Multi-Node Spot Instances Availability Score Collection System

Sungkyu Cheon*
Department of Computer Science
Kookmin University
Seoul, South Korea
wsx2138@kookmin.ac.kr

Kyumin Kim*
Department of Computer Science
Kookmin University
Seoul, South Korea
okkimok123@kookmin.ac.kr

Kyunghwan Kim*
Department of Computer Science
Kookmin University
Seoul, South Korea
bryan9801@kookmin.ac.kr

Moohyun Song
Department of Computer Science
Kookmin University
Seoul, South Korea
mhsong@kookmin.ac.kr

Kyungyong Lee†
Department of Data Science
Hanyang University
Seoul, South Korea
kyungyong@hanyang.ac.kr

ABSTRACT

Spot instances let users access unused cloud resources at significantly reduced costs. While cloud vendors offer availability information, existing tools like Spotlake only provide single-node availability data, which falls short for modern distributed applications. This paper highlighted the limitations of single-node availability data and introduced a multi-node availability dataset collection system. We analyzed the collected data and enhanced Spotlake to share these multi-node datasets publicly for broader use.

KEYWORDS

cloud computing, spot instance, spot instance datasets

ACM Reference Format:
Sungkyu Cheon, Kyumin Kim, Kyunghwan Kim, Moohyun Song, and Kyungyong Lee. 2025. Multi-Node Spot Instances Availability Score Collection System. In *The 34th International Symposium on High-Performance Parallel and Distributed Computing (HPDC '25), July 20–23, 2025, Notre Dame, IN, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3731545.3735122

1 INTRODUCTION

Spot instances are services in which cloud vendors offer unused resources at discounts of up to 90%. These services are provided by most major cloud vendors, including AWS, Azure, GCP, Alibaba, and IBM. When using spot instances, cost savings and stability are critical considerations for users due to the dynamically changing spot price and interrupt events. To assist users, public cloud vendors offer datasets on spot prices and availability. For example, AWS offers interrupt ratios for the past month and real-time availability data, such as AWS Spot Placement Score (SPS) [3]. This dataset indicates the immediate availability of spot instances, without disclosing internal details. Users can access this dataset via the management console or API, though there are significant query limitations. To overcome this limitation and facilitate easier access to the data, a web service called Spotlake has been proposed [5].

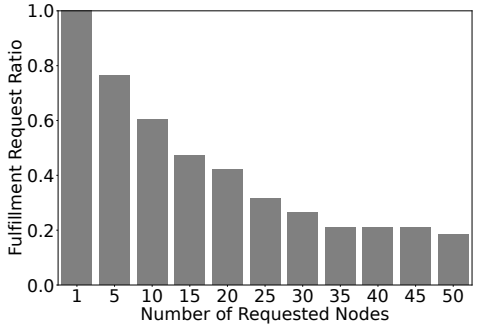


Figure 1: The ratio of fulfilled instances when requesting a different number of nodes whose single-node availability score is high.

Currently, modern cloud applications often require distributed environments with multiple GPU-equipped nodes [9] and large-scale computing resources [6], both of which lead to substantial costs. To reduce these costs, using multi-node spot instances is becoming increasingly common [1, 2, 7, 8], thereby highlighting the growing need for multi-node spot instance availability information. However, Spotlake provides SPS information only for single-node spot instances and does not offer SPS scores for scenarios in which multiple instances are requested simultaneously.

In this paper, we explore the limitations of using a single-node availability score and propose a solution to overcome these limitations. To explore whether single-node SPS values can reflect availability in multi-node requests, we randomly selected 32 instance types with the maximum SPS score of 3 and tested their success rates across increasing instance counts. As shown in Figure 1, the success rate dropped sharply from 100% at 1 node to only 20% at 50 nodes. This result confirms that single-node SPS is not a reliable indicator for multi-node provisioning.

To address these limitations, we proposed a multi-node availability dataset collection system and conducted a thorough analysis

*Equal contribution
†Corresponding author

HPDC '25, July 20–23, 2025, Notre Dame, IN, USA

ACM ISBN 979-8-4007-1869-4/25/07
https://doi.org/10.1145/3731545.3735122

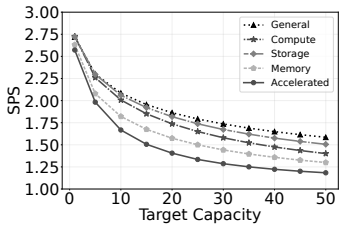
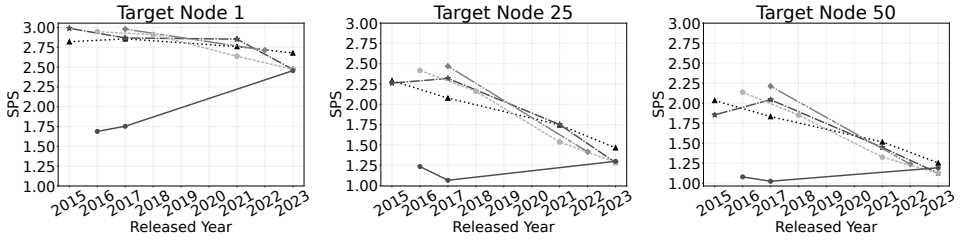


Figure 2: SPS score distribution



(a) Target node number : 1 **(b) Target node number : 25** **(c) Target node number : 50**

Figure 3: Temporal changes of SPS for different capacity grouped by instance categories

of the collected data. Furthermore, we enhanced the Spotlake platform by integrating these multi-node availability scores and making them publicly available as a web service.[1] Our solution empowers users to make cost-efficient and reliable decisions for large-scale distributed applications in cloud environments.

2 IMPLEMENTATION & DATA ANALYSIS

Implementation. We conducted queries every 10 minutes and collected data for various target node counts. For price data, we directly utilized the price dataset provided by AWS. The collected dataset is stored in *S3*, available for direct access by users. The files are saved with a Year-Month-Date-Time naming convention to include historical data.

Data Analysis. We analyzed the collected multi-node SPS dataset, which spans from July 1, 2024, to January 15, 2025, covering 844 unique instance types across 17 regions.

Figure 2 shows a line chart of SPS variations by instance category. The X-axis represents target capacity, and the Y-axis shows average SPS values. Each line corresponds to an instance category. SPS values generally decrease as the requested capacity increases, indicating reduced availability for large requests. Accelerated Computing instances show the steepest decline, likely due to high demand for GPU-equipped instances [4]. Other categories also show varying degrees of reduction. These results highlight the importance of selecting suitable instance types for large-scale resource pools.

We further examined the relationship between instance release dates and SPS values across instance categories. Figure 3 plots SPS trends, with the horizontal axis representing the release year and the vertical axis showing average SPS values. Each subplot differentiates target node counts, grouped by each instance category.

For most categories, newer instance types tend to have lower SPS values, and this effect intensifies as the number of target nodes increases (see Figures 3b and 3c). However, in the Accelerated Computing category, newer instances exhibit higher SPS values; the *p5.48xlarge*, equipped with NVIDIA H100 GPUs, demonstrates improved availability. Since SPS calculation details are not disclosed, the exact cause of this pattern is unclear. One plausible explanation is that upgrading to newer CPU-based instances is easier due to fewer software dependencies compared to GPUs, leading to higher adoption. Additionally, the high cost of the latest GPU instances may limit their usage, resulting in higher SPS values.

[1] https://spotlake.ddps.cloud/

3 CONCLUSION

We addressed the limitations of single-node availability score datasets for spot instances in a multi-node setup and proposed a multi-node availability dataset collection system. However, the current system only supports AWS spot instances, and future work should extend compatibility to other cloud providers like Microsoft Azure.

ACKNOWLEDGMENTS

This work is supported by Institute of Information & communications Technology Planning & Evaluation (IITP) Grant funded by the Korean Government (MSIT) : RS-2022-00144309.

REFERENCES

[1] Navraj Chohan, Claris Castillo, Mike Spreitzer, Malgorzata Steinder, Asser Tantawi, and Chandra Krintz. 2010. See spot run: Using spot instances for {MapReduce} workflows. In *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10)*.

[2] Jiangfei Duan, Ziang Song, Xupeng Miao, Xiaoli Xi, Dahua Lin, Harry Xu, Minjia Zhang, and Zhihao Jia. 2024. Parcae: Proactive, Liveput-Optimized DNN Training on Preemptible Instances. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. USENIX Association, Santa Clara, CA, 1121–1139. https://www.usenix.org/conference/nsdi24/presentation/duan

[3] AWS What is New. 2021. Introducing Amazon EC2 Spot placement score. https://aws.amazon.com/about-aws/whats-new/2021/10/amazon-ec2-spot-placement-score/

[4] K. Lee and M. Son. 2017. DeepSpotCloud: Leveraging Cross-Region GPU Spot Instances for Deep Learning. In *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*. 98–105. https://doi.org/10.1109/CLOUD.2017.21

[5] S. Lee, J. Hwang, and K. Lee. 2022. SpotLake: Diverse Spot Instance Dataset Archive Service. In *2022 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE Computer Society, Los Alamitos, CA, USA, 242–255. https://doi.org/10.1109/IISWC55918.2022.00029

[6] Josep Sampé, Marc Sánchez-Artigas, Gil Vernik, Ido Yehekzel, and Pedro García-López. 2023. Outsourcing Data Processing Jobs With Lithops. *IEEE Transactions on Cloud Computing* 11, 1 (2023), 1026–1037. https://doi.org/10.1109/TCC.2021.3129000

[7] Myungjun Son, Gulsum Gudukbay Akbulut, and Mahmut Taylan Kandemir. 2024. SpotVerse: Optimizing Bioinformatics Workflows with Multi-Region Spot Instances in Galaxy and Beyond. In *Proceedings of the 25th International Middleware Conference* (Hong Kong, Hong Kong) *(Middleware '24)*. Association for Computing Machinery, New York, NY, USA, 74–87. https://doi.org/10.1145/3652892.3700750

[8] P. Varshney and Y. Simmhan. 2019. AutoBoT: Resilient and Cost-Effective Scheduling of a Bag of Tasks on Spot VMs. *IEEE Transactions on Parallel & Distributed Systems* 30, 07 (jul 2019), 1512–1527. https://doi.org/10.1109/TPDS.2018.2889851

[9] Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. 2022. *DeepSpeed Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale.* Technical Report MSR-TR-2022-21. Microsoft. https://www.microsoft.com/en-us/research/publication/deepspeed-inference-enabling-efficient-inference-of-transformer-models-at-unprecedented-scale/