STM JOURNALS

International Journal of
Algorithms Design and Analysis Review

Volume 1, Issue 1, 2023
January–June
DOI (Journal): 10.37591/IJADAR

https://journals.stmjournals.com/ijadar

Review IJADAR

Effects of Cluster Computing on Big Data Analysis and Network Topology

Sanchi S. Achalkhamb[1,*], Krishna T. Madrewar[2]

Abstract

The rapid expansion of big data has posed substantial difficulties for conventional computing systems. As a result, cluster computing has grown to be a potent method for effective large data processing. Cluster computing involves multiple interconnected nodes functioning as a unified system, pooling together their processing, storage, and memory resources. These nodes are typically connected through high-speed networks such as ethernet or InfiniBand, facilitating efficient data sharing and communication among them. Big data has made cluster computing frameworks like Apache Hadoop and Apache Spark very popular. These frameworks provide accessible tools and libraries that make creating and running parallel computing tasks on a cluster easier. Additionally, they have fault tolerance methods to ensure system resilience in the event of node failures, protecting data integrity and allowing computation to continue without interruption. By leveraging interconnected computers working in unison, cluster computing enables parallel processing, leading to faster and more scalable data analysis. This paper examines the effects of cluster computing on both big data analysis and network topology.

Keywords: Big data analysis, cluster computing, network topology, computing systems, high-performance computing, high bandwidth

INTRODUCTION

In recent years, the explosion of data volumes and the increasing complexity of data analysis have posed significant challenges for traditional computing systems. The emergence of big data, characterized by its massive scale, high velocity, and diverse formats, has necessitated the development of new approaches and technologies to effectively process and extract valuable insights from these vast data sets [1]. Cluster computing has emerged as a powerful solution to address the computational requirements of big data analysis, revolutionizing the field of data processing, and transforming the way organizations handle and derive value from their data as shown in Figure 1.

Cluster computing refers to the utilization of interconnected computers, or nodes, working together as a unified system to solve complex computational problems. By harnessing the processing power and storage capacity of multiple machines, cluster computing provides a scalable and high-performance infrastructure for handling large-scale data sets. This distributed computing paradigm allows for parallel processing, enabling faster data processing and analysis compared to traditional single-node systems [2].

***Author for Correspondence**
Sanchi S. Achalkhamb
E-mail: sanchiachalkhamb@gmail.com

[1]Student, Department of Electronics and Telecommunication Engineering, Deogiri Institute of Engineering and Management Studies College, Chhatrapati Sambhajinagar, Maharashtra, India
[2]Assistant Professor, Department of Electronics and Telecommunication Engineering, Deogiri Institute of Engineering and Management Studies College, Chhatrapati Sambhajinagar, Maharashtra, India

Received Date: June 21, 2023
Accepted Date: June 30, 2023
Published Date: July 29, 2023

Citation: Sanchi S. Achalkhamb, Krishna T. Madrewar. Effects of Cluster Computing on Big Data Analysis and Network Topology. International Journal of Algorithms Design and Analysis Review. 2023; 1(1): 31–39p.

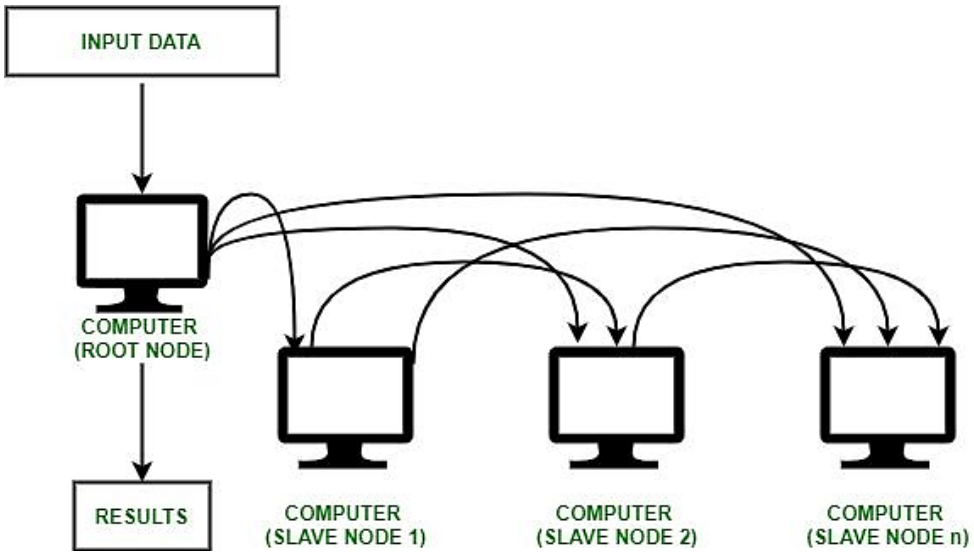

Figure 1. Cluster computing layout.

The effects of cluster computing on big data analysis have been profound. With its ability to divide data processing tasks across multiple nodes, cluster computing offers a distributed and parallel approach to data analysis, leading to significant reductions in processing time. Complex algorithms and computationally intensive tasks that would have been infeasible with single-node systems can now be executed efficiently by dividing the workload among the nodes in a cluster.

HISTORY OF CLUSTER COMPUTING

Cluster computing, which involves interconnecting multiple computers or servers to function as a unified system, has a history that dates back to the mid-1970s. Over the years, it has undergone significant evolution, resulting in increased processing power, improved performance, and enhanced reliability [3].

Early Years (1970s–1980s)

In the 1970s, computer scientists began experimenting with the concept of connecting multiple computers to work together on a task. These early systems often relied on specialized hardware and proprietary software.

One notable project during this period was the Ethernet-based Xerox Alto computer, developed at Xerox PARC in the mid-1970s. The Alto featured networked workstations that could communicate and collaborate on tasks.

In the 1980s, the advent of local area networks (LANs) provided a foundation for cluster computing. Workstations and servers could be interconnected using network technologies like ethernet, enabling resource sharing and distributed computing.

Beowulf Clusters (1990s)

The term “Beowulf cluster” was coined in 1994 by Thomas Sterling and Donald Becker [4] at NASA. They accomplished the development of a high-performance computing (HPC) architecture by utilizing readily available components and open-source software.

Beowulf clusters became popular in the 1990s as a cost-effective solution for building supercomputers. These clusters utilized commodity hardware, typically x86-based computers, and a Linux operating system [5].

The Beowulf project influenced the development of open-source software tools and libraries for parallel computing, such as Message Passing Interface (MPI).

High-performance Computing Clusters (2000s–2010s)

In the 2000s and 2010s, cluster computing gained prominence in the field of high-performance computing. HPC clusters were built using powerful servers interconnected with high-speed networks like InfiniBand as shown in Figure 2.

Supercomputers, such as those in the TOP500 list, often employed cluster architectures. These systems consisted of thousands or even millions of processing cores, enabling extensive computational capabilities [6].

Parallel programming models and frameworks, such as OpenMP and OpenCL, were developed to facilitate software development for HPC clusters. These frameworks allowed developers to write code that could take advantage of the distributed computing resources.

Cloud Computing and Big Data (2010s–Present)

The advent of cloud computing in the late 2000s and early 2010s revolutionized the cluster computing landscape, as cloud providers began offering virtualized resources and scalable computing power. This enabled users to easily create and manage clusters as needed. Furthermore, the emergence of big data processing frameworks like Apache Hadoop and Apache Spark played a crucial role in popularizing the utilization of cluster computing. These frameworks enabled distributed processing of large datasets across clusters, providing scalability and fault tolerance [7].

BIG DATA ANALYSIS AND NETWORK TOPOLOGY

Big data analysis refers to the process of extracting valuable insights, patterns, and knowledge from large and complex datasets. Network topology, on the other hand, encompasses the physical or logical arrangement of nodes and connections within a computer network. The choice of network topology can have an impact on the performance and scalability of big data analysis systems [8]. Here is how network topology can influence big data analysis.

Figure 2. Technicians working on a cluster consisting of several servers.

Data Distribution and Parallel Processing

Big data analysis often involves distributing the dataset across multiple nodes or servers for parallel processing. The network topology affects how data is distributed and accessed by the processing nodes.

In a centralized topology, where all the data is stored in a single location, the network can become a bottleneck as the data needs to be transferred across the network for processing. This can lead to latency and reduced performance.

Distributed network topologies, such as a star or mesh network, can offer better performance for big data analysis. In these topologies, data can be stored and processed locally on each node, reducing network traffic, and improving processing speed.

Data Movement and Communication

Big data analysis often involves data movement and communication between different processing stages or nodes. The network topology can impact the efficiency of these data transfers.

A well-designed network topology with high-bandwidth links can facilitate fast and efficient data transfers between nodes. This aspect becomes particularly crucial when handling large volumes of data in the context of big data analysis.

Network topologies with low-latency connections, such as a fully connected mesh or a high-speed interconnect like InfiniBand, can minimize communication delays and improve the overall performance of big data analysis systems as shown in Figure 3.

EFFECTS OF CLUSTER COMPUTING ON BIG DATA ANALYSIS AND NETWORK TOPOLOGY

Cluster computing has a significant impact on big data analysis and network topology. Here are some effects of cluster computing on these areas.

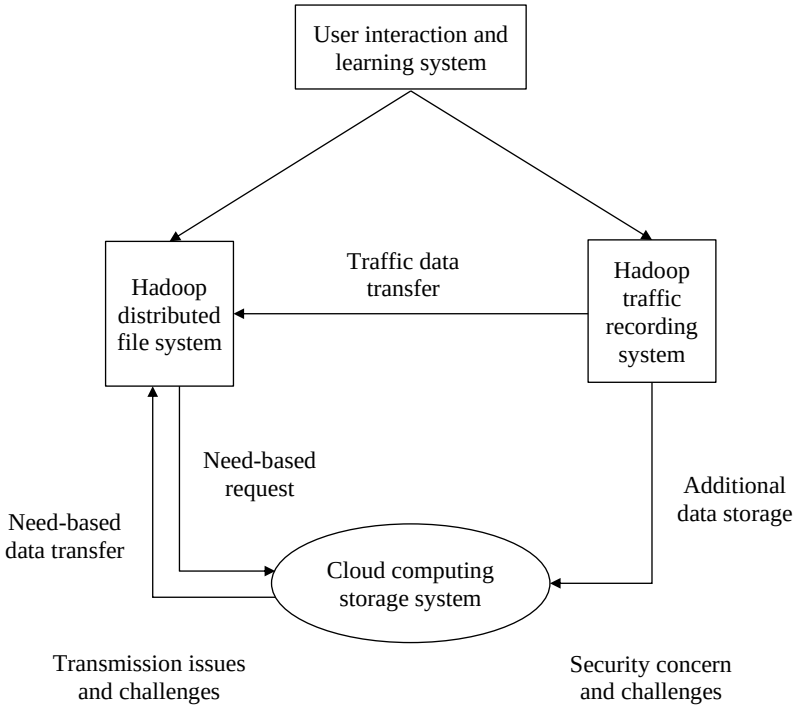

Figure 3. Network topology for big data analytics.

Enhanced Processing Power

Cluster computing combines the computational resources of multiple machines, allowing for increased processing power. This heightened processing capability facilitates accelerated data processing and analysis in big data applications.

With cluster computing, big data analysis tasks can be distributed across multiple nodes or servers in the cluster, enabling parallel processing. This parallelization speeds up the overall data analysis process, leading to improved efficiency.

Scalability

Cluster computing provides scalability for big data analysis. As the volume of data grows or the processing job becomes more complicated, more nodes can be added to the cluster to meet the increased demand.

Scalable network topologies, such as mesh or tree topologies, are often used in cluster computing environments. These topologies allow for easy expansion by adding more nodes to the network without significant disruption.

To store and handle massive datasets, cluster computing systems frequently use distributed file systems like Hadoop Distributed File System (HDFS). These distributed storage systems are designed to handle big data and provide fault tolerance and high availability.

Cluster computing facilitates data locality by distributing data over numerous nodes in the cluster. This means that data is processed on the same node where it is stored, decreasing network data transit and enhancing performance as shown in Figure 4.

Improved Fault Tolerance and Reliability

Cluster computing offers improved fault tolerance and reliability for big data analysis. If a node or server in the cluster fails, the workload can be automatically transferred to other functioning nodes, ensuring continuous processing without significant interruptions. Network topologies in cluster computing environments often incorporate redundancy and alternative paths to mitigate the impact of node or link failures. This redundancy enhances the fault tolerance and resilience of the network infrastructure [9].

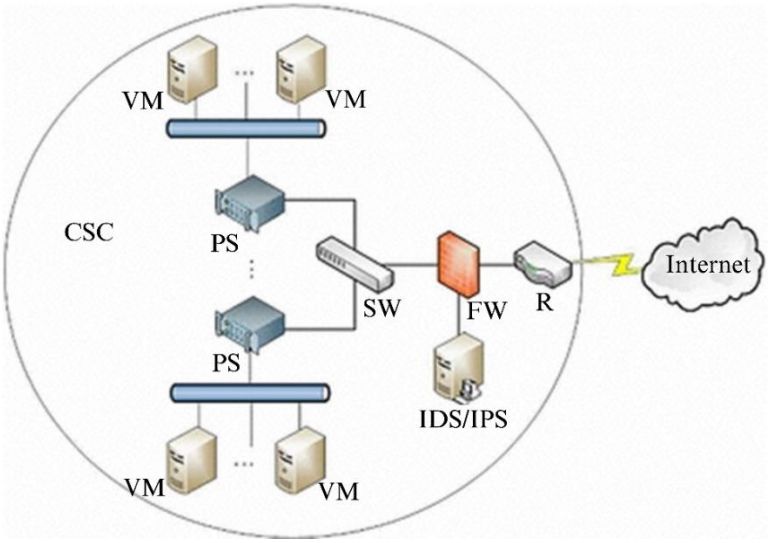

Figure 4. Network topology of a cloud server cluster.

Efficient Data Communication

Cluster computing systems employ high-speed network interconnects, such as InfiniBand or 10/25/40/100 gigabit ethernet, to ensure efficient data communication between nodes in the cluster.

The choice of network topology in cluster computing environments is influenced by the need for low-latency and high-bandwidth communication [10].

COST REQUIREMENT IN INDIA

Presently, there are commercial turn-key clusters available at remarkably affordable prices, starting as low as $10,000. These clusters typically include 8 to 12 cores, an interconnect, discs for storage, and an operating system, providing a compact yet powerful solution for effective cluster computing as shown in Table 1 [11].

Table 1. A comparison of different systems.

System	Job processing type	QoS attributes	Job composition	Resource allocation control	Platform support	Evaluation method	Process migration
Enhanced MOSIX	Parallel	Cost	Single task	Decentralized	Heterogeneous	User-centric	Yes
Gluster	Parallel	Reliability (no point of failure)	Parallel task	Decentralized	Heterogeneous	N/A	Yes
Faucets	Parallel	Time, cost	Parallel task	Centralized	Heterogeneous	System-centric	Yes
DQS	Batch	CPU memory sizes, hardware architecture and OS versions.	Parallel task	Decentralized	Heterogeneous	System-centric	No
Tycoon	Sequential	Time, cost	Multiple task	Decentralized	Heterogeneous	User-centric	No
Cluster-on-demand	Sequential	Cost in terms of time	Independent	Decentralized	Heterogeneous	User-centric	No
Kerrighed	Sequential	Ease of use, high performance, high availability, efficient resources management, and high customizability of the OS	Multiple task	Decentralized	Homogeneous	System-centric	Yes
Open SSI	Parallel	Availability, scalability and manageability	Multiple task	Decentralized	Heterogeneous	System-centric	Yes
Libra	Batch, sequential	Time, cost	Parallel	Centralized	Heterogeneous	System-centric, user-centric	Yes
PVM	Parallel, concurrent	Cost	Multiple task	Centralized	Heterogeneous	User-centric	Yes

Condor	Parallel	Throughput, productivity of computing environment	Multiple task	Centralized	Platform support	System-centric	Yes
REXEC	Parallel, sequential	Cost	Independent, single task	Decentralized	Homogeneous	User-centric	No
GNQS	Batch, parallel	Computing power	Parallel processing	Centralized	Heterogeneous	System-centric	No
Load Leveler	Parallel	Time, high availability	Multiple task	Centralized	Heterogeneous	System-centric	Yes
LSF	Parallel batch	Job submission simplification, setup time reduction and operation errors	Multiple task	Centralized	Heterogeneous	System-centric	Yes
SLURM	Parallel	Simplicity, scalability, portability and fault tolerance	Multiple Task	Centralized	Homogeneous	System-centric, user-centric	No
PBS	Batch	Time, jobs queuing	Multiple task	Centralized	Heterogeneous	System-centric	Yes

ADVANTAGES OF CLUSTER COMPUTING ON BIG DATA ANALYSIS AND NETWORK TOPOLOGY

1. *Enhanced performance:* Cluster computing allows for parallel processing and distributed computing, which significantly improves the performance of big data analysis and network operations. It enables the processing of large volumes of data in a shorter time and facilitates faster network data transfer and communication.
2. *Scalability:* Cluster computing provides scalability for both big data analysis and network topology. As the data volume or network traffic increases, additional nodes can be added to the cluster, allowing for increased processing power and network capacity. This scalability ensures that organizations can handle growing data demands and network loads effectively.
3. Clusters are meticulously designed with fault tolerance as a primary consideration, ensuring high availability of services and data. In the event of a node failure or network issue, the workload or network traffic can be automatically rerouted to other nodes, ensuring uninterrupted data analysis and network connectivity. This fault tolerance and high availability improve the system's reliability and resilience.
4. *Cost efficiency:* Cluster computing offers cost efficiency for both big data analysis and network topology. By utilizing commodity hardware and leveraging parallel processing, organizations can achieve high performance at a lower cost compared to investing in expensive high-end servers or network equipment.

DISADVANTAGES OF CLUSTER COMPUTING ON BIG DATA ANALYSIS AND NETWORK TOPOLOGY

1. *Complexity and management overhead:* Implementing and managing cluster computing for big data analysis and network topology can be complex and require specialized knowledge and skills. Configuring and maintaining the cluster, troubleshooting issues, and optimizing performance can be challenging and time-consuming.
2. *Communication overhead:* In a cluster, data needs to be distributed across nodes, and network communication is required for coordination and synchronization. This introduces communication overhead, which can impact overall performance and introduce potential bottlenecks.

3. *Single point of failure:* While cluster computing provides fault tolerance, there is still a risk of a single point of failure within the system. If the central management node or a critical network component fails, it can disrupt the entire cluster or network, affecting both data analysis and network operations.
4. *Resource allocation and optimization:* Proper resource allocation and optimization are crucial for efficient cluster computing. Ensuring that resources are distributed effectively, balancing workloads, and optimizing network traffic require careful planning and ongoing management. Inadequate resource allocation or optimization can lead to underutilization of resources or performance degradation.

CONCLUSION

In conclusion, cluster computing offers significant advantages for both big data analysis and network topology. It enhances performance by enabling parallel processing and distributed computing, allowing organizations to process large volumes of data quickly and facilitate faster network operations. Scalability is another benefit, as clusters can be expanded by adding nodes to handle growing data demands and network loads effectively.

Fault tolerance and high availability are requirements for cluster computing. The ability to automatically reroute workloads or network traffic in case of failures ensures uninterrupted data analysis and network connectivity, enhancing the reliability and resilience of the system. Furthermore, cluster computing can be cost-efficient by utilizing commodity hardware and achieving high performance at a lower cost compared to expensive high-end servers or network equipment.

However, implementing and managing cluster computing for big data analysis and network topology can be complex and require specialized knowledge and skills. Communication overhead, potential single points of failure, and the need for resource allocation and optimization are challenges that organizations must address. Furthermore, resolving problems and optimizing performance in a dispersed environment might take time.

Acknowledgements

This article was supported/partially supported by Deogiri Institute of Engineering and Management Studies. We thank our colleagues from DIEMS who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

We thank STM Journals for assistance with particular technique, methodology, and Prof. K. T. Madrewar for comments that greatly improved the manuscript.

We would also like to express our gratitude to the Prof. Dr. R. M. Autee for sharing his pearls of wisdom with us during the course of this article, and we thank three “anonymous” reviewers for their so-called insights. We are also immensely grateful to for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons.

REFERENCES

1. Buyya R, Vecchiola C, Selvi ST. Mastering Cloud Computing: Foundations and Applications Programming. Cambridge, MA: Morgan Kaufmann; 2013.
2. Mayer-Schönberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston, MA: Houghton Mifflin Harcourt; 2013.
3. Medhi D, Ramasamy K. Network Routing: Algorithms, Protocols, and Architectures. Cambridge, MA: Morgan Kaufmann; 2017.

4. Ridge D, Becker D, Merkey P, Sterling T. Beowulf: harnessing the power of parallelism in a pile-of-PCs. In: 1997 IEEE Aerospace Conference, Snowmass, CO, USA, February 13, 1997. Volume 2, pp. 79–91.
5. Lin J, Dyer C. Data-intensive text processing with MapReduce. In: Hirst G, series editor. Synthesis Lectures on Human Language Technologies #7. Kentfield, CA: Morgan & Claypool Publishers; 2010.
6. Sa-Ngasoongsong A, Kunthong J, Sarangan V, Cai X, Bukkapatnam ST. A low-cost, portable, high-throughput wireless sensor system for phonocardiography applications. Sensors. 2012; 12 (8): 10851–10870.
7. Miller TC, Stirlen C, Nemeth E. satool – a system administrator's cockpit, an implementation. In: Seventh System Administration Conference: LISA 1993, Monterey, CA, USA, November 5, 1993. pp. 119–130.
8. MarketWide Research. Cluster computing market analysis -– industry size, share, research report, insights, covid-19 impact, statistics, trends, growth and forecast 2023-2030. [Online]. 2023. MarkWide Research. 2023. Available at https://markwideresearch.com/cluster-computing-market/
9. Saturn Cloud. Hadoop how to unit test filesystem. [Online]. 2023. Available at https://saturncloud.io/blog/hadoop-how-to-unit-test-filesystem/
10. NAKIVO Team. High availability vs fault tolerance vs disaster recovery. [Online]. NAKIVO Team. 2018. Available at https://www.nakivo.com/blog/disaster-recovery-vs-high-availability-vs-fault-tolerance/
11. Mosley D. Network topology definitions – designing infrastructure Windows Server 2003. [Online]. 2023. Windows Server Brain. Available at https://www.serverbrain.org/designing-infrastructure-2003/network-topology-definitions.html

