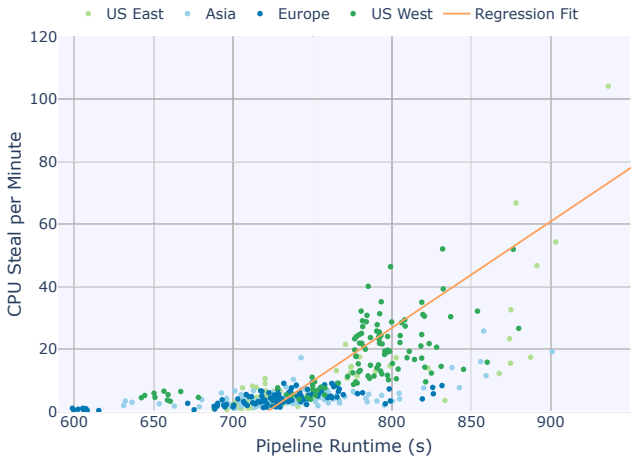

Figure 5: x86_64 CPU steal linear regression with runtime

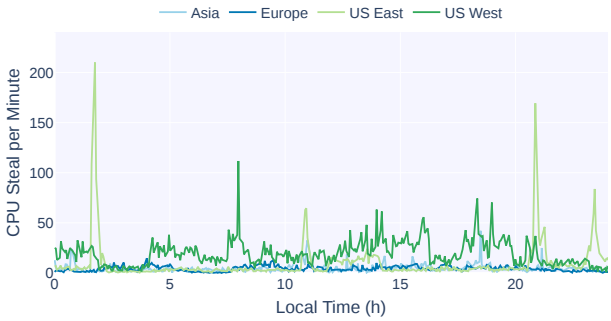

Figure 6: CPU architecture x86_64 CPU steal time in four regions.

runtime on Intel processors from 10:00 a.m. to 12:00 p.m. was 6% slower in contrast to 6:00 a.m. to 8:00 a.m., and 3.3% slower on ARM processors from 2:00 p.m. to 4:00 p.m. vs. 12:00 a.m. to 2:00 a.m. Figure 8 shows that NLP pipeline runtime tended to be faster outside regular business hours on any CPU architecture and region for our 24-hour study.

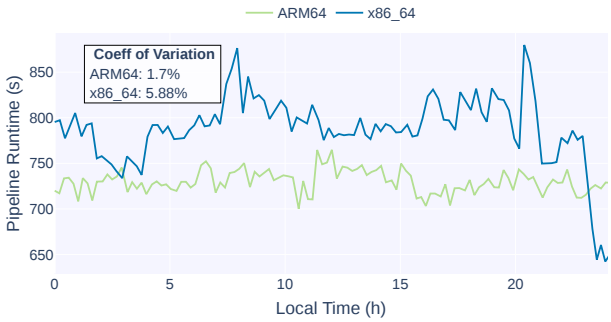

Figure 7: CPU architecture x86_64 versus ARM64 container runtimes in US West, Oregon.

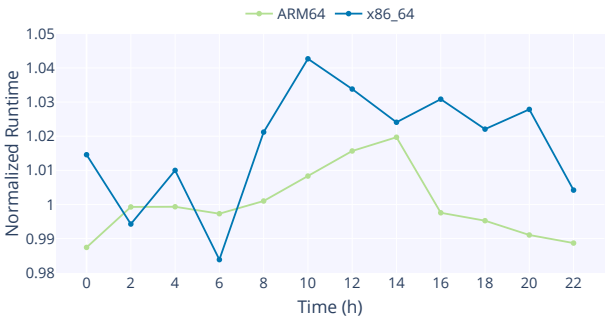

Figure 8: Global two-hour average NLP pipeline runtime normalized against global ARM64 average runtime.

5 CONCLUSIONS

Our study has helped identify performance differences of the ARM64 vs. x86_64 architectures on AWS Lambda while identifying diurnal performance patterns globally. The ARM64 architecture provided faster and more consistent runtime performance on average than the x86_64 architecture.

We summarize our findings with respect to the research questions as follows:

(RQ-1): Researchers and practitioners should be encouraged to adopt the ARM64 architecture on AWS Lambda when possible due not only to the discounted cost, but also because higher resource contention for x86_64 resources further exacerbates the cost differential. ARM64 cost savings, however, may only be temporary if the discount drives more users to adopt this architecture. We demonstrated up to **33.4% cost savings** for our NLP pipeline by leveraging ARM64 CPUs vs. x86_64 CPUs in us-west-2, the region exhibiting the highest resource contention.

(RQ-2): Researchers and practitioners running non-latency sensitive workloads may consider redirecting their workloads to leverage regions outside regular business hours. For example, we observed a **6% global average runtime differential** across four regions from 6:00 to 8:00 a.m. vs. 10:00 to 12:00 p.m.

ACKNOWLEDGMENTS

This research is supported by the NSF Advanced Cyberinfrastructure Research Program (OAC-1849970), National Institutes of Health (NIH) National Institute of General Medical Sciences (NGMS) grant 5R01GM126019, and the AWS Cloud Credits for Research program.

REFERENCES

[1] Alexandru Agache, Marc Brooker, Alexandra Iordache, Anthony Liguori, Rolf Neugebauer, Phil Piwonka, and Diana-Maria Popa. 2020. Firecracker: Lightweight Virtualization for Serverless Applications. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. USENIX Association, Santa Clara, CA, 419–434. https://www.usenix.org/conference/nsdi20/presentation/agache

[2] Hitesh Ballani, Paolo Costa, Thomas Karagiannis, and Ant Rowstron. 2011. Towards Predictable Datacenter Networks. In *Proceedings of the ACM SIGCOMM 2011 Conference* (Toronto, Ontario, Canada) *(SIGCOMM '11)*. Association for Computing Machinery, New York, NY, USA, 242–253. https://doi.org/10.1145/2018436.2018465

[3] Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural language processing with Python.

[4] Robert Cordingly, Wen Shu, and Wes J. Lloyd. 2020. Predicting Performance and Cost of Serverless Computing Functions with SAAF. In *2020 IEEE Intl Conf on*

Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE, 640–649. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00111

[5] Mohammadbagher Fotouhi, Derek Chen, and Wes J. Lloyd. 2019. Function-as-a-Service Application Service Composition: Implications for a Natural Language Processing Application. In *Proceedings of the 5th International Workshop on Serverless Computing* (Davis, CA, USA) *(WOSC '19)*. Association for Computing Machinery, New York, NY, USA, 49–54. https://doi.org/10.1145/3366623.3368141

[6] Andrei Frumusanu. 2020. Amazon's arm-based Graviton2 against AMD and Intel: Comparing cloud compute. https://www.anandtech.com/show/15578/cloud-clash-amazon-graviton2-arm-against-intel-and-amd

[7] Samuel Ginzburg and Michael J. Freedman. 2020. Serverless Isn't Server-Less: Measuring and Exploiting Resource Variability on Cloud FaaS Platforms. In *Proceedings of the 2020 Sixth International Workshop on Serverless Computing* (Delft, Netherlands) *(WoSC'20)*. Association for Computing Machinery, New York, NY, USA, 43–48. https://doi.org/10.1145/3429880.3430099

[8] Brendan Gregg. 2017. AWS EC2 Virtualization 2017: Introducing Nitro. https://www.brendangregg.com/blog/2017-11-29/aws-ec2-virtualization-2017.html. Accessed: 2022-01-25.

[9] Xinlei Han, Raymond Schooley, Delvin Mackenzie, Olaf David, and Wes J Lloyd. 2020. Characterizing public cloud resource contention to support virtual machine co-residency prediction. In *2020 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, IEEE, 162–172.

[10] Eric Jonas, Johann Schleier-Smith, Vikram Sreekanti, Chia-Che Tsai, Anurag Khandelwal, Qifan Pu, Vaishaal Shankar, Joao Carreira, Karl Krauth, Neeraja Yadwadkar, Joseph E. Gonzalez, Raluca Ada Popa, Ion Stoica, and David A. Patterson. 2019. Cloud Programming Simplified: A Berkeley View on Serverless Computing. https://doi.org/10.48550/ARXIV.1902.03383

[11] Artur Klauser. 2020. Building Multi-Architecture Docker Images With Buildx. https://medium.com/@artur.klauser/building-multi-architecture-docker-images-with-buildx-27d80f7e2408

[12] Rohit Kulkarni. 2017. A Million News Headlines. https://www.kaggle.com/therohk/million-headlines.

[13] Philipp Leitner and Jürgen Cito. 2016. Patterns in the chaos—a study of performance variation and predictability in public iaas clouds. *ACM Transactions on Internet Technology (TOIT)* 16, 3 (2016), 1–23.

[14] Wes Lloyd, Shrideep Pallickara, Olaf David, Mazdak Arabi, and Ken Rojas. 2017. Mitigating resource contention and heterogeneity in public clouds for scientific modeling services. In *2017 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 159–166. https://doi.org/10.1109/IC2E.2017.29

[15] Wes Lloyd, Shruti Ramesh, Swetha Chinthalapati, Lan Ly, and Shrideep Pallickara. 2018. Serverless computing: An investigation of factors influencing microservice performance. In *2018 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 159–169. https://doi.org/10.1109/IC2E.2018.00039

[16] David Perez, Ling-Hong Hung, Sonia Xu, Ka Yee Yeung, and Wes Lloyd. 2020. Characterizing Performance Variation of Genomic Data Analysis Workflows on the Public Cloud. In *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE, IEEE, 680–683.

[17] David Perez, Ling-Hong Hung, Sonia Xu, Ka Yee Yeung, and Wes Lloyd. 2020. An Investigation on Public Cloud Performance Variation for an RNA Sequencing Workflow. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (Virtual Event, USA) *(BCB '20)*. Association for Computing Machinery, New York, NY, USA, Article 96, 7 pages. https://doi.org/10.1145/3388440.3414859

[18] Danilo Poccia. 2020. New for AWS Lambda – Container Image Support. https://aws.amazon.com/blogs/aws/new-for-aws-lambda-container-image-support

[19] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[20] Jörg Schad, Jens Dittrich, and Jorge-Arnulfo Quiané-Ruiz. 2010. Runtime measurements in the cloud: observing, analyzing, and reducing variance. *Proc. VLDB Endow.* 3 (2010), 460–471.

[21] Robert Schmitz, III and Danielle Lambion. 2022. *FaaS performance variability study using topic-modeling*. https://github.com/lambiond/faas_variability_topic_modeling

[22] Alexandru Uta and Harry Obaseki. 2018. A Performance Study of Big Data Workloads in Cloud Datacenters with Network Variability. In *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering* (Berlin, Germany) *(ICPE '18)*. Association for Computing Machinery, New York, NY, USA, 113–118. https://doi.org/10.1145/3185768.3186299

[23] Liang Wang, Mengyuan Li, Yinqian Zhang, Thomas Ristenpart, and Michael Swift. 2018. Peeking Behind the Curtains of Serverless Platforms. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*. USENIX Association, Boston, MA, 133–146. https://www.usenix.org/conference/atc18/presentation/wang-liang