Azerbaijan Journal of High Performance Computing, Vol 6, Issue 2, 2023, pp. 163-170
https://doi.org/10.32010/26166127.2023.6.2.163.170

Survey of Usage Artificial Intelligence Mechanism in the Load Balancer

Nigar Ismayilova

Azerbaijan State Oil and Industry University, Baku, Azerbaijan, nigar.ismailova@asoiu.edu.az

*Correspondence: Nigar Ismayilova, Azerbaijan State Oil and Industry University, Baku, Azerbaijan, nigar.ismailova@asoiu.edu.az

Abstract

Nowadays, there is no way to imagine artificial intelligence applications without using high-performance computing systems. The huge amount of processing data, the complex structure of learning technology, time limitations, and the necessity of real-time operation require powerful computational resources and parallel algorithms. This paper analyzed another direction of convergence between high-performance computing and artificial intelligence: using artificial intelligence techniques in one of the main problems of distributed systems load balancing. The primary objective of this work is to examine the necessity of using AI concepts in load balancing and the definition of providing facilities for load balancers.

Keyword: Load Balancer, Convergence of HPC and AI, Dynamic Load Balancer, Task Scheduling, Artificial Intelligence

1. Introduction

By increasing the usage of HPC technologies in various fields of science and industry optimization of load balancers, finding the best assignment between processes and resources has become challenging. The challenges with optimization methods for task scheduling in different types of distributed computing systems are growing with the progress of computational resources and scientific inventions. The optimal distribution of tasks between resources in cluster computing systems with the standing quantity of requests and computational machines based on general network optimization methods such as bipartite matching, minimax criteria, and finite element methods have been successfully applied (Chu, W. C., Yang, D. L., Yu, J. C., & Chung, Y. C., 2001; Shen, C. C., & Tsai, W. H., 1985; Harvey, N. J., Ladner, R. E., Lovász, L., & Tamir, T., 2006). On the other hand, new approaches based on artificial intelligence must be applied to dynamic environments with unstable resources and requests. The goal of a load balancer in cloud computing systems is the optimization of comprehensive computing capacity, which differs from the load balancer's maximization of computing mission in other distributed systems such as cluster computing, grid computing, peer-to-peer computing, and exascale computing (Ramya & Senthilselvi, 2021). With the popularization of modern wireless communications, load balancers in fog computing systems are necessary. The goal of load balancers in such systems is to minimize execution time and energy consumption.

This work investigates different approaches for optimizing load balancers based on

artificial intelligence techniques, indicates the advantages and disadvantages of proposed methods and their applications, and points out challenges for getting better results in optimizing load balancers.

Section 2 analyses the classification of load balancers in computing systems using different parameters, the third section is concerned with different approaches for minimization of execution time and cost based on machine learning algorithms, and the fourth section presents the application of fuzzy logic-based methods for task scheduling, pros and cons of these approaches. Section five has demonstrated using genetic algorithms to find the best solutions in task assignments for distributed computing systems. The paper ends with a discussion and conclusion sections.

2. Classification of Load Balancers

The mechanism of load balancing activity can be modeled by the following formula (Bakhishoff, U., Khaneghah, E. M., Aliev, A. R., & Showkatabadi, A. R., 2020):

Accordingly, this formula describes conditions for optimal load balancing. It is an assignment process of processes to computing machines where each process in the system must be scheduled, and the activity of all resources should be 100 percent.

There are different approaches and parameters for the classification of load balancer strategies. Some 3researchers, as essential parameters, use the location for load balancers and classify them as centralized and distributed (Waraich, 2008) or centralized, decentralized, and hierarchical load balancers (Al-Rayis & Kurdi, 2013). Several methods and algorithms have been proposed for implementing these strategies in multi-scale computing systems, and their priorities were justified by experiments (Barazandeh & Mortazavi). There is also a large body of work considering the classification of load balancers into local and global classes based on the information needed for distributing requests. The superiority of distributed algorithms for load balancers, which gets information from the global state, has been demonstrated over local load balancers, showing unscalable distribution in the computing system (Gasmelseed & Ramar, 2019).

A more general classification of load balancers as static and dynamic has been proposed based on system characteristics. For static load balancers where several available resources and solving requests are persistent were proposed application of famous scheduling algorithms such as Round Robin, least connections approach, and graph matching (Alankar, B., Sharma, G., Kaur, H., Valverde, R., & Chang, V., 2020; Wei, L. F., Ji, J. W., & Zhao, L. Q., 2011; Kaur, M., & Mohana, R., 2019; Devi, R. K., & Muthukannan, M., 2018, October). Some studies proposed a partitioning approach to distributing unmanageable loads between resources in static systems (Meyerhenke, H., Monien, B., & Sauerwald, T., 2009; Sevilla, M. A., et al., 2015, November). All these approaches have their advantages and limitations. The main disadvantage of the round-robin approach is an assumption about equality of resource abilities and the nonexistence of perspectives for AI applications (Aza & Urrea, 2019). Randomized load balancing established on the base of the Poisson process or Markov process application of statistical inference algorithms

and state-based search AI methods will provide efficient results for load balancing optimization (Bramson et al., 2010). Centralized load balancers can improve their activity by applying different clustering methods for classification and learning the processes in the computing system. The machine learning approach allows us to optimize the load balancing in dynamic, decentralized Grid systems by threshold approach (Rathore, N., 2016; Goldsztajn, D., et al., 2022; Lin, W., Wang, J. Z., Liang, C., & Qi, D., 2011; Rathore, N., & Chana, I., 2015). The main problem here, as usual in systems with dynamic structures, is the recognition and classification of the unlabeled data. AI forecasting methods open wide areas for high-performance load-balancing models based on least-connection scheduling algorithms (Choi, D., Chung, K. S., & Shon, J., 2010, December; Ren, X., Lin, R., & Zou, H., 2011, September). State-based techniques of AI, especially statistical inference methods, can effectively be applied to the organization of load balancing in distributed systems by local queue algorithms and central queue algorithms (Sharma, S., Singh, S., & Sharma, M., 2008).

Progress and intensive usage of heterogeneous computing environments such as grid computing, P2P computing, and cloud computing, as well as challenges in Exascale computing systems, demonstrate the necessity and effectiveness of proposed dynamic load balancers for the distribution of load (Chandakanna, V. R., & Vatsavayi, V. K., 2016; Rajavel, R., Somasundaram, T. S., & Govindarajan, K., 2010). Several investigations used artificial intelligence techniques to assign requests to suitable resources where one or both have a dynamic nature (Hongvanthong, S., 2020, May; Nadaph, A., & Maral, V., 2015, February). The following sections have discussed different approaches for optimizing load balancers in computing systems with a dynamic nature based on artificial intelligence methods.

3. Machine Learning Based Methods for Load Balancer

The use of machine learning algorithms for the selection of the most effective load-balancing algorithm in heterogeneous HPC systems with requests demonstrating dynamic and unknown behavior has been discussed by a significant number of authors (Oikawa, C. A. V., Freitas, V., Castro, M., & Pilla, L. L., 2020, March). Researchers attempted to reduce execution time and communication load and guarantee real-time scheduling by application of different types of neural networks, as well as reinforcement learning. Generally, in these methods, optimal load distribution by the load balancer and real-time decision-making are implemented by learning based on information received from a dynamic computing environment and feedback information. Ahmed et al. use classification methods based on different features extracted from the characteristics of scientific applications for load balancing in heterogeneous computing systems. For training and testing a database comprising execution results of parallel applications. This approach reduces execution time, increases the resource utilization ratio, and performs better than other scheduling algorithms. The main limitation of these methods is the impossibility of increasing the database to the desired volume to improve the accuracy of the classification system.

The application of clustering algorithms for load balancer optimization is widely reported in the literature by researchers. The authors used a k-means clustering algorithm to reduce resource costs and execution time. Conceptually similar work implemented by Sun and others, which used improved k-means clustering according to resource attributes information, which helps to reduce the search environment during task scheduling (Sun et al., 2014). Despite the success of the mentioned approach for load balancer optimization in distributed computing environments, it is still limited by the narrow nature of resource attributes. Using fuzzy c-means clustering and applying this method to classify resources and tasks can be considered for more 'soft' task scheduling in heterogeneous computing systems.

Mao and others reported on a new approach based on the Bayesian model for optimization of resource usage in cloud environments. Their work traces the advantages of predicting prior probabilities using posterior information for optimization load balancer's work (Mao et al., 2014). Bayesian model allows the management work of load balancers in heterogeneous systems. On the other hand, the optimization process is realized by considering a few characteristics of the resources and processes. There is a necessity for a detailed analysis of process and resource characteristics for their appropriate classification.

4. Fuzzy Logic-based Approaches for Load Balancer

Fuzzy inference system (FIS) and adaptive neuro-fuzzy inference system are the main methods based on fuzzy sets theory for optimizing the load distribution in dynamic and heterogeneous environments. Computing with fuzzy logic-based methods and the possibility of handling uncertainties using these algorithms allows us to make real-time decisions in diverse computing systems. Using logic-based artificial intelligence techniques ensures clear information about optimizing the load-balancing process, which can easily be improved through regular monitoring. For, Setia and others applied fuzzy logic-based load balancing for parallel master-slave implementations with linguistic variables for input (estimated load and estimated delay in the nodes) and output (takeover capacity of the system) (Setia et al., 2009).

As has been investigated in some works, the application of logic-based techniques of artificial intelligence, such as using fuzzy logic controllers for fault tolerance management in cloud systems, managing traffic in fog computing systems, and stabilizing load among computational resources in multiprocessor systems, gives appropriate solutions, this also has been used ANFIS load balancer for optimization in Big Data. This approach can demonstrate better performance and require less effort if the *parameters of the system can be determined from the learning process. However, it still needs a considerable amount of data about executed applications.

5. Algorithms Based on Evolutionary Algorithms Used for Load Balancer Optimization

One of the most well-known approaches for handling decentralized computing systems

is using different naturally inspired algorithms based on swarm intelligence. Dasgupta and others examined the optimization of load balancers in cloud computing using genetic algorithms. This approach has been selected as the best scheduling by minimizing the execution time of requests (Dasgupta et al., 2013). Different approaches based on genetic algorithms are used for multicriteria scheduling tasks, regularization of traffic in the network, minimization of energy consumption, and another objective of the load balancer.

G. Sivashanmugam and others focused on the problems of load balancer development in cloud computing systems; they have demonstrated that besides the primary goal of the load balancer to assign tasks to resources, we also need problem-solving and computing abilities. For this purpose, authors have proposed a naturally inspired load balancer algorithm entirely based on eagle behavior (Sivashanmugam et al., 2019). Although approaches based on evolutionary algorithms for task scheduling in computing environments demonstrate success in achieving the primary goal, search in the big environment cannot propose an optimal solution. For this purpose, combining logic-based approaches with evolutionary algorithms can give better results.

6. Discussion

As mentioned in the central part of the work, there are many opportunities for improving the task scheduling process in distributed computing systems by using different methods and algorithms of artificial intelligence. However, the application results of these approaches could be more satisfactory. They have various disadvantages that can be eliminated by using different combinations of the mentioned approaches or proposing novel ideas for handling load balancing in computing systems with a dynamic nature drawing on logic-based artificial intelligence techniques.

Conclusion

This study aimed to evaluate different approaches for load balancers in distributed computing systems, especially systems with dynamic resources and requests, using artificial intelligence mechanisms. The methods reported here, with their priorities and shortcomings, give new objectives for improving the task scheduling process by proposing innovative conceptions. More research for real-time decision-making in heterogeneous computing systems using intelligent approaches is needed to optimize the job scheduling process. This study suggests using intelligent methods for minimization of execution time and computing cost in multi-scale computing systems, which will be able to demonstrate good learning performance using a smaller amount of historical information and environmental feedback. For this reason, applying Bayesian inference methods can demonstrate high accuracy by recovering missing data. On the other hand, using soft clustering methods can reduce search space for tackling the load balancer.

References

Al-Rayis, E., & Kurdi, H. (2013). Performance Analysis of Load Balancing Archi-

tectures in Cloud Computing [Proceedings Paper]. Uksim-Amss Seventh European Modelling Symposium on Computer Modelling and Simulation (Ems 2013), 520-524. https://doi.org/10.1109/ems.2013.10

Alankar, B., Sharma, G., Kaur, H., Valverde, R., & Chang, V. (2020). Experimental setup for investigating the efficient load balancing algorithms on virtual cloud. Sensors, 20(24), 7342.

Aza, E. F., & Urrea, J. P. (2019). Implementation of Round-Robin load balancing scheme in a wireless software defined network [Proceedings Paper]. 2019 Ieee Colombian Conference on Communications and Computing (Colcom 2019), 6.

Bakhishoff, U., Khaneghah, E. M., Aliev, A. R., & Showkatabadi, A. R. (2020). DTHMM ExaLB: discrete-time hidden Markov model for load balancing in distributed exascale computing environment. Cogent Engineering, 7(1), 1743404.

Barazandeh, I., & Mortazavi, S. S. (2009, Dec 28-30). Two Hierarchical Dynamic Load Balancing Algorithms in Distributed Systems.International Conference on Computer and Electrical Engineering ICCEE [Second international conference on computer and electrical engineering, vol 1, proceedings]. 2nd International Conference on Computer and Electrical Engineering, Dubai, U ARAB EMIRATES.

Bramson, M., Lu, Y., Prabhakar, B., & Acm. (2010). Randomized Load Balancing with General Service Time Distributions [Proceedings Paper]. Sigmetrics 2010: Proceedings of the 2010 Acm Sigmetrics International Conference on Measurement and Modeling of Computer Systems, 38(1), 275-286.

Chandakanna, V. R., & Vatsavayi, V. K. (2016). A QoS-aware self-correcting observation based load balancer. Journal of Systems and Software, 115, 111-129.

Choi, D., Chung, K. S., & Shon, J. (2010, December). An improvement on the weighted least-connection scheduling algorithm for load balancing in web cluster systems. In International Conference on Grid and Distributed Computing (pp. 127-134). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chu, W. C., Yang, D. L., Yu, J. C., & Chung, Y. C. (2001). UMPAL: an unstructured mesh partitioner and load balancer on World Wide Web. J. Inf. Sci. Eng., 17(4), 595-614.

Dasgupta, K., Mandal, B., Dutta, P., Mondal, J. K., & Dam, S. (2013). A Genetic Algorithm (GA) based Load Balancing Strategy for Cloud Computing [Proceedings Paper]. First International Conference on Computational Intelligence: Modeling Techniques and Applications (Cimta) 2013, 10, 340-347. https://doi.org/10.1016/j.protcy.2013.12.369

Devi, R. K., & Muthukannan, M. (2018, October). Mobile Agent-based Secure Cloud Data Center Exploration for Load Data Retrieval Using Graph Theory. In Proceedings of the 2018 International Conference on Cloud Computing and Internet of Things (pp. 1-6).

Gasmelseed, H., & Ramar, R. (2019). Traffic pattern-based load-balancing algorithm in software-defined network using distributed controllers [Article]. International

Journal of Communication Systems, 32(17), 14, Article e3841. https://doi.org/10.1002/dac.3841

Goldsztajn, D., et al. (2022). Self-learning threshold-based load balancing. INFORMS Journal on Computing, 34(1), 39-54.

Harvey, N. J., Ladner, R. E., Lovász, L., & Tamir, T. (2006). Semi-matchings for bipartite graphs and load balancing. Journal of Algorithms, 59(1), 53-78.

Hongvanthong, S. (2020, May). Novel four-layered software defined 5g architecture for ai-based load balancing and qos provisioning. In 2020 5th International Conference on Computer and Communication Systems (ICCCS) (pp. 859-863). IEEE.

Kaur, M., & Mohana, R. (2019). Static load balancing technique for geographically partitioned public cloud. Scalable Computing: Practice and Experience, 20(2), 299-316.

Lin, W., Wang, J. Z., Liang, C., & Qi, D. (2011). A threshold-based dynamic resource allocation scheme for cloud computing. Procedia Engineering, 23, 695-703.

Mao, H. Y., Yuan, L., & Qi, Z. W. (2014). A Load Balancing and Overload Controlling Architecture in Clouding Computing [Proceedings Paper]. 2014 Ieee 17th International Conference on Computational Science and Engineering (Cse), 1589-1594. https://doi.org/10.1109/cse.2014.293

Meyerhenke, H., Monien, B., & Sauerwald, T. (2009). A new diffusion-based multilevel algorithm for computing graph partitions. Journal of Parallel and Distributed Computing, 69(9), 750-761.

Nadaph, A., & Maral, V. (2015, February). Methodical analysis of various balancer conditions on public cloud division. In 2015 International Conference on Computing Communication Control and Automation (pp. 40-46). IEEE.

Oikawa, C. A. V., Freitas, V., Castro, M., & Pilla, L. L. (2020, March). Adaptive load balancing based on machine learning for iterative parallel applications. In 2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP) (pp. 94-101). IEEE.

Rajavel, R., Somasundaram, T. S., & Govindarajan, K. (2010). Dynamic load balancer algorithm for the computational grid environment. In Information and Communication Technologies: International Conference, ICT 2010, Kochi, Kerala, India, September 7-9, 2010. Proceedings (pp. 223-227). Springer Berlin Heidelberg.

Ramya, K., & Senthilselvi, A. (2021). Performance Improvement in Cloud Computing Environment by Load Balancing-A Comprehensive Review. Revista Geintec-Gestao Inovacao E Tecnologias, 11(2), 1386-1399.

Rathore, N. (2016). Dynamic threshold based load balancing algorithms. Wireless Personal Communications, 91(1), 151-185.

Rathore, N., & Chana, I. (2015). Variable threshold-based hierarchical load balancing technique in Grid. Engineering with computers, 31, 597-615.

Ren, X., Lin, R., & Zou, H. (2011, September). A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast. In 2011 IEEE

international conference on cloud computing and intelligence systems (pp. 220-224). IEEE.

Setia, A., Swarup, V. M., Kumar, S., Singh, L., & Ieee. (2009). A Novel Adaptive Fuzzy Load Balancer for Heterogeneous LAM/MPI Clusters Applied to Evolutionary Learning in Neuro-Fuzzy Systems [Proceedings Paper]. 2009 Ieee International Conference on Fuzzy Systems, Vols 1-3, 68-+. https://doi.org/10.1109/fuzzy.2009.5277322

Sevilla, M. A., et al. (2015, November). Mantle: a programmable metadata load balancer for the ceph file system. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-12).

Sharma, S., Singh, S., & Sharma, M. (2008). Performance analysis of load balancing algorithms. International Journal of Civil and Environmental Engineering, 2(2), 367-370.

Shen, C. C., & Tsai, W. H. (1985). A graph matching approach to optimal task assignment in distributed computing systems using a minimax criterion. IEEE Transactions on Computers, 100(3), 197-203.

Sivashanmugam, G., Shantharajah, S. P., & Iyengar, N. (2019). Avian Based Intelligent Algorithm to Provide Zero Tolerance Load Balancer for Cloud Based Computing Platforms [Article]. International Journal of Grid and High Performance Computing, 11(4), 42-67. https://doi.org/10.4018/ijghpc.2019100104

Sun, X. Y., Fu, X. L., Hu, H., & Gui, T. (2014). The Cloud computing tasks scheduling algorithm based on improved K-Means. Applied Science, Materials Science and Information Technologies in Industry, 513-517, 1830-1834. https://doi.org/10.4028/www.scientific.net/AMM.513-517.1830

Waraich, S. S. (2008). Classification of dynamic load balancing strategies in a network of workstations [Proceedings Paper]. Proceedings of the Fifth International Conference on Information Technology: New Generations, 1263-1265.

Wei, L. F., Ji, J. W., & Zhao, L. Q. (2011). The Research and Design of Two-Level Load Balancer Based on Web Server Cluster. Advanced Materials Research, 282, 765-769.

Submitted 21.09.2023
Accepted 15.11.2023