-speed services and content criticality. Cluster computing resolves the need for content criticality and process services in a quicker approach. As Internet Service Providers look for enhanced availability in a scalable approach, cluster computing will provide this.

And even, this technology is the heavy need for the film industry as they require this for rendering extended quality of graphics and cartoons. Implementation of the cluster through the Beowulf method also resolves the requirement of statistics, fluid dynamics, genetic analysis, astrophysics, economics, neural networks, engineering, and finance.

Many of the organizations and IT giants are implementing this technology to augment their scalability, processing speed, availability and resource management at the economic prices.

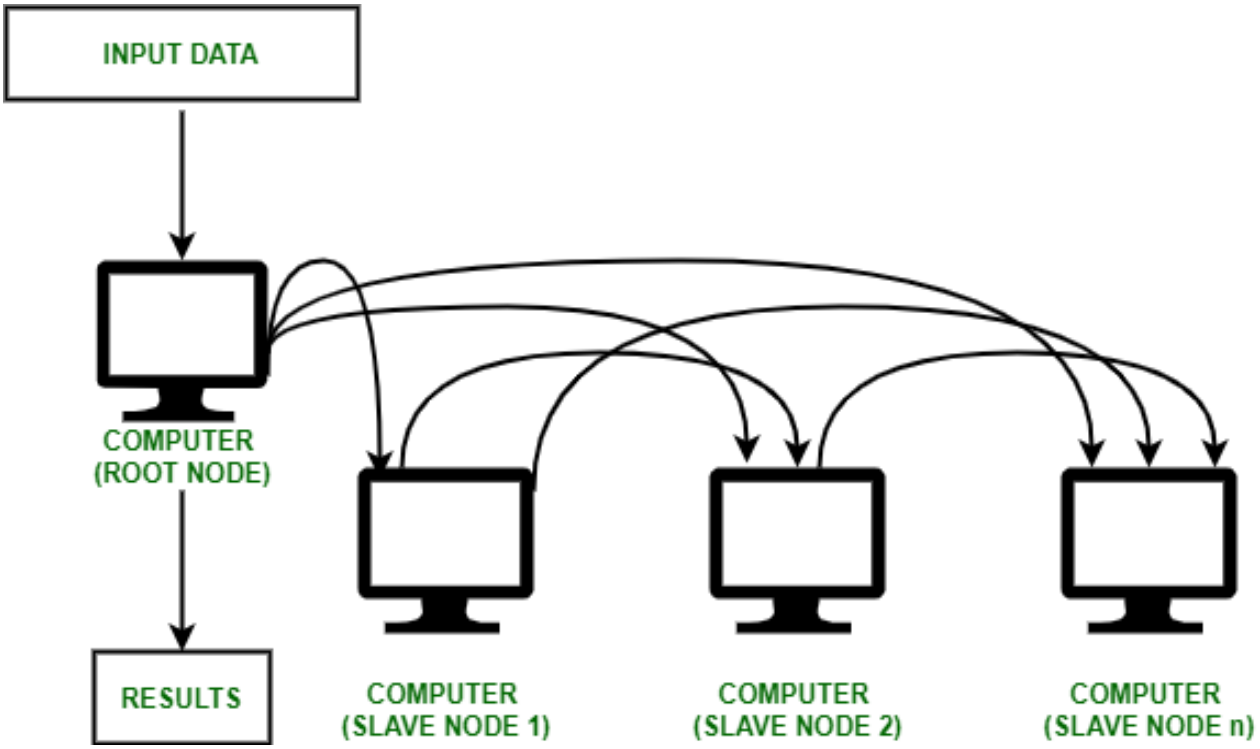

Figure 5: Cluster Computing

Clusters or combinations of clusters are used when the content is critical, and services need to be available. Internet Service Providers and e-commerce sites demand high availability and load balancing in a scalable manner. The parallel clusters are being extensively used in the film industry as they need high-quality graphics and animations. Talking about the Beowulf clusters, they are dominantly used in science, engineering, and finance to perform various critical projects. Researchers, organizations, and businesses use clusters to demand enhanced scalability, resource management, availability, and processing at affordable price.

6. Classification of Cluster

6.1. Open Cluster:

IPs are needed by every node and those are accessed only through the internet or web. This type of cluster causes enhanced security concerns.

6.2. Close Cluster:

The nodes are hidden behind the gateway node, and they provide increased protection. They need fewer IP addresses and are good for computational tasks.

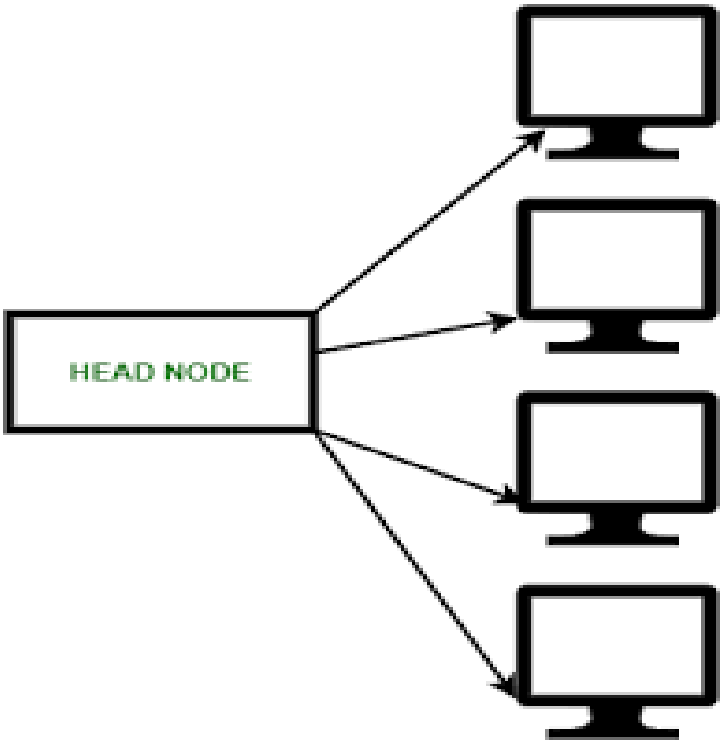

Figure 6: Classification of Cluster

7. Cluster Computing Architecture

It is designed with an array of interconnected individual computers and the computer systems operating collectively as a single standalone system.

It is a group of workstations or computers working together as a single, integrated computing resource connected via high-speed interconnects.

A node – Either a single or a multiprocessor network having memory, input and output functions and an operating system.

Two or more nodes are connected on a single line or every node might be connected individually through a LAN connection.

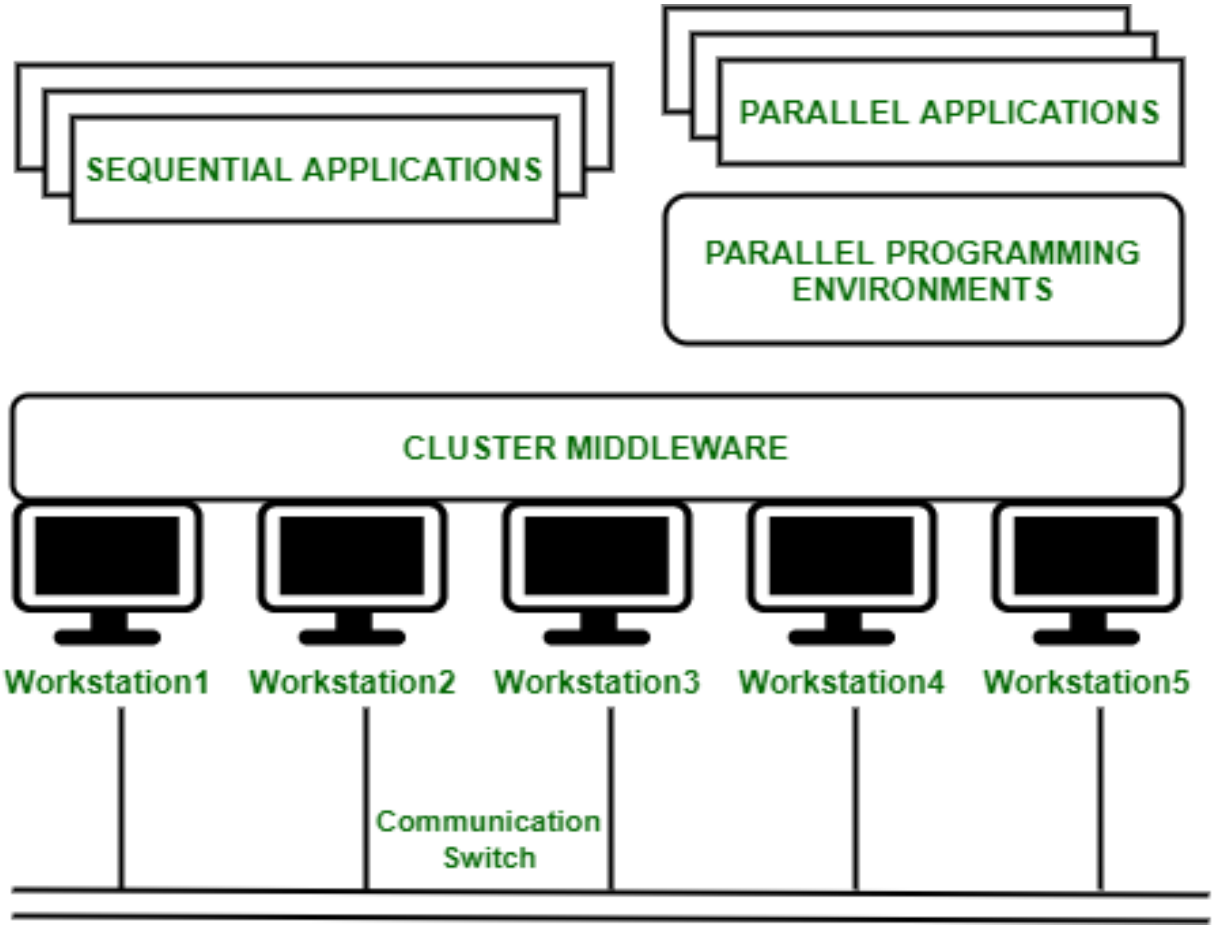

Figure 7: Cluster Computing Architecture

8. Components of a Cluster Computing

- Cluster Nodes
- Cluster Operating System
- The switch or node interconnect
- Network switching hardware

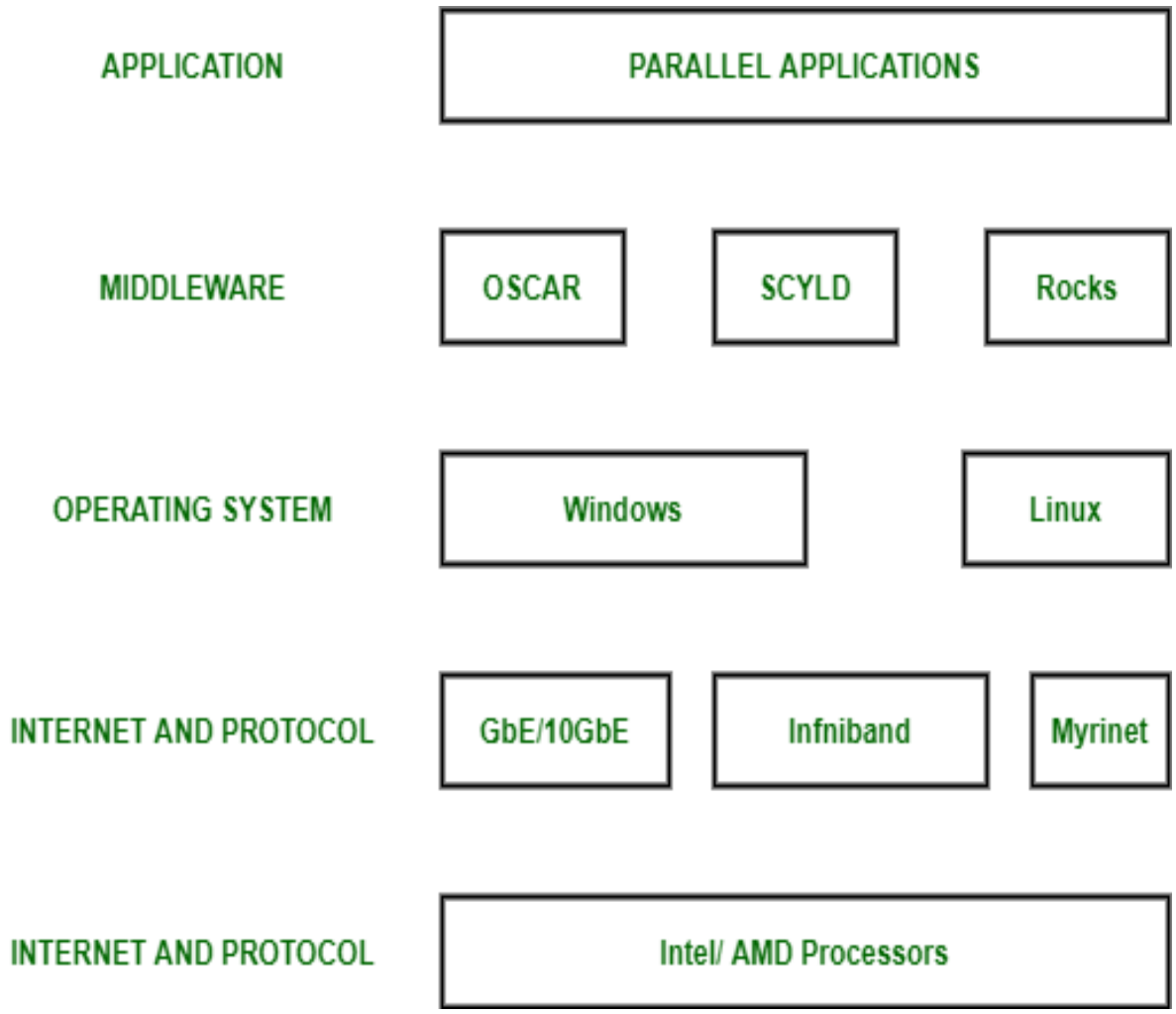

Figure 8: Components of a Cluster Computing

9. Disadvantages of Cluster Computing

9.1. High cost:

It is not so much cost-effective due to its high hardware and its design.

9.2. Problem in finding fault:

It is difficult to find which component has a fault.

9.3. More space is needed:

Infrastructure may increase as more servers are needed to manage and monitor.

10. Applications of Cluster Computing

- Various complex computational problems can be solved.
- It can be used in the applications of aerodynamics, astrophysics and in data mining.
- Weather forecasting.
- Image Rendering.
- Various e-commerce applications.
- Earthquake Simulation.
- Petroleum reservoir simulation.
- Data mining
- Applications of aerodynamics and astrophysics
- Weather forecasting
- Image rendering and processing
- E-commerce applications
- Earthquake and tornado simulation
- Petroleum reservoir simulation
- Electromagnetics
- Nuclear simulations

11. Conclusion

Cluster computing is cost-effective, expandable, and ensures the high availability of resources. They can be either loosely or tightly coupled to bind them together so that they can work as a single system to achieve the task.

Clusters can be created based on the requirements (high performance, high availability, or load balancing). Which makes it easier for the users to create it based on the system's needs. Also, it is possible to achieve load balancing along with high availability.

Cluster computing is cost-effective, expandable, and ensures the high availability of resources. They can be either loosely or tightly coupled to bind them together so that they can work as a single system to achieve the task. Clusters can be created based on the requirements (high performance, high availability, or load balancing). Which makes it easier for the users to create it based on the system's needs. Also, it is possible to achieve load balancing along with high availability.

12. Reference

1. **https://www.google.com**
2. **https://www.geeksforgeeks.org/an-overview-of-cluster-computing/**
3. **https://en.wikipedia.org/wiki/Computer_cluster**
4. **https://www.tutorialspoint.com/what-is-cluster-computing**
5. **https://www.sciencedirect.com/topics/computer-science/cluster-computing**