Chapter 9
Cyberinfrastructure for Advanced Research with High Performance Edge Computing

Dharm Singh Jat, Arpit Jain, and Kamwayamunzi Fortune Sepo

Abstract The technological revolution brings rapid change in scientific and computational approaches. A huge amount of data is generated due to concurrent request-response processes and computation on real-time data. And this data requires to store as historical data for future reference and data analysis. For the computational work with intelligent and dynamic algorithm processes in communication networks, the infrastructure requires many scientific and networking equipment such as High Performance Computing (HPC) cyberinfrastructure. Researchers working with the complex problem need a small, state-of-the-art HPC system for their research. Researchers also require HPC administration expertise and identify and install the required tools, system software. Most of the time, the researcher would install the required tools and software that will be expensive. Undoubtedly, there is a need for a fast and low-cost ready-to-use HPC system that can be straightway put to utilisation by researchers and users. This paper presents a comprehensive literature review about the high performance edge computing (HPEC) technologies of cyberinfrastructure and other existing related initiatives around the world. Further, the paper also presents a case study of an affordable supercomputing solution named PARAM !ARUB which offers ready-to-use supercomputing facility based on edge computing and AI technologies hardware resources for complex problems. This provides a support research tool for analysis, design and development.

Keywords HPC · Edge computing · High performance edge computing · HPEC

D. S. Jat · A. Jain (✉) · K. F. Sepo
Namibia University of Science and Technology, Windhoek, Namibia
e-mail: ajain@nust.na

D. S. Jat
e-mail: dsingh@nust.na

R. Mathur et al. (eds.), *Emerging Trends in Data Driven Computing and Communications*,
Studies in Autonomic, Data-driven and Industrial Computing,
https://doi.org/10.1007/978-981-16-3915-9_9

1 Introduction

High Performance Computing (HPC) Systems are usually extensive systems that require big space and specialised infrastructure. Due to several reasons HPC systems are unaffordable and inaccessible for researchers when needed for short duration [1]. The HPC will strengthen the application development of packages in smart, secure environments such as Climatology, Bioinformatics, Agriculture, Preventive health care system, Astrophysics and Transportation systems required in smart cities. This will accelerate research and development (R&D) in the war against global diseases such as COVID-19 through faster simulations, medical imaging and forecasting.

High Performance Computing (HPC) can process data and perform complex calculations at extremely high speeds [2]. The HPC is to combine multiple processors to create a unified system that can process a tremendous amount of work within a short period. To cope up with the fourth industrial revolution, nations must build infrastructures that will handle the big data associated with the resulting technologies. However, for most developing countries, the HPC facilities that support artificial intelligence and big data analytics are limited or non-existent. The lack of HPC facilities poses a severe problem in the research fraternity and the entire community. As we approach the era of the Internet of Things (IoT), more and more data will be generated daily. Figure 1 shows a prediction by the international telecommunication union (2015) for the expected data traffic in exabytes per month from 2020 to 2030 [3].

The figure shows the traffic will be growing at an annual rate of around 55% in 2020–2030. The global traffic per month is estimated to reach 607 exabytes (EB) in 2025 and 5016 EB in 2030. Thus, the major challenges are the facilities that will store and process such a large traffic volume. It is required to build widespread available, affordable and accessible communication networks with enhanced hardware HPC

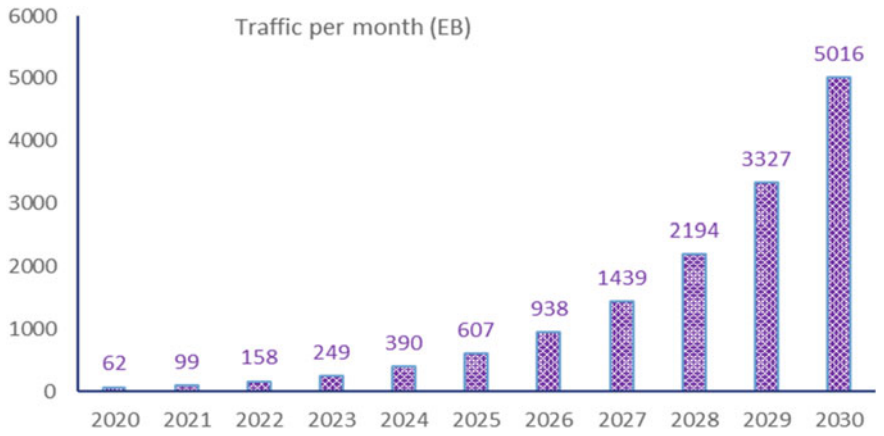

Fig. 1 Estimates of global traffic from the year 2020–2030 [3]

infrastructure to meet edge's current demand for intelligent services. High Performance computing (HPC) takes on challenges that will not fit in a single desktop/server configuration or be solved by a simple system in an acceptable time frame. In the HPC setup, multiple computers known as cores are clustered together to function in unison as a single high processing machine. This setup is highly needed for a researcher who collects and processes big data that requires large storage and rapid processing power.

There is multiple software that makes an HPC run effectively, efficiently and securely. HPC software's can be broken into system software and application software. The system software is used to build and manages the HPC system, while the application software supports the various application running on the HPC system. Mostly HPC infrastructure are developed to serve multiple and concurrent users in a cloud platform, thus there are unique softwares required for each set of users.

The general profile of HPC applications is constituted by an extensive collection of compute-intensive tasks that need to be processed in a short period of time. Historically, HPC was only available and used by key institutions such as scientific laboratories, university research and the military. Today, HPC is used to solve a range of problems, including applications in Bioinformatics, Climatology, Astrophysics, Computational Fluid dynamics, Computational Chemistry, Molecular Dynamics, Finite Element Analysis, Agriculture, Big data analytics, Artificial Intelligence, Internet of Things, etc. HPC applications are always provisioned using cloud computing.

2 Related Works

In this study, autotuning tools are emphasised to facilitate High performance computing, which helps the programmer write a set of codes efficiently without putting many efforts. It eliminates the manual efforts in the implementation of low-level architecture, which results in programmer productivity. The study analyses the powerful features of autotuning such as forecasting, predictability and integrational aspect into the applications running in HPC. The autotuning tools are much suitable for heavy computational aspects like deep learning, data exploration, data visualisation and data analytics for a huge number of datasets [3].

As the concurrent processes and computational overhead are executing in High performance computing components like clusters, base nodes, master nodes and clouds, where complex problem-solving and decision-making algorithms consumes much energy in continuous or longer execution. To optimise the energy consumption, the study proposed a novel energy scheduling algorithm based on the concept of maximum execution time with respect to minimum energy consumption. In this study, the optimised scheduling algorithm is tested with statistical analysis of variance followed by post hoc analysis to verify the energy efficiency and effectiveness [4].

The study analyses the challenges and techniques to overcome these issues in communication. It also introduces the efficient technique in edge AI communication system with an intelligent algorithm and inferences computational task at edge network. Further, the study presents an efficient solution in edge AI applications to train the AI models at the edge layer with several order optimisation algorithms. It also summarises the types of edge-based AI model architecture includes data partition and model division training approach. The study concludes with an analysis of computational latency, offloading and inference at edge node with an AI edge framework [5]. In this study, the intelligent system, coupled with edge computing, is analysed in the cyber human evolution and concludes with the Edge AI environment's challenges. The study analyses Edge Intelligence's fabric with critical components like sensing substrate, edge network with AI functionalities complimented by HPC. This study is also focussed on the investigation gaps in edge intelligence infrastructure of smart-city scale. Also shown are the negative aspects due to complexity in AI applications and machine learning processes for which new efficient techniques are required [6].

In this study, the singularity container-based technology is investigated to analyse the CPU's performance, memory and bandwidth of the network. Through this container-based technology, user and HPC data centres have the flexibility to utilise and distribute software environment. According to the given analysis, the study shows compatibility, mobility and security in the computing environment [7].

The study shows the vulnerability and security challenges in High Performance Computing systems, which have increased due to advanced integration and computing techniques. Further, the study employs three-dimensional integrated circuits with case studies to analyse the latest security threats. The study also highlighted the security measure to cover up the loopholes and provide approximations on computing. 3D integration plays a vital role in HPC with the added advantage of high device density, low power consumption and high bandwidth. The study shows the security threats on the hardware like 3D chips, specifically hardware trojans, with many other types of attacks [8].

The study proposed a framework named Edgent for Deep Neural Network (DNN) inferences with Edge Computing. To optimise the DNN inference latency, the study introduced two forms of design. One is the partition between device and edge computing. Another is the right-sizing design to describe the exit mechanism. The study designed collaboration aspects with different modes of networks. The fluctuation in bandwidth can be stable using a regression-based forecasting or prediction model known as a static network environment. The bandwidth changes rapidly, called dynamic. Further, the study implements to achieve the low latency edge intelligence with the help of a prototype developed in Raspberry pi [9].

The study shows the integration of blockchain technology with edge computing and implemented in blockchain-based architecture, which keeps track of the users accessing the data stored in distributed databases in the form of data analysis. This data analysis is concluded with edge artificial intelligence using the Ethereum blockchain technology. Further, the study emphasises the secure database and

distributed trust with AI at the edge with the reduction in resource consumption [10].

The study analyses the limitations and advantages of working with edge architectures compared to cloud computing to execute AI algorithms or applications. With the help of a hardware edge accelerator, the edge-based AI workload is tested with benefit analysis of distributed or split processing, including model splitting and model compression. Split processing enables the deep learning process to be split across multiple nodes, whereas compression is the alternative method that provides the smaller compressed form of split processing with the lowest memory and least resources. Further, the study concludes with two analyses: firstly the edge accelerator can serve the concurrency of multitenant applications and secondly the disadvantage of isolation mechanism required for the edge computing environment [11].

3 HPC Edge and AI

The study shows the deployment of edge-based AI applications characterised by face recognition video analytics using machine learning and open-source infrastructure. The application is built using machine learning and artificial intelligence algorithm with many accelerators, which has limited software infrastructure. Further, the study shows the pre and post-processing code integration with multiple inferences stage. It reveals the system level implications on AI application which expose AI tax for overall computing for CPU performance. The study concludes with the analysis of upcoming challenges of accelerated AI [12].

Figure 2 shows the edge computing environment in which the High Performance Computing device interacts with edge devices and cloud computing. The HPC edge node is equipped with the data analysis machine learning algorithm, through which the real-time decision-making and resource management task can be achieved. The collaboration of edge computing and AI concludes the term edge AI, which enhances the system by learning the intelligence with data and programs. The various benefits of edge computing like optimised bandwidth and low latency make the added advantage to Edge nodes' designed algorithm. With the help of many datasets, AI can take out the inferences and make decisions according to the business requirements.

Due to the high demand for applications using IoT devices, the cloud extension in the form of edge computing can serve with AI components. The real-time data is used for predictive analysis from the sensors and IoT devices through this computing enhancement. To ensure the quality of Edge AI model inference results, it is required to evaluate the performance with such parameters as communication delay or latency, privacy and integrity of data, the accuracy of the transaction and communication overhead. The advancement in technology results in benefits and some challenges associated with the current system [13].

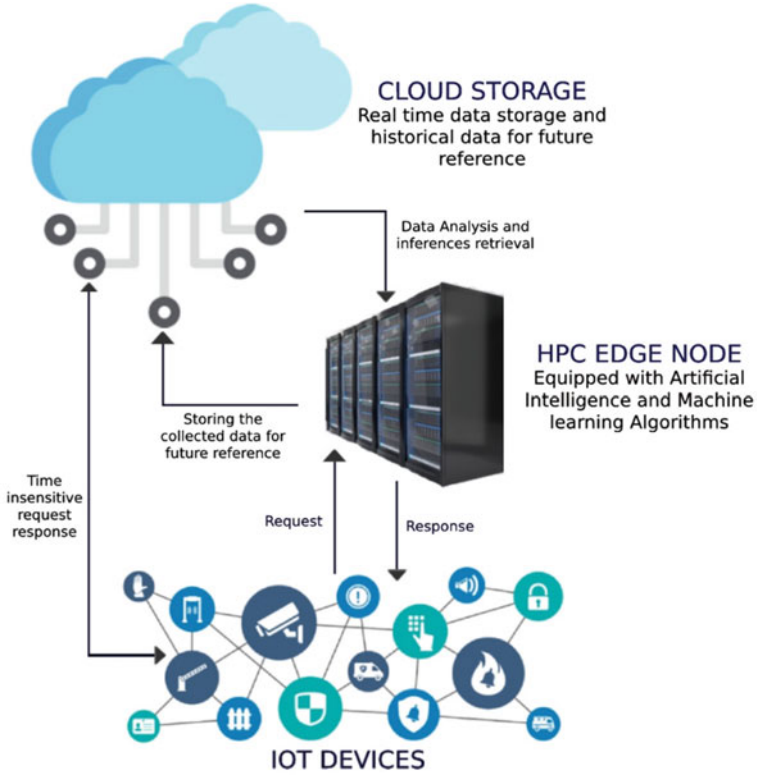

Fig. 2 HPC edge computing environment

4 PARAM !ARUB Architecture

An affordable supercomputing solution for a complex problem named PARAM !ARUB [14], which offers ready-to-use supercomputing facilities based on edge computing and AI technologies hardware resources is hosted in the Namibia University of Science and Technology (NUST). This provides a support research tool for analysis, design and development. PARAM !arub system is based on processor Intel Xeon Gold 6130. The cluster consists of compute nodes connected with INTEL OPA 100Gbps interconnect. The PARAM !ARUB Supercomputer aims to provide a computational resource in Namibia with edge artificial intelligence and other advanced technologies to perform complex tasks for industry, academic, scientific, technology, engineering programmes. PARAM !ARUB high performance edge computing (HPEC) is an affordable supercomputer facility with pre-installed required software, and the various applications from selected engineering and scientific domains are ready-to-use. Edge and artificial intelligence (AI) based PARAM !ARUB HPC system is designed at an affordable cost to perform complex and

high-end computations for the scientific, technology, engineering and academic programmes to solve the complex problem by using modelling, simulation and data analysis.

Most researchers working with complex problem need a small state-of-the-art HPC system for their research. Researchers also require HPC administration expertise and identify and install the required tools, system software. Most of the time, the researcher would consume time to install the required tools and software and be expensive. Undoubtedly, there is a need for a fast and low-cost ready-to-use HPC system that can be straightway put to utilisation by researchers and users. The PARAM !ARUB supercomputing solution provides scalability and power efficiency. This paper presents a comprehensive literature review about the HPEC and other existing related initiatives around the world.

4.1 Machine Learning/Deep Learning Application Development

Most of the popular python based machine learning/deep learning libraries are installed on PARAM !arub system. While developing and testing their applications, users have the option to choose virtual environment-based python libraries or conda runtime based python libraries [14] (Fig. 3).

Fig. 3 PARAM !ARUB architecture at NUST, Namibia [15]

5 Conclusion

Today, a huge amount of data is generated from the Internet and other various sources. For the experimental work in communication networks research, the experimental setup requires many scientific and networking equipment such as high performance computing cyberinfrastructure. These equipments are normally geographically scattered with different capabilities around the nation, region and globe. For most of the researchers, this experimental setup is unavailable or hardly accessible. This paper explores how to enable wider accessibility of the HPC platform to potential researchers and learners and provide an affordable HPEC system named PARAM !ARUB.

Acknowledgements The authors would like to acknowledge the Centre for Development of Advanced Computing (C-DAC), Ministry of Electronics & Information Technology, Government of India, particularly the HPEC-Technologies Group, who have contributed to the development.

References

1. Agrawal S, Das S, Valmiki M, Wandhekar S, Moona R (2017) A case for PARAM shavak: ready-to-use and affordable supercomputing solution. In: 2017 international conference on high performance computing & simulation (HPCS), Genoa, 2017, pp 396–401. https://doi.org/10.1109/HPCS.2017.66. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8035105
2. Cappello F, Gentzsch W, Valero M, Nygård M (2013) HPCS 2013 panel: the era of exascale sciences: challenges, needs and requirements. In: 2013 international conference on high performance computing & simulation (HPCS), Helsinki, Finland, 2013, pp 1–12. https://doi.org/10.1109/HPCSim.2013.6641380. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6641380
3. ITU 2015.IMT traffic estimates for the years 2020–2030. https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2370-2015-PDF-E.pdf
4. Orris BON, Uduc RIV, IEEE M (2018) Autotuning in high- performance computing applications 106(11)
5. Biswas T, Kuila P, Ray AK (2021) A novel energy efficient scheduling for high performance computing systems
6. Shi Y, Yang K, Member GS (2020) Communication-efficient edge AI: algorithms and systems 22(4):2167–2191
7. Rausch T (2019) Edge intelligence: the convergence of humans, things, and AI.https://doi.org/10.1109/IC2E.2019.00022
8. Wkh S, Ri S, Ljk IRU (2019) ([sorulqj wkh shuirupdqfh ri 6lqjxodulw\ iru +ljk 3huirupdqfh &rpsxwlqj 6fhqdulrv. https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00362
9. Yellu P, Zhang Z, Mezanur M, Monjur R, Abeysinghe R, Yu Q (2019) Emerging applications of 3D integration and approximate computing in high-performance computing systems: unique security vulnerabilities.
10. Li E, Zeng L, Zhou Z, Chen X (2020) Edge AI: on-demand accelerating deep neural network inference via edge computing 19(1):447–457
11. Nawaz A, Gia TN, Pe J (n.d.) Edge AI and blockchain for privacy-critical and data-sensitive applications, 2–3
12. Liang Q, Irwin D (2020) AI on the edge: characterizing AI-based IoT applications using specialized edge architectures, 145–156. https://doi.org/10.1109/IISWC50251.2020.00023

13. Richins D, Doshi D, Blackmore M, Nair AT, Pathapati N, Patel A (2020) Missing the forest for the trees: end-to-end AI application performance in edge data centers, 515–528. https://doi.org/10.1109/HPCA47549.2020.00049
14. Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J (2019) Edge intelligence: paving the last mile of artificial intelligence with edge computing, 107(8)
15. Centre for Development of Advanced Computing (C-DAC) (2021) PARAM !Arub—user guide