*International Journal of Engineering Applied Sciences and Technology, 2019*
**Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 168-170**
Published Online December 2019 in IJEAST (http://www.ijeast.com)

# STRUCTURAL ANALYSIS OF HPC'S FOR BIG DATA ANALYTICS

Srivatsa Raju S
Department of Computer Science and Engineering
B.M.S Institute of Technology & Management,
Bengaluru, Karnataka, India

Dr. Anjan K Koundinya
Department of Computer Science and Engineering
B.M.S Institute of Technology & Management
Bengaluru, Karnataka, India

***Abstract*— Data analytics possess challenges in various scenarios, designing a system is one of them. This paper provides an insightful preview of the challenges and possible outcome one needs to consider. The High-Performance Computing system for analysis of complex computer-intensive data needs to be focused on architecture design, network hardware, clustering, I/O systems, Metadata, cost and how this analysis can be used in Machine Learning and Big Data analytics techniques.**

*Keywords*— **High-Performance computing, Architecture, Network Hardware, Clustering, I/O systems, Machine Learning.**

## I. INTRODUCTION

This paper sheds light about how HPC is involved in the generation of huge data in every second and scaling data centers according to Moore's law, where the performance and functionality of digital informatics grows twice the present every 2 years with cost, energy requirements and area. This paper also shares insights about possible technologies which can be utilized for better performance characteristics for Big data analytics with architecture, cost-efficient network, clusters, I/O variability, Metadata Management, Platform selection, resource management for HPC and Machine Learning point of view.

## II. HPC'S ARCHITECTURE SCALABILITY

As the size of information and data grows on the internet exponentially the existing infrastructures take a toll on managing this data specially the emerging accelerating technologies for the field of Machine learning the workload push the limits of the computing infrastructure for a better bandwidth and energy efficient technologies for the upscales.

The Moore's law where the growth doubles for every 2 years are transforming the existing architectures for the ready to use networking and storage devices which are not anymore a statistically fulfilling solution to respond the needs of being scalable and dynamically adaptable memory fabrics that uses the classics of photonics which can shape the future of vast datacentre and HPC architectures. [1]

On the basis of such considerations, the algorithm uses a different color image multiplied by the weighting coefficients of different ways to solve the visual distortion, and by embedding the watermark, wavelet coefficients of many ways, enhance the robustness of the watermark.

## III. COST-EFFECTIVE ROUTING FOR HPC NETWORKS

With the scaled-out data centers and HPC which has large volume of nodes it is critical to have efficient network of distributed computing and memory resources, these kinds of network require high computation bandwidth, capacity, and internetwork parallelism, which makes a challenge to meet cost efficient, energy efficient, and reliability.

Building these networks with higher radix switches, where the signal rates are costlier than the packets this technology is cost efficient than the lower radix routers with low bisection bandwidth and path diversity. The Multiport Binding Tile-based Router (MBTR) proves to be effective and best alternative to off the shelf routers. [2]

## IV. CONTAINER CLUSTERS FOR HPC'S

In the adaptive environment the cloud technologies support the microservices-based applications for scalability, dynamic and manageability by container concept for workloads with resources infrastructure High-Performance Data Analytics (HPDA) to process higher volumes of data generated from different applications. The utilization of SmartX Intelligence Cluster for running containerized HPDA workloads which can provide Hyper-converged style resources with integrated network support, storage and computing.

With the SmartX Intelligence the usage of parallel file system with high-performance within the work node has increase in the performance with the best integration of software for deep learning workloads. [3]

## V. I/O VARIABILITY FOR HPC STORAGE SYSTEM

With the storage devices shared between numerous applications and managed in the best way, the I/O is often a major concern to be addressed which can be resolved by implementing messaging-based re-routing together with throttling at mid-level which can solve QoS-less HPC storage system and runtime scheduling that can be scalable.

The I/O re-routing stabilizes the balance by re-directing traffic to less congested devices which majorly happens at the messaging layer of storage process. [4]

The system logs are useful resources to analyse the system behavior at various level of the stack, the individual analysis can lead to in-consistent results due to limited information of each stack logging at multiple layer is helpful for detection of anomalous activity, this activities can be detected by capturing patterns from non-linear log data, heterogeneous log data and high-dimensional log data by using Luster Monitoring Tool(LMT) to log I/O activities at each node while the NetFlow tool logs transfer activities on which the learning based techniques can be applied to label the analysis with the presence of human expert to employ Machine Learning technologies.[5]

## VI. METADATA MANAGEMENT IN HPC'S

The large volume of metadata consisting of entities such as files, jobs and users where the existing systems can manage these using POSIX data of the files which are critical for the support of advance data management functions such as identifying data sources and identifying parameters of a result, auditing usage data, analysing the transformation of inputs to outputs.

The metadata are heterogeneous which the attributes associated with them makes it complicated for individuals to understand, these data are obtained from various sources and distinct formats which needs to be integrated uniformly to avoid redundancy across management tool, with the help of highly efficient query language to perform data management tasks graph-based metadata management systems are implemented called GraphMeta for HPC's.

The main functionality of GraphMeta is to unify all metadata into heterogeneous graph property and support an outcome infrastructure for managing this graph to meet scaling and performance requirements of the HPC's. [6]

## VII. COST COMPARISON BETWEEN EC2 AND HPC'S

The organization take advantage of the High-Performance Computing (HPC) resources to visualize, analyse and model the growing data volume to growing market. The HPC are innovative and competitive essential for innovation but the Total Cost of Ownership (TCO) in HPC are too high and manpower and skills are required to operate these systems. But with infrastructure as a service (IAAS) in the cloud computing are the best suited for the HPC workload, which attracts more attention due to attractive advertisement for cost-effective approach foreseen in IAAS where the performance bottleneck is nonnegligible due to hypervisor at every middleware in cloud infrastructure.

The statics-based comparison is provided in period of year 2014 with the usage of logs form the batch scheduler where the nodes were allocated to jobs based on the prices. The actual performance of the jobs capable of calculating the annual cost of operating the EC2 and in-house cluster in almost equivalent with the objective to evaluate cost of cloud-computing (CC) and in-house facility with cutting-edge HPC technologies (Direct Liquid Cooling system, InfiniBand EDR interconnect and so on.). [7]

## VIII. HPC'S FOR BIG DATA ANALYTICS

The HPC infrastructure attracts ways to improve the performance but the collocation of HPC and Big-Data is not easy because of the differences in concepts. The HPC job rigidity create holes in batch scheduler we can use these idle resources as dynamic adaptability for Big-Data workload with the help of Resource and Job Management System's (RJMS) configured to communicate with both Big-Data Systems and HPC's using prolog techniques.

Using the RJMS middle-ware OAR and Hadoop YARN from HPC and Big-Data ecosystem with Grid5000 platform the experiments shows HPC workload with 69% utilization and Big-data batch scheduler fills holes to reach optimal capacity of 100%. With the mean waiting time as penalty for HPC job below 17% and effectiveness of Big-Data greater than 67% in average.

The BeBiDa approach which is polished for end users and requires configuration of the cluster administrator. [8]

## IX. CONCLUSIONS

This paper discusses about the architecture scalability in account of Moore's law where the HPC system must be dynamically scalable to adapt workload and distribute load by using the Multiport Binding Tile-based Router in the network layer and the utilization of SmartX intelligence clustering for containers, Luster monitoring tool (LMT) and NetFlow tool to obtain QoS-less HPC storage system I/O and GraphMeta to manage large volume of metadata to unify. Infrastructure cost comparisons between EC2 and HPC to analysis the cost to deploy hypervisor nodes and finally with Resource and Job management system to and Hadoop YARN with Grid5000 platform for best utilization of the available resources.

## X. REFERENCES

[1] J. Shalf, "*HPC Interconnects at the End of Moore's Law*," 2019 Optical Fiber Communications Conference and Exhibition (OFC), San Diego, CA, USA, 2019.

[2] Y. Dai, K. Lu, L. Xiao and J. Su, "*A Cost-Efficient Router Architecture for HPC Inter-Connection Networks: Design and Implementation*," in IEEE Transactions on Parallel and

Distributed Systems, vol. 30, no. 4, pp. 738-753, 1 April 2019.

[3] J. Kwon, N. L. Kim, M. Kang and J. WonKim, "*Design and Prototyping of Container-Enabled Cluster for High Performance Data Analytics*," 2019 International Conference on Information Networking (ICOIN), Kuala Lumpur, Malaysia, 2019.

[4] D. Huang et al., "*Can I/O Variability Be Reduced on QoS-Less HPC Storage Systems?*," in IEEE Transactions on Computers, vol. 68, no. 5, pp. 631-645, 1 May 2019.

[5] J. Choi and A. Sim, "*Spatio-Temporal Analysis of HPC I/O and Connection Data*," 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), Vienna, 2018, pp. 1585-1588.

[6] D. Dai, R. B. Ross, P. Carns, D. Kimpe and Y. Chen, "*Using Property Graphs for Rich Metadata Management in HPC Systems*," 2014 9th Parallel Data Storage Workshop, New Orleans, LA, 2014, pp. 7-12.

[7] J. Emeras, S. Varrette and P. Bouvry, "*Amazon Elastic Compute Cloud (EC2) vs. In-House HPC Platform: A Cost Analysis*," 2016 IEEE 9th International Conference on Cloud Computing (CLOUD), San Francisco, CA, 2016, pp. 284-293.

[8] M. Mercier, D. Glesser, Y. Georgiou and O. Richard, "*Big data and HPC collocation: Using HPC idle resources for Big Data analytics*," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 347-352.

[9] A. Gupta, P. Faraboschi, F. Gioachin, L. Kale, R. Kaufmann, B.-S. Lee, V. March, D. Milojicic, C. Suen, "*Evaluating and improving the performance and scheduling of hpc applications in cloud*", dCloud Computing IEEE Transactions on, no. 99, pp. 1-1, 2014.

[10] B. Iordanov, "*HyperGraphDB: A Generalized Graph Database*," in Web-Age Information Management. Springer, 2010, pp. 25-36. Wucherl Yoo, Michelle Koo, Yi Cao, Alex Sim, Peter Nugent, Kesheng Wu, "PATHA: performance analysis tool for HPC applications", 34th IEEE International Performance Computing and Communications Conference IPCCC 2015, pp. 1-8, December 14–16,2015.