REFERENCES

[1] Aslanpour, M.S., Ghobaei-Arani, M. and Nadjaran Toosi, A. (2017) 'Auto-scaling web applications in clouds: A cost-aware approach', Journal of Network and Computer Applications, 95, pp. 26–41. doi:10.1016/j.jnca.2017.07.012.

[2] Goli, A. et al. (2021) 'A holistic machine learning-based autoscaling approach for Microservice Applications', Proceedings of the 11th International Conference on Cloud Computing and Services Science [Preprint]. doi:10.5220/0010407701900198.

[3] Nikravesh, A.Y., Ajila, S.A. and Lung, C.-H. (2015) 'Towards an autonomic auto-scaling prediction system for cloud resource provisioning', 2015 IEEE/ACM 10th International Symposium on Software Engineering for Adaptive and Self-Managing Systems [Preprint]. doi:10.1109/seams.2015.22.

[4] Khaleq, A.A. and Ra, I. (2021) 'Intelligent autoscaling of microservices in the cloud for real-time applications', IEEE Access, 9, pp. 35464–35476. doi:10.1109/access.2021.3061890.

[5] Lv, J., Wei, M. and Yu, Y. (2019) 'A container scheduling strategy based on machine learning in Microservice architecture', 2019 IEEE International Conference on Services Computing (SCC) [Preprint]. doi:10.1109/scc.2019.00023.

[6] Dragoni, N. et al. (2018) 'Microservices: How to make your application scale', Lecture Notes in Computer Science, pp. 95–104. doi:10.1007/978-3-319-74313-4_8.

[7] Bushong, V. et al. (2021) 'On microservice analysis and Architecture Evolution: A Systematic Mapping Study', Applied Sciences, 11(17), p. 7856. doi:10.3390/app11177856.

[8] Lorido-Botran, T., Miguel-Alonso, J., & Lozano, J. (2014). A review of auto-scaling techniques for elastic applications in cloud environments. Journal of Grid Computing, 12(4), 559-592.

[9] Yu, G., Chen, P. and Zheng, Z. (2019) 'Microscaler: Automatic scaling for microservices with an online learning approach', 2019 IEEE International Conference on Web Services (ICWS) [Preprint]. doi:10.1109icws.2019.00023.

[10] Abdullah, M. et al. (2021) 'Predictive autoscaling of Microservices hosted in Fog Microdata Center', IEEE Systems Journal, 15(1), pp. 1275–1286. doi:10.1109/jsyst.2020.2997518.

[11] Hasan, M.Z. et al. (2012) 'Integrated and Autonomic Cloud Resource Scaling', 2012 IEEE Network Operations and Management Symposium [Preprint]. doi:10.1109/noms.2012.6212070.

[12] Zhao, H. et al. (2019) 'Predictive container auto-scaling for cloud-native applications', 2019 International Conference on Information and Communication Technology Convergence (ICTC) [Preprint]. doi:10.1109/ictc46691.2019.8939932.

[13] Li, Q. et al. (2021) 'Rambo: Resource Allocation for microservices using bayesian optimization', IEEE Computer Architecture Letters, 20(1), pp. 46–49. doi:10.1109/lca.2021.3066142.

[14] Xu, C.-Z., Rao, J. and Bu, X. (2012) 'URL: A unified reinforcement learning approach for autonomic cloud management', Journal of Parallel and Distributed Computing, 72(2), pp. 95–105. doi:10.1016/j.jpdc.2011.10.003.

[15] Xing, J. et al. (2018) 'AsIDPS: Auto-scaling intrusion detection and prevention system for cloud', 2018 25th International Conference on Telecommunications (ICT) [Preprint]. doi:10.1109/ict.2018.8464855.

[16] Wajahat, M. et al. (2016a) 'Using machine learning for black-box autoscaling', 2016 Seventh International Green and Sustainable Computing Conference (IGSC) [Preprint]. doi:10.1109/igcc.2016.7892598.

[17] Qu, C., Calheiros, R.N. and Buyya, R. (2018) 'Auto-scaling web applications in clouds', ACM Computing Surveys, 51(4), pp. 1–33. doi:10.1145/3148149.

[18] Cheng, K. et al. (2023) 'Proscale: Proactive autoscaling for Microservice with time-varying workload at the edge', IEEE Transactions on Parallel and Distributed Systems, 34(4), pp. 1294–1312. doi:10.1109/tpds.2023.3238429.

[19] Marathe, N., Gandhi, A. and Shah, J.M. (2019) 'Docker Swarm and kubernetes in cloud computing environment', 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) [Preprint]. doi:10.1109/icoei.2019.8862654.

[20] Lombardi, F. (2018) 'A proactive Q-learning approach for autoscaling Heterogeneous cloud servers', 2018 14th European Dependable Computing Conference (EDCC) [Preprint]. doi:10.1109/edcc.2018.00038.

[21] Nguyen, T.-T. et al. (2020) 'Horizontal pod autoscaling in Kubernetes for Elastic Container Orchestration', Sensors, 20(16), p. 4621. doi:10.3390/s20164621.

[22] Abdel Khaleq, A. and Ra, I. (2019) 'Agnostic approach for microservices autoscaling in cloud applications', 2019 International Conference on Computational Science and Computational Intelligence (CSCI) [Preprint]. doi:10.1109/csci49370.2019.00264.

[23] Sarma, S.K. (2021) 'Metaheuristic based auto-scaling for microservices in a cloud environment: A new container-aware application scheduling', International Journal of Pervasive Computing and Communications, 19(1), pp. 74–96. doi:10.1108/ijpcc-12-2020-0213.

[24] Marie-Magdelaine, N. and Ahmed, T. (2020) 'Proactive autoscaling for Cloud-Native Applications using machine learning', GLOBECOM 2020 - 2020 IEEE Global Communications Conference [Preprint]. doi:10.1109/globecom42002.2020.9322147.

[25] Radhika, E.G., Sudha Sadasivam, G. and Fenila Naomi, J. (2018) 'An efficient predictive technique to Autoscale the resources for web applications in private cloud', 2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) [Preprint]. doi:10.1109/aeeicb.2018.8480899.

[26] Shariffdeen, R.S. et al. (2016) 'Adaptive workload prediction for Proactive Auto Scaling in paas systems', 2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech) [Preprint]. doi:10.1109/cloudtech.2016.7847713.

[27] Iqbal, W., Erradi, A. and Mahmood, A. (2018) 'Dynamic workload patterns prediction for proactive auto-scaling of web applications', Journal of Network and Computer Applications, 124, pp. 94–107. doi:10.1016/j.jnca.2018.09.023.

[28] Roy, N., Dubey, A. and Gokhale, A. (2011) 'Efficient autoscaling in the cloud using predictive models for workload forecasting', 2011 IEEE 4th International Conference on Cloud Computing [Preprint]. doi:10.1109/cloud.2011.42.

[29] Srirama, S.N., Adhikari, M. and Paul, S. (2020) 'Application deployment using containers with auto-scaling for microservices in cloud environment', Journal of Network and Computer Applications, 160, p. 102629. doi:10.1016/j.jnca.2020.102629.

[30] Liu, B., Buyya, R. and Nadjaran Toosi, A. (2018) 'A fuzzy-based auto-scaler for web applications in cloud computing environments', Service-Oriented Computing, pp. 797–811. doi:10.1007/978-3-030-03596-9_57.

[31] Heimerson, A., Eker, J. and Årzén, K.-E. (2022) 'A proactive cloud application auto-scaler using reinforcement learning', 2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC) [Preprint]. doi:10.1109/ucc56403.2022.00040.

[32] Bibal Benifa, J.V. and Dejey, D. (2018) 'RLPAS: Reinforcement learning-based proactive auto-scaler for resource provisioning in cloud environment', Mobile Networks and Applications, 24(4), pp. 1348–1363. doi:10.1007/s11036-018-0996-0.

[33] Biswas, A. et al. (2014) 'Automatic resource provisioning: A machine learning based proactive approach', 2014 IEEE 6th International Conference on Cloud Computing Technology and Science [Preprint]. doi:10.1109/cloudcom.2014.147.

[34] Sood, S.K. and Sandhu, R. (2015) 'Matrix based proactive resource provisioning in Mobile Cloud Environment', Simulation Modelling Practice and Theory, 50, pp. 83–95. doi:10.1016/j.simpat.2014.06.004.

[35] Messias, V.R. et al. (2015) 'Combining time series prediction models using genetic algorithm to autoscaling web applications hosted in the Cloud Infrastructure', Neural Computing and Applications, 27(8), pp. 2383–2406. doi:10.1007/s00521-015-2133-3.

[36] Caglar, F. et al. (2013) 'Model-driven performance estimation, deployment, and Resource Management for Cloud-hosted services', Proceedings of the 2013 ACM workshop on Domain-specific modeling [Preprint]. doi:10.1145/2541928.2541933.

[37] Bunch, C. et al. (2012) 'A pluggable Autoscaling Service for Open Cloud Paas Systems', 2012 IEEE Fifth International Conference on Utility and Cloud Computing [Preprint]. doi:10.1109/ucc.2012.12.

[38] Aws auto-scaling: http://aws.amazon.com/documentation/autoscaling

[39] Rudrabhatla, C.K. (2020) 'A quantitative approach for estimating the scaling thresholds and step policies in a distributed microservice architecture', IEEE Access, 8, pp. 180246–180254. doi:10.1109/access.2020.3028310.

[40] Abdullah, M. et al. (2022) 'Burst-aware predictive autoscaling for containerized microservices', IEEE Transactions on Services Computing, 15(3), pp. 1448–1460. doi:10.1109/tsc.2020.2995937.

[41] Bouabdallah, R., Lajmi, S. and Ghedira, K. (2016) 'Use of reactive and proactive elasticity to adjust resources provisioning in the cloud provider', 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS) [Preprint]. doi:10.1109/hpcc-smartcity-dss.2016.0162.

[42] Al Qassem, L.M. et al. (2023) 'Proactive random-forest autoscaler for Microservice Resource Allocation', IEEE Access, 11, pp. 2570–2585. doi:10.1109/access.2023.3234021.

[43] Prachitmutita, I. et al. (2018) 'Auto-scaling microservices on iaas under SLA with cost-effective framework', 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI) [Preprint]. doi:10.1109/icaci.2018.8377525.

[44] ZargarAzad, M. and Ashtiani, M. (2023) 'An auto-scaling approach for microservices in cloud computing environments', Journal of Grid Computing, 21(4). doi:10.1007/s10723-023-09713-7.

[45] Sun, Y. et al. (2016) 'Automated QoS-oriented cloud resource optimization using containers', Automated Software Engineering, 24(1), pp. 101–137. doi:10.1007/s10515-016-0191-0.

[46] Nguyen, H.X., Zhu, S. and Liu, M. (2022) 'Graph-PHPA: Graph-based proactive horizontal pod Autoscaling for microservices using LSTM-GNN', 2022 IEEE 11th International Conference on Cloud Networking (CloudNet)[Preprint]. doi:10.1109/cloudnet55617.2022.9978781.

[47] Khan, M.N. et al. (2015) 'Modeling the Autoscaling Operations in Cloud With Time Series Data', 2015 IEEE 34th Symposium on Reliable Distributed Systems Workshop (SRDSW) [Preprint]. doi:10.1109/srdsw.2015.20.

[48] Ghobaei-Arani, M., Jabbehdari, S. and Pourmina, M.A. (2016) 'An autonomic approach for Resource Provisioning of Cloud Services', Cluster Computing, 19(3), pp. 1017–1036. doi:10.1007/s10586-016-0574-9.

[49] Aws SDK for Java: https://docs.aws.amazon.com/sdk-for-java/?icmpid=docs_homepage_sdktoolkits/index.html

[50] AWS Cloudwatch: https://docs.aws.amazon.com/cloudwatch/?icmpid=docs_homepage_mgmtgov

[51] Aws ECS: https://docs.aws.amazon.com/ecs/