Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

Contemporary High-Performance Computing for Big Data Applications

Dr.S.Ayyasamy

Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore-632014, Tamil Nadu, India.

E-mail: ayyasamyphd@gmail.com

Abstract

High-performance computing (HPC) involves leveraging parallel data processing to enhance computer performance and handle difficult tasks. HPC meets these aims by pooling computing capacity, enabling efficient, reliable, and prompt execution of even complex programs according to user demands and expectations. The rapid growth of HPDA in many sectors has led to the extension of the HPC market into new territory. HPC as well as Big Data systems differ not just in terms of technology but also in ecosystems. Extensive research in this sector has led to the emergence of various Big Data analytics models in recent years. As Big Data analytics spreads across several fields, new challenges about the usefulness of analytical paradigms also emerge. This article discusses the key analytical models, as well as the difficulties and challenges associated with high-performance data analytics. This research work aims to identify the factors influencing the integration of HPC with big data, including present and future trends. The study also proposes an architecture for big data with HPC convergence based on design principles.

Keywords: High-Performance Data Analytics, Big Data, HPC Convergence, Efficiently, Reliably.

1. Introduction

High Performance Computing (HPC) refers to systems that can process data and perform calculations at significantly faster rates than ordinary computers [1]. This collective

DOI: https://doi.org/10.36548/jitdw.2023.4.004
Received: 02-12-2023. received in revised form: 23-12-2023, accepted: 11-01-2024, published: 22-01-2024

processing capacity helps many research, commercial, and engineering organizations to tackle enormous challenges that could have been unsolvable [2]. HPC is the method of combining computing resources to achieve significantly more performance than a standard desktop or place of employment, in order to address complicated issues in research, engineering, and business.

Need of High-Performance Computing

- The need of high-performance computing develops when computational needs surpass the capability of traditional computer systems, allowing academics, scientists, and enterprises to tackle complicated issues and advance in their respective professions.
- It will accomplish an operation with a short deadline and conduct a large number of tasks per second.
- It increases processing speeds, that can be crucial for a variety of computer procedures, applications, and workflows.
- HPC is used for commercial applications, data warehousing, and transaction processing.
- Running complicated simulations and models to effectively anticipate complex patterns require high-performance computing.
- Training deep learning models and processing large datasets in domains like image recognition, self-driving cars, and natural language processing, need a lot of computing resources.

1.1 Main Constraints on Working of HPC

Parallel Processing with HPC for Big Data

High-performance computing (HPC) is becoming a required ability for students in computer science programmes. Parallel computing is becoming a more prominent topic in traditional curriculum. Big data's volume, diversity, and velocity, together with its important information, have led to the development of novel parallel data processing technologies beyond standard database management systems.

Systems can reduce programme execution time by spreading a task's various components among several processors. Parallel processing is enabled by multi-core processors, which are often found in modern computers, as well as any system using more than one CPU.

The amount of data and the necessity to gather, store, and process it have created several technological problems for Big Data applications. Technical problems include data variation, inconsistency and insufficiency, privacy and data ownership, size, and timeliness. Parallel and distributed computing is critical, particularly for addressing scale and timeliness concerns.

Advantages

Performance: One of the key benefits of parallel computing remains its ability to dramatically boost performance. Parallel computing, which distributes jobs over numerous processing units, may perform complicated computations and data-intensive activities quicker than sequential computing.

Scalability

Parallel computing has great scalability, which means it can effectively manage increasing workloads when the number of computational units grows.

Resource Utilization

Parallel computing maximises resource utilisation by utilising many processing units simultaneously. It guarantees that no processing power is wasted, maximising the efficiency using hardware resources.

Realtime Processing

Parallel computing enables these applications to satisfy real-time processing requirements, resulting in effortless and dynamic user experiences.

Limitations

- Synchronization Overhead
- Complexity
- Cost and Infrastructure.

HPC Clusters

High-performance computing (HPC) clusters address complicated problems that need enormous processing capacity. They are made up of numerous networked computers that run computations and simulations in simultaneously, allowing vast volumes of data to be processed more quickly and efficiently.

These clusters collaborate to offer the processing capacity required for analysing and processing enormous data volumes, simulating complicated systems, and solving challenging scientific and technical challenges. HPC as well as cloud computing make advantage of distributed computing systems, which include several computers collaborating to tackle complicated problems. HPC clusters have grown increasingly popular as organisations seek to process massive volumes of data fast and effectively.

Organizations acquire large volumes of data, which may be difficult to process. HPC clusters can analyse big data efficiently and effectively, allowing organisations to obtain insights into their data on real time. This is especially important in fields like banking and healthcare, where massive data volumes must be analysed fast in order to make sound judgements.

They have several uses, include scientific research, the field of engineering, financial evaluation, health care, and machine learning. With the rise of big data and the rising complexity associated with scientific and technical challenges, the need for HPC clusters will only continue to expand in the next years.

High-Performance Components

High-Performance Computing (HPC) systems for big data applications usually include a mix of hardware and software components intended to handle enormous amounts of data effectively.

All of the extra computing resources within an HPC cluster networking, storage, memory, along file system are high-speed, high-throughput, low-latency components that can keep up with the nodes while optimizing the cluster's computational power and performance.

Influence of HPC in Cloud

HPC in the cloud known as HPC as a service, or HPCaaS—provides a substantially quicker, more scalable, and more cost-effective option for businesses to leverage HPC [4]. HPCaaS often involves access to HPC clusters as well as infrastructure located in a cloud service provider's data centre, as well as ecosystem skills (including AI and data analytics), as well as HPC expertise.

You can implement your solution using private or public clouds. Most major cloud providers provide HPC services that are either packaged or build-your-own packages.

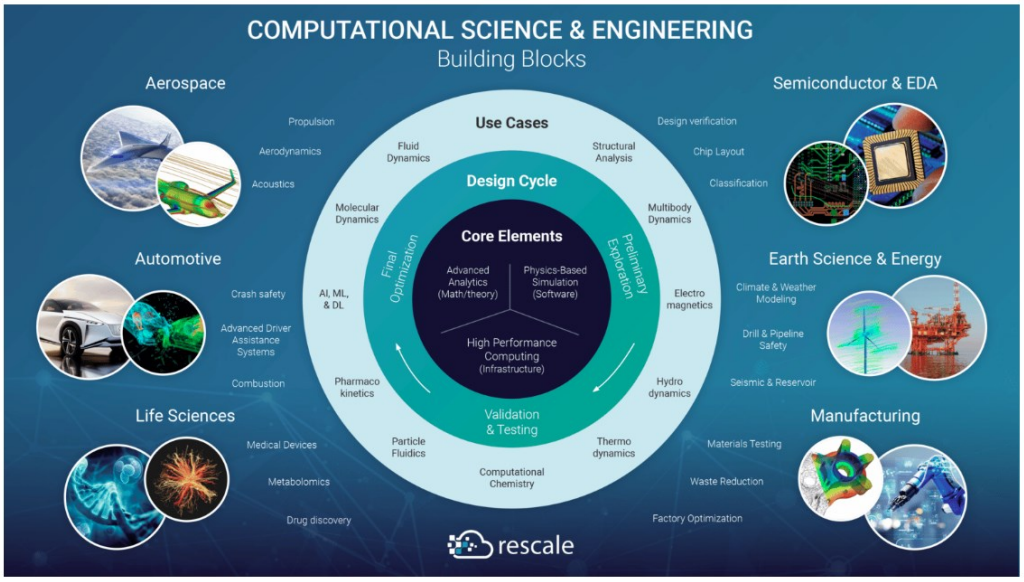

Figure 1. Various Applications of HPC [8]

Because of its agility and flexibility in deployment and management, cloud HPC has transformed how businesses supply assets for advancement activities. Cloud HPC can be acquired and deployed faster than on-premises supercomputers, HPC clusters, as well as workstations, and it can be quickly adjusted to meet the demands of the organisation at any time.

Advantages of HPC in the Cloud

Distribution of workload: Cloud computing allows you to employ containerised microservices, which can aid in the seamless orchestration of task distribution. Containers can also make it easier to migrate current workloads and tools between platforms.

Availability: HPC in the cloud provides high availability and reduces the likelihood of your workload getting interrupted.

2. Working of HPC on Big Data

High-Performance Computing (HPC) along with Big Data are complementary technologies that frequently work together to solve complicated computing problems related with processing and analysing large datasets. HPC and Big Data combine the advantages of parallel cutting, high-speed connects, as well as distributed computing in order to effectively handle the large-scale data needs of today's applications.

Big data processing requires improved performance and speed, which HPC gives. It reduces workload run time, lowering total expenses while improving productivity. It also mitigates the dangers involved with unsuccessful experiments. Because an experiment fails in shorter time, it is less expensive to conduct trials or creative tests.

Processing of Big Data based on HPC

- HPC systems excel at managing enormous amounts of data on a continual basis thanks to their parallel processing capabilities.
- Automated batch processing enables the real-time summarising and classification of data that comes in streams. HPC clusters can successfully parallelize processes, allowing for quick data processing and the creation of summaries or classifications as new data becomes accessible.
- Sending data streams directly to the cloud allows for distributed processing, enabling HPC systems to efficiently handle large-scale data and respond to dynamic workloads.

- HPC's capacity for handling streams in real time is important. Automated signal detection and response techniques can be used to recognise trends, anomalies, or specific occurrences quickly, allowing for faster decision-making and reaction.

- This abundance of resources enables the efficient development of a wide range of data visualizations, allowing decision-makers and analysts to acquire more insight.

Challenges of HPC in Big Data and Cloud Computing

- To optimise output from installing cloud HPC, big data analysts must track utilisation and workloads simultaneously on-premise and in the cloud, and analyse past workload trends, scheduling, and cloud bursting policies.

- HPC applications along with Big Data frameworks frequently employ diverse programming paradigms and software stacks. Integrating these innovations seamlessly can be difficult.

- Despite their impressive powers, supercomputers have a serious challenge: a very short lifespan. In fact, the lifespan for a supercomputer is lower than that of traditional consumer devices such as desktop computers.

- Big Data applications frequently include remote storage and processing, which increases the risk of data corruption. Implementing methods for data integrity becomes a major problem.

- Big Data apps create and process large volumes of data. It might be difficult to efficiently move this data between storage and computation resources in an HPC setting.

3. Real-time Applications

Augmented Reality and Virtual Reality: Augmented Reality (AR) and Virtual Reality (VR) can greatly benefit from the integration of High-Performance Computing (HPC) in the context of Big Data. While the combination of AR, VR, and HPC for Big Data applications has enormous promise, it also presents obstacles such as optimising algorithms for parallel processing, assuring low-latency interactions, and dealing with the complexities of handling

massive datasets in immersive settings. The CAVE is frequently used to visualise information related to fluid dynamics, structural mechanics, architectural modelling, and media arts, providing a look into the potential of working in a virtual reality environment.

Engineering: Engineering is all about improving a machine's real-world performance, yet testing prototypes is costly and sometimes risky. To get around this, engineers frequently test new ideas in huge computer simulations. High-speed computing is helping to transform several parts of the automobile industry, form crash test simulation to self-driving cars.

Eg: Autonomous vehicles, fuel conservation, lighter aircraft.

Healthcare: Healthcare professionals employ HPC for a variety of reasons, including enhancing screening procedures, producing more accurate patient diagnoses, as well as expediting administrative tasks. Medicine and computers are as inextricably linked as DNA's double helix. Computers currently keep personal medical data, monitor vital signs, and assess treatment efficacy.

Eg: Cancer Diagnosis, Cardiovascular Issues,

Financial services: High-performance computing is extremely useful for the financial services business. This cutting-edge technology is used for a variety of purposes, like automated trading, fraud detection, and tracking real-time stock movements.

Eg: Fraud detection, cryptocurrency.

Meteorology: High-performance computer greatly enhances weather pattern predictions and tracking. It improves climate model accuracy, allowing for more exact predictions of the weather. Furthermore, HPC helps better to understand the origins and implications of diverse meteorological occurrences.

Eg: Tornado Visualisation, Solar Weather Monitoring

4. Future of HPC in Big Data Application

Discussing current HPC advances without addressing artificial intelligence (AI) would be difficult. In the last decade with IoT, 5G, and various other data-driven technologies, the quantity of data accessible for significant, life-changing AI has increased sufficiently for AI to

have an influence on HPC, and vice versa. Companies can put their HPC data centre on-premises, in the cloud, at the "edge" (however that term is defined), or a mix of the above. The volume of data produced, along with the need for quicker and more accurate analytics, makes big data and HPC a perfect combination. Until a more affordable or efficient solution is created, big data as well as HPC are going to become increasingly coupled. One future successor for this pair might be quantum computing, however, this technology has yet to become widely available.

5. Conclusion

High-performance computing (HPC) for Big Data are critical technologies for progress in research, business, and industry. It is expected to remain a crucial technology, with an increasing need for computer capacity. HPC's capabilities provide powerful answers to the difficulties of data transportation, storage, scalability, and various workloads, enabling for parallel processing, optimised resource utilisation, and effective handling of mixed tasks. As requests for more rapid, effective, and scalable data processing advances, HPC emerges as a critical technology for realizing the full promise of Big Data applications.

References

[1] https://www.geekboots.com/story/parallel-computing-and-its-advantage-and-disadvantage

[2] https://www.purestorage.com/it/knowledge/what-is-an-hpc-cluster.html

[3] Wu, Y., Xiang, Y., Ge, J., & Muller, P. (2018). High-Performance Computing for Big Data Processing. Future Generation Computer Systems, 88, 693–695. doi:10.1016/j.future.2018.07.054

[4] Robey, Robert, and Yuliana Zamora. *Parallel and high performance computing*. Simon and Schuster, 2021.

[5] https://www.hpe.com/in/en/what-is/high-performance-computing.html

[6] Yin, Fei, and Feng Shi. "A Comparative Survey of Big Data Computing and HPC: From a Parallel Programming Model to a Cluster Architecture." International Journal of Parallel Programming 50, no. 1 (2022): 27-64.

[7] LIAO, Wei-keng, and Alok CHOUDHARY. "High performance big data clustering." Cloud Computing and Big Data 23 (2013): 192.

[8] Bautista Villalpando, Luis Eduardo, Alain April, and Alain Abran. "Performance analysis model for big data applications in cloud computing." *Journal of Cloud Computing* 3, no. 1 (2014): 1-20.

[9] Abid, Mohamed Riduan. "HPC (high-performance the computing) for big data on cloud: Opportunities and challenges." *International Journal of Computer Theory and Engineering* 8, no. 5 (2016): 423.

[10] Al Shehri, Waleed, Maher Khemakhem, F. Eassa, A. Basuhail, and F. E. Eassa. "Evaluation of high-performance computing techniques for big data applications." *Science International* 31, no. 1 (2019): 149-163.

[11] Navaux, Philippe Olivier Alexandre, Arthur Francisco Lorenzon, and Matheus da Silva Serpa. "Challenges in high-performance computing." Journal of the Brazilian Computer Society 29, no. 1 (2023): 51-62.

[12] https://rescale.com/blog/dispelling-the-top-5-myths-of-cloud-hpc-introduction-myth-1-hpc-is-a-niche-application/

[13] Usman, Sardar, Rashid Mehmood, and Iyad Katib. "Big data and hpc convergence for smart infrastructures: A review and proposed architecture." *Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies* (2020): 561-586.

[14] Xenopoulos, Peter, Jamison Daniel, Michael Matheson, and Sreenivas Sukumar. "Big data analytics on HPC architectures: Performance and cost." In 2016 IEEE International Conference on Big Data (Big Data), pp. 2286-2295. IEEE, 2016.