REFERENCES

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility,” *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, Jun. 2009, doi: 10.1016/j.future.2008.12.001.

[2] Gartner, “Gartner forecasts worldwide public cloud end-user spending to reach nearly $600 billion in 2023,” *https://www.gartner.com/en/newsroom/press-releases/2022- 10-31-gartner-forecasts-worldwide-public-cloud-end-user-spending-to- reach-nearly-600-billion-in-2023*, Oct. 2022.

[3] Gartner, “Gartner says worldwide iaas public cloud services market grew 41.4% in 2021,” *https://www.gartner.com/en/newsroom/press-releases/2022-06-02-gartner-says-worldwide-iaas-public-cloud-services-market-grew-41- percent-in-2021*, Jun. 2022.

[4] C. Wu, R. Buyya, and K. Ramamohanarao, “Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges,” *ACM Computing Surveys*, vol. 52, no. 6. Association for Computing Machinery, Oct. 01, 2019. doi: 10.1145/3342103.

[5] E. Cortez, M. Russinovich, A. Bonde, M. Fontoura, A. Muzio, and R. Bianchini, “Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms?,” in *SOSP 2017 - Proceedings of the 26th ACM Symposium on Operating Systems Principles*, Association for Computing Machinery, Inc, Oct. 2017, pp. 153–167. doi: 10.1145/3132747.3132772.

[6] L. Lin, L. Pan, and S. Liu, “Methods for improving the availability of spot instances: A survey,” *Computers in Industry*, vol. 141. Elsevier B.V., Oct. 01, 2022. doi: 10.1016/j.compind.2022.103718.

[7] T. P. Pham, S. Ristov, and T. Fahringer, “Performance and Behavior Characterization of Amazon EC2 Spot Instances,” in *IEEE International Conference on Cloud Computing, CLOUD*, IEEE Computer Society, Sep. 2018, pp. 73–81. doi: 10.1109/CLOUD.2018.00017.

[8] M. Mao and M. Humphrey, *Auto-Scaling to Minimize Cost and Meet Application Deadlines in Cloud Workflows*. ACM, 2011.

[9] A. J. Sanad and M. Hammad, “Combining Spot Instances Hopping with Vertical Auto-scaling to Reduce Cloud Leasing Cost,” in *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies, 3ICT 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/3ICT51146.2020.9311955.

[10] K. Oh and M. Song, “Cocoa: Towards a Scalable Compute Cost-aware Data Analytics System,” in *Proceedings - 2021 IEEE International Conference on*

Cloud Engineering, IC2E 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 110–117. doi: 10.1109/IC2E52221.2021.00025.

[11] Monge, David A, and Gar, “Autoscaling scientific workflows on the cloud by combining on-demand and spot instances,” *Computer Systems Science and Engineering*, vol. 32, no. 4, pp. 291–306, 2017.

[12] R. Cushing, S. Koulouzis, A. S. Z. Belloum, and M. Bubak, *Prediction-based Auto-scaling of Scientific Workflows*. ACM, 2011.

[13] L. Versluis, M. Neacsu, and A. Iosup, “A trace-based performance study of autoscaling workloads of workflows in datacenters,” in *Proceedings - 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2018*, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, pp. 223–232. doi: 10.1109/CCGRID.2018.00037.

[14] B. Baliś, A. Broński, and M. Szarek, “Auto-scaling of Scientific Workflows in Kubernetes,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 33–40. doi: 10.1007/978-3-031-08754-7_5.

[15] S. Henning and W. Hasselbring, “Demo Paper: Benchmarking Scalability of Cloud-Native Applications with Theodolite,” in *Proceedings - 2022 IEEE International Conference on Cloud Engineering, IC2E 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 275–276. doi: 10.1109/IC2E55432.2022.00037.

[16] M. A. Tamiru, J. Tordsson, E. Elmroth, and G. Pierre, “An Experimental Evaluation of the Kubernetes Cluster Autoscaler in the Cloud,” in *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom*, IEEE Computer Society, Dec. 2020, pp. 17–24. doi: 10.1109/CloudCom49646.2020.00002.

[17] Z. Wu, X. Liu, Z. Ni, D. Yuan, and Y. Yang, “A market-oriented hierarchical scheduling strategy in cloud workflow systems,” *Journal of Supercomputing*, vol. 63, no. 1, pp. 256–293, Jan. 2013, doi: 10.1007/s11227-011-0578-4.

[18] L. F. Bittencourt and E. R. M. Madeira, “HCOC: A cost optimization algorithm for workflow scheduling in hybrid clouds,” *Journal of Internet Services and Applications*, vol. 2, no. 3, pp. 207–227, Dec. 2011, doi: 10.1007/s13174-011-0032-0.

[19] L. Ramakrishnan, J. S. Chase, D. Gannon, D. Nurmi, and R. Wolski, “Deadline-sensitive workflow orchestration without explicit resource control,” *J Parallel Distrib Comput*, vol. 71, no. 3, pp. 343–353, Mar. 2011, doi: 10.1016/j.jpdc.2010.11.010.

[20] S. Abrishami, M. Naghibzadeh, and D. H. J. Epema, “Deadline-constrained workflow scheduling algorithms for Infrastructure as a Service Clouds,” *Future Generation Computer Systems*, vol. 29, no. 1, pp. 158–169, Jan. 2013, doi: 10.1016/j.future.2012.05.004.

[21] M. T. Islam, S. N. Srirama, S. Karunasekera, and R. Buyya, “Cost-efficient dynamic scheduling of big data applications in apache spark on cloud,” *Journal of Systems and Software*, vol. 162, Apr. 2020, doi: 10.1016/j.jss.2019.110515.

[22] S. Abrishami, M. Naghibzadeh, and D. H. J. Epema, “Deadline-constrained workflow scheduling algorithms for Infrastructure as a Service Clouds,” *Future Generation Computer Systems*, vol. 29, no. 1, pp. 158–169, Jan. 2013, doi: 10.1016/j.future.2012.05.004.

[23] P. Ambati and D. Irwin, “Optimizing the Cost of Executing Mixed Interactive and Batch Workloads on Transient VMs,” *In Proc. ACM Meas. Anal. Comput. Syst*, vol. 3, p. 24, 2019, doi: 10.1145/3326143.

[24] Z. Wei-guo, M. Xi-lin, and Z. Jin-zhong, “Research on kubernetes’ resource scheduling scheme,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2018. doi: 10.1145/3290480.3290507.

[25] Z. Zhong and R. Buyya, “A Cost-Efficient Container Orchestration Strategy in Kubernetes-Based Cloud Computing Infrastructures with Heterogeneous Resources,” *ACM Trans Internet Technol*, vol. 20, no. 2, May 2020, doi: 10.1145/3378447.

[26] O. M. Ungureanu, C. Vlădeanu, and R. Kooij, “Kubernetes cluster optimization using hybrid shared-state scheduling framework,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jul. 2019. doi: 10.1145/3341325.3341992.

[27] Q. Han, L. Niu, G. Quan, S. Ren, and S. Ren, “Energy efficient fault-tolerant earliest deadline first scheduling for hard real-time systems,” *Real-Time Systems*, vol. 50, no. 5–6, pp. 592–619, 2014, doi: 10.1007/s11241-014-9210-z.

[28] M. A. Haque, H. Aydin, and D. Zhu, “On Reliability Management of Energy-Aware Real-Time Systems Through Task Replication,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 3, pp. 813–825, 2017, doi: 10.1109/TPDS.2016.2600595.

[29] A. Ejlali, B. M. Al-Hashimi, and P. Eles, “Low-energy standby-sparing for hard real-time systems,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 3, pp. 329–342, Mar. 2012, doi: 10.1109/TCAD.2011.2173488.

[30] M. Salehi, A. Ejlali, and B. M. Al-Hashimi, "Two-Phase Low-Energy N-Modular Redundancy for Hard Real-Time Multi-Core Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 5, pp. 1497–1510, May 2016, doi: 10.1109/TPDS.2015.2444402.

[31] F. Mireshghallah, M. Bakhshalipour, M. Sadrosadati, and H. Sarbazi-Azad, "Energy-Efficient Permanent Fault Tolerance in Hard Real-Time Systems," 2019.

[32] A. M. Sampaio and J. G. Barbosa, *Enhancing Reliability of Compute Environments on Amazon EC2 Spot Instances*. 2019.

[33] S. Shastri and D. Irwin, "HotSpot: Automated server hopping in cloud spot markets," in *SoCC 2017 - Proceedings of the 2017 Symposium on Cloud Computing*, Association for Computing Machinery, Inc, Sep. 2017, pp. 493–505. doi: 10.1145/3127479.3132017.

[34] Y. Yan, Y. Gao, Y. Chen, Z. Guo, B. Chen, and T. Moscibroda, "TR-Spark: Transient computing for big data analytics," in *Proceedings of the 7th ACM Symposium on Cloud Computing, SoCC 2016*, Association for Computing Machinery, Inc, Oct. 2016, pp. 484–496. doi: 10.1145/2987550.2987576.

[35] P. Sharma, T. Guo, X. He, D. Irwin, and P. Shenoy, "Flint: Batch-interactive data-intensive processing on transient servers," in *Proceedings of the 11th European Conference on Computer Systems, EuroSys 2016*, Association for Computing Machinery, Inc, Apr. 2016. doi: 10.1145/2901318.2901319.

[36] F. Xu, H. Zheng, H. Jiang, W. Shao, H. Liu, and Z. Zhou, "Cost-effective cloud server provisioning for predictable performance of big data analytics," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 5, pp. 1036–1051, May 2019, doi: 10.1109/TPDS.2018.2873397.

[37] Z. Xu, C. Stewart, N. Deng, and X. Wang, "Blending on-demand and spot instances to lower costs for in-memory storage," in *Proceedings - IEEE INFOCOM*, Institute of Electrical and Electronics Engineers Inc., Jul. 2016. doi: 10.1109/INFOCOM.2016.7524348.

[38] D. Poola, K. Ramamohanarao, and R. Buyya, "Enhancing reliability of workflow execution using task replication and spot instances," in *ACM Transactions on Autonomous and Adaptive Systems*, Association for Computing Machinery, Feb. 2016. doi: 10.1145/2815624.

[39] R. Dewi and R. Munir, "Software Availability Enhancement in Preemptible Instance Kubernetes Cluster."

[40] J. Von Kistowski, S. Eismann, N. Schmitt, A. Bauer, J. Grohmann, and S. Kounev, "TeaStore: A micro-service reference application for benchmarking, modeling and resource management research," in *Proceedings - 26th IEEE International Symposium on Modeling, Analysis and Simulation of Computer*

and Telecommunication Systems, MASCOTS 2018, Institute of Electrical and Electronics Engineers Inc., Nov. 2018, pp. 223–236. doi: 10.1109/MASCOTS.2018.00030.

[41] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw Pract Exp*, vol. 41, no. 1, pp. 23–50, Jan. 2011, doi: 10.1002/spe.995.