SWAN: Distributed Analytics and Smart Workload Prediction with Scalable Wide Area Network

Anshuman Das Mohapatra
Graduate Student, Computer Science

1. Introduction

Geo-distributed data analytics (GDA) plays an important role for many Internet applications to mine meaningful insights from large-scale geo-distributed data generated in a multiple-cloud environment. It comprises tens of geo-distributed data centers (DCs) and hundreds of smaller scattered edge clusters of cloud providers like Amazon, Microsoft, and Google. For example, Facebook and Twitter analyze highly diffused users posts and systems logs to query global trends, make advertising decisions, and check overall cluster health. Minimizing GDA query latency is a critical workload for these Internet applications as it can affect their revenues significantly.

Since these queries can be sent at any time, i.e., ad-hoc queries, which may cause peak workload, existing GDA systems typically deploy overprovisioned compute resources a prior, i.e., redundant virtual machines (VMs) in multiple DCs, to process incoming queries without a performance bottleneck. This approach is simple and allows queries to be handled promptly. However, this approach would incur a significant monetary cost ($), i.e., cost-bottleneck, for redundant VM instances that are charged throughout allocation even though they are idle.

To avoid additional cost for idle compute resources, many scalable systems [12, 13, 16] have been introduced to determine optimal compute resource configurations by predicting applications workloads, i.e., deploying additional VM instances to handle peak workloads and terminating them when the queries are done. These systems, unfortunately, may encounter a performance bottleneck for latency-sensitive workloads because they have ignored unavoidable overhead of VM, i.e., boot-up time (> 60 seconds) [29]. In addition, these systems may encounter a cost-bottleneck as well because they only considered a single DC setting, where the network is not a performance-bottleneck [9, 10, 46] and data transfer is free of charge. Thus, these systems cannot work well for GDA that requires large data transfer via a wide area network (WAN), one of the most expensive and scarce resources in a multi-cloud environment.

This proposal aims to determine optimal cloud configurations (how much compute resources at each DC) to handle GDA queries in a timely and cost-efficient way while avoiding performance- and cost-bottlenecks. To achieve this goal, the proposed research agenda includes predicting applications workloads by building a prediction model and determining the best cloud configurations for queries based on prediction. In addition, a newly emerging compute resource called serverless (SL), which offers agility, i.e., very small boot-up time (< 100 ms), will be considered for optimal cloud configurations to avoid the significant boot-up overhead of VM.

The proposed ideas and approaches will be evaluated to show their efficacy by building and deploying a prototype implementation in both simulated and real multi-cloud environments and comparing them with state-of-the-art approaches as baselines. Results from this research effort will be used to design a scalable GDA system, which exploits diverse and heterogeneous cloud resources in a multi-cloud environment to achieve the GDA applications' desired cost-performance goals easily.

2. Related Work and Motivation

To achieve the goal in this proposal, precisely predicting GDA workloads is one of the most critical tasks. This section presents existing state-of-the-art techniques for workload prediction and other related works.

2.1. Workload Prediction

Table 1 shows state-of-the-art solutions for workload prediction to determine optimal compute resource configurations. Ernest [45] and Optimus [33] attempt to model the application parameters and then use a Non-Negative Least Squares (NNLS) algorithm to estimate optimal values. Cherrypick [1] uses a Bayesian Optimizer [11, 19] and assumes the performance model of a distributed cloud application as a black box. CrystalLP [38] uses a deep learning model of Long Short-Term Memory (LSTM) to generate a future set of data server workloads. Since these systems have considered VM only for optimal cloud resource configurations, they will encounter a performance bottleneck while handling latency-sensitive workloads due to the boot-up latency of VM. To avoid the overhead of VM, recent works, MArk [50] and Spock [17] have applied LSTM and Linear Regression (LR) respectively to a combination of VM and SL instances based on the ML-Inference application attributes. These existing workload

Table 1. State-of-the-art Workload Prediction Systems

Systems	***VMs***	***SLs***	***Technique***	***Application***	WAN	Tradeoff
Ernest [45]	Yes	No	NNLS	Advanced Analytics	No	No
Optimus [33]	Yes	No	NNLS	Deep Learning	No	No
Cherrypick [1]	Yes	No	Bayesian	Recurring Data Analytics	No	No
CrystalLP [38]	Yes	No	LSTM	Storage Systems	No	No
MArk [50]	Yes	Yes	LSTM	ML Inference	No	No
Spock [17]	Yes	Yes	LR	ML Inference	No	No

prediction systems, unfortunately, will not work well for GDA as they have not considered WAN that causes both performance- and cost-bottlenecks in GDA.

2.2. Serverless-enabled data analytics systems

SL is appealing for data analytics to handle peak workloads thanks to its agility, i.e., minimal boot-up time (< 100 ms), and thus numerous SL-enabled data analytics (SDA) systems [24, 27, 28, 35, 40, 42] have been introduced. These systems focused on addressing the limitations of SL and showed that using SL can handle peak workloads quickly and cost-efficiently without over-provisioning VMs. However, these systems may encounter performance and cost bottlenecks because they have ignored the more expensive per unit time cost ($) and worse performance of SL compared to VM [32]. To get composite benefits from heterogeneous compute resources, i.e., VM and SL, recent SDA systems [22, 23, 32, 36] tried to utilize VM and SL together to handle queries. However, none of the existing SDA systems considered WAN, and thus they may encounter performance- and cost-bottlenecks.

2.3. Geo-distributed Data Analytics Systems

Many GDA systems [18, 21, 34, 47, 48] have been proposed to reduce the overall makespan of GDA queries by overcoming the limitations of the WAN. In recent work, Kimchi [31] incorporates data transfer cost into task placement decisions and handles changes in network dynamics with little or no impact on the job makespan. However, these GDA systems were designed based on simple assumptions, both/either infinite compute resources and/or non-scalable compute resources, which would cause a cost-bottleneck due to redundant compute resources and a performance bottleneck due to lack of compute resources. That is, none of the GDA systems have predicted GDA queries' workloads to determine optimal cloud configurations in a multi-cloud environment.

To summarize, the previous postulations have overlooked WAN characteristics in the context of workload prediction. Additionally, GDA systems on VM and SL mix need the hour to handle dynamic workloads efficiently. Hence, we propose a Scalable WAN (SWAN) model to exploit network properties in optimal cloud resource prediction and explore geo-distributed heterogeneous compute resources for handling dynamic workloads.

3. Methodology

Fig. 1 presents the overall architecture. The new model will use Bayesian Optimization (BO) to determine the optimal number of instances for executing a query with minimum cost or time. These determinations will be used by a Workload Handler (WH) module to schedule incoming tasks onto VM and SL instances. We will choose BO because it incorporates black-box optimization, which will take care of the implicit non-linear relationships, which otherwise would have been difficult to model [8, 30, 39, 41]. Other works that have tried establishing mathematical relations between various features of a distributed application [6, 18, 33, 45] are either incompatible with other applications [17, 50] or miss out on specific key characteristics. The choice of BO, however, presents another challenge. It requires model fitting with real objective function, to explore/exploit the search space for global optima. This can be achieved with representative workloads or test runs orchestrated on actual VM and SL instances, but both of these are inefficient and costly. Therefore, we will use Random Forest (RF) [7, 14, 25] to model the heterogeneous platform of VMs and SLs in a WAN setting. Several network and instance properties will be embedded into the training of this model, so that the predictions accurately capture network latency and cost. The choice of RF over other deep learning neural networks [38] is because it is less computationally intensive and requires less training data [2, 15, 26, 44]. Although it is possible to iteratively parse all the available instance configurations through the RF sub-module only, this operation would be costly. Considering 10 DCs and instance configuration for each DC in the range 0–10, the search space would consist of a significant number of

$(11^{10} - 1)$ combinations. Thus, the use of RF will help BO navigate through the plausible configurations efficiently.

The monitoring module and the WH task scheduler of SWAN will be implemented in Spark [49], while the SWP and remainder of the WH module will be designed and implemented as separate modules written in Python. The SWAN prototype implementation will be deployed and evaluated both in a simulated WAN environment (CloudLab) [37] and public cloud environments (e.g., AWS and Azure). For serverless compute resources, we will deploy Apache OpenWhisk [3] on CloudLab and utilize AWS Lambda [4] and Azure Functions [5]. For workloads, we will use popular big data benchmarks, e.g., HiBench [20] (the bigdata micro-benchmark suite) and TPC-DS [43] (a standard decision support benchmark). We will compare the prediction latency and precision with the state-of-the-art approaches presented in Section 2, e.g., MArk [50].

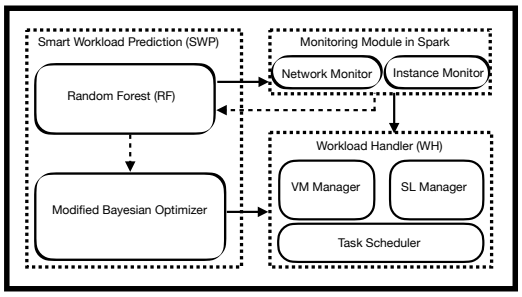

Fig. 1. Overview of the Proposed Research

4. Challenges

Considering SL instances for optimal resource determination increases the problem's dimensionality , which necessitates prediction efficiency and accuracy. When coupled with WAN bandwidth, the complexity for decision-making is further increased and geo-locality predictions must account for data locality in map stages and the minimum shuffle overhead with lesser data transferred over the network. Additionally, the tradeoff space between monetary budget and query latency needs to be carefully explored so that the high per-unit cost of SL instances do not overshadow the cost of overprovisioned instances.

5. Project Timeline

The detailed project timeline is presented below.

Tasks	Activities	Timeline
Design	• Plan interaction among various modules and sub-modules in Spark. • Determine the features to be used for finding optimal cloud configuration in a WAN setting along with a plan to capture these metrics.	May-July
Execute & Test	• Implement the proposed changes. • Deploy the prototype in cloud and compare the results with recent works.	June-August
Project Report	• Document the proposition, approach, challenges, implementation and experimental results. • Work with Dr. Oh on the final report.	May-August

6. Student/Faculty Mentor Roles

As the principal investigator of this research proposal, Anshuman Das Mohapatra will be responsible for designing, modeling and implementing the proposed idea. Anshuman will also be accountable for running experiments on simulated and real test beds for validation of the implemented prototype. The research advisor, Dr. Kwangsung Oh, will review the progress through weekly meetings and address any unforeseen challenges. In addition, Dr. Oh will assess the experimental results.

7. Outcomes

To recapitulate, our contribution through this project will be as follows:

- To the best of our knowledge, SWAN will be the first working model (involving intelligent workload prediction) to use both VMs and SLs for distributed analytics in WAN.
- We will explore the tradeoffs between fastest query resolution versus budgeted query resolution.
- We will develop a prototype for the prediction model in Python and integrate it with Spark. The prototype implementation will be deployed and evaluated in public cloud environments (e.g., AWS and Azure) and in scientific infrastructure (CloudLab) using popular big data benchmarks such as TPC-DS and HiBench.

8. Budget Justification

The budget requested for this research project is $5,000, details of which are given below. Anshuman has a graduate assistantship that supports him over the Fall and Spring terms. The following stipend will allow him to pursue research activities during the summer.

Item	Justification	Amount
Graduate Student Stipend	Summer stipend (May 2022 - Aug 2022 for Anshuman Das Mohapatra who will spend approximately 375 hours on this project.)	$4,500
Experimental Expenses	For thorough evaluation, the working prototype will be deployed on cloud instances rented from Amazon and Microsoft.	$500

References

1. O. Alipourfard, H. H. Liu, J. Chen, S. Venkataraman, M. Yu, and M. Zhang. Cherrypick: Adaptively unearthing the best cloud configurations for big data analytics. In *Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation*, NSDI'17, pages 469–482, Berkeley, CA, USA, 2017. USENIX Association.
2. O. Anisfeld, E. Biton, R. Milshtein, M. Shifrin, and O. Gurewitz. Scaling of cloud resources-principal component analysis and random forest approach. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–5, 2018.
3. Apache OpenWhisk, 2017. http://openwhisk.org/.
4. AWS Lambda, 2017. https://aws.amazon.com/lambda/.
5. Azure Functions, 2017. https://azure.microsoft.com/en-us/scrvices/functions/.
6. J. Bhimani, N. Mi, M. Leeser, and Z. Yang. Fim: Performance prediction for parallel computation in iterative data processing applications. In *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, pages 359–366, 2017.
7. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
8. E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, 2010.
9. M. Calder, X. Fan, Z. Hu, E. Katz-Bassett, J. Heidemann, and R. Govindan. Mapping the expansion of google's serving infrastructure. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, page 313–326, New York, NY, USA, 2013. Association for Computing Machinery.
10. M. Cardosa, C. Wang, A. Nangia, A. Chandra, and J. Weissman. Exploring mapreduce efficiency with highly-distributed data. In *Proceedings of the Second International Workshop on MapReduce and Its Applications*, MapReduce '11, page 27–34, New York, NY, USA, 2011. Association for Computing Machinery.
11. Y. Fan, J. Hu, and T. Sun. Performance prediction of network-intensive systems in cloud environment: A bayesian approach. In *2019 International Conference on Electronic Engineering and Informatics (EEI)*, pages 361–364, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society.
12. F. Fargo, C. Tunc, Y. Al-Nashif, A. Akoglu, and S. Hariri. Autonomic workload and resources management of cloud computing services. In *2014 International Conference on Cloud and Autonomic Computing*, pages 101–110, 2014.
13. F. Fargo, C. Tunc, Y. Al-Nashif, and S. Hariri. Autonomic performance-per-watt management (apm) of cloud resources and services. In *Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference*, CAC '13, New York, NY, USA, 2013. Association for Computing Machinery.
14. R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.
15. S. Ghosh and C. Banerjee. A predictive analysis model of customer purchase behavior using modified random forest algorithm in cloud environment. In *2020 IEEE 1st International Conference for Convergence in Engineering (ICCE)*, pages 239–244, 2020.
16. J. R. Gunasekaran, M. Cui, P. Thinakaran, J. Simons, M. T. Kandemir, and C. R. Das. Multiverse: Dynamic vm provisioning for virtualized high performance computing clusters. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pages 131–141, 2020.
17. J. R. Gunasekaran, P. Thinakaran, M. T. Kandemir, B. Urgaonkar, G. Kesidis, and C. Das. Spock: Exploiting serverless functions for slo and cost aware resource procurement in public cloud. In *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, pages 199–208, 2019.
18. B. Heintz, A. Chandra, R. K. Sitaraman, and J. Weissman. End-to-end optimization for geo-distributed mapreduce. *IEEE Transactions on Cloud Computing*, 4(3):293–306, 2016.
19. C.-J. Hsu, V. Nair, V. W. Freeh, and T. Menzies. Arrow: Low-level augmented bayesian optimization for finding the best cloud vm. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 660–670, 2018.

20. S. Huang, J. Huang, J. Dai, T. Xie, and B. Huang. The hibench benchmark suite: Characterization of the mapreduce-based data analysis. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pages 41–51, 2010.
21. C.-C. Hung, G. Ananthanarayanan, L. Golubchik, M. Yu, and M. Zhang. Wide-area analytics with multiple resources. In *Proceedings of the Thirteenth EuroSys Conference*, EuroSys '18, pages 12:1–12:16, New York, NY, USA, 2018. ACM.
22. A. Jain, A. F. Baarzi, G. Kesidis, B. Urgaonkar, N. Alfares, and M. Kandemir. Splitserve: Efficiently splitting apache spark jobs across faas and iaas. In *Proceedings of the 21st International Middleware Conference*, Middleware '20, page 236–250, New York, NY, USA, 2020. Association for Computing Machinery.
23. J. Jarachanthan, L. Chen, F. Xu, and B. Li. Astra: Autonomous serverless analytics with cost-efficiency and qos-awareness. In *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 756–765, 2021.
24. E. Jonas, Q. Pu, S. Venkataraman, I. Stoica, and B. Recht. Occupy the cloud: Distributed computing for the 99%. In *Proceedings of the 2017 Symposium on Cloud Computing*, page 445–451, New York, NY, USA, 2017. Association for Computing Machinery.
25. L. Khaidem, S. Saha, and S. R. Dey. Predicting the direction of stock market prices using random forest, 2016.
26. V. Khandelwal, A. K. Chaturvedi, and C. P. Gupta. Amazon ec2 spot price prediction using regression random forests. *IEEE Transactions on Cloud Computing*, 8(1):59–72, 2020.
27. Y. Kim and J. Lin. Serverless data analytics with flint. In *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*, pages 451–455, 2018.
28. A. Klimovic, Y. Wang, P. Stuedi, A. Trivedi, J. Pfefferle, and C. Kozyrakis. Pocket: Elastic ephemeral storage for serverless analytics. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'18, pages 427–444, Berkeley, CA, USA, 2018. USENIX Association.
29. M. Mao and M. Humphrey. A performance study on the vm startup time in the cloud. In *2012 IEEE Fifth International Conference on Cloud Computing*, pages 423–430, 2012.
30. J. Mockus. Bayesian approach to global optimization: Theory and applications, 2012.
31. K. Oh, A. Chandra, and J. Weissman. A network cost-aware geo-distributed data analytics system. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pages 649–658, 2020.
32. K. Oh and M. Song. Cocoa: Towards a scalable compute cost-aware data analytics system. In *Proceedings of the 9th IEEE International Conference on Cloud Engineering*. IEEE Computer Society Conference Publishing Services, Oct. 2021. *To Appear*.
33. Y. Peng, Y. Bao, Y. Chen, C. Wu, and C. Guo. Optimus: An efficient dynamic resource scheduler for deep learning clusters. In *Proceedings of the Thirteenth EuroSys Conference*, EuroSys '18, New York, NY, USA, 2018. Association for Computing Machinery.
34. Q. Pu et al. Low latency geo-distributed data analytics. *SIGCOMM Comput. Commun. Rev.*, 45(4):421–434, Aug. 2015.
35. Q. Pu, S. Venkataraman, and I. Stoica. Shuffling, fast and slow: Scalable analytics on serverless infrastructure. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, pages 193–206, Boston, MA, Feb. 2019. USENIX Association.
36. M. M. Rahman and M. Hasibul Hasan. Serverless architecture for big data analytics. In *2019 Global Conference for Advancement in Technology (GCAT)*, pages 1–5, 2019.
37. R. Ricci, E. Eide, and C. Team. Introducing cloudlab: Scientific infrastructure for advancing cloud architectures and applications. *login Usenix Mag.*, 39, 2014.
38. L. Ruan, Y. Bai, S. Li, S. He, and L. Xiao. Workload time series prediction in storage systems: a deep learning based approach. *Cluster Computing*, pages 1–11, 01 2021.
39. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
40. V. Shankar, K. Krauth, Q. Pu, E. Jonas, S. Venkataraman, I. Stoica, B. Recht, and J. Ragan-Kelley. numpywren: serverless linear algebra, 2018.
41. J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms, 2012.
42. Spark on Lambda. https://github.com/qubole/spark-on-lambda/.
43. TPC-DS. http://www.tpc.org/tpcds/.
44. R. B. Uriarte, F. Tiezzi, and S. A. Tsaftaris. Supporting autonomic management of clouds: Service clustering with random forest. *IEEE Transactions on Network and Service Management*, 13(3):595–607, 2016.
45. S. Venkataraman, Z. Yang, M. Franklin, B. Recht, and I. Stoica. Ernest: Efficient performance prediction for large-scale advanced analytics. In *Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation*, NSDI'16, pages 363–378, Berkeley, CA, USA, 2016. USENIX Association.
46. R. Viswanathan, G. Ananthanarayanan, and A. Akella. CLARINET: wan-aware optimization for analytics queries. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016.*, pages 435–450, 2016.

47. A. Vulimiri, C. Curino, P. B. Godfrey, T. Jungblut, K. Karanasos, J. Padhye, and G. Varghese. Wanalytics: Geo-distributed analytics for a data intensive world. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, page 1087–1092, New York, NY, USA, 2015. Association for Computing Machinery.
48. A. Vulimiri, C. Curino, P. B. Godfrey, T. Jungblut, J. Padhye, and G. Varghese. Global analytics in the face of bandwidth and regulatory constraints. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*, pages 323–336, Oakland, CA, May 2015. USENIX Association.
49. M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, page 10, USA, 2010. USENIX Association.
50. C. Zhang, M. Yu, W. Wang, and F. Yan. Mark: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, pages 1049–1062, Renton, WA, July 2019. USENIX Association.

University of Nebraska Omaha

Dr. Kwangsung Oh
PKI 283B
Omaha, Nebraska 68182
Phone: 402.554.2862
kwangsungoh@unomaha.edu

Department of Computer Science
College of Information Science and Technology

November 11, 2021

Subject: Anshuman Das Mohapatra GRACA Award

Dear Member of the GRACA Grant Selection Committee:

I am writing this letter for Anshuman Das Mohapatra to express my enthusiastic support for his GRACA award in 2022. I have started working with Anshuman from the Fall/2021 semester as a research advisor for his masters thesis. I recommend Anshuman for a GRACA because he has an innovative research question, and I strongly believe he has a passion and a potential for the research project.

I have been impressed by his research progress within a short period of time. Anshuman has tried to model and solve a research problem for predicting data analytics workloads in a single data center (DC) setting. I believe that he can achieve publishable results soon. He has proven himself as a strongly motivated researcher with strong research background knowledge and is perpetually eager to learn research topics and pursue new ideas.

Anshuman's research will be concerned with challenges in geo-distributed data analytics (GDA) for mining valuable information from geo-distributed data, one of the most essential workloads for many Internet applications such as Netflix, Facebook, and Airbnb that run in a multi-cloud environment. In this project, he will primarily focus on the research question: *"How to predict diverse GDA applications' workloads to determine optimal cloud compute resources configuration in a multi-cloud environment."* as a part of our ongoing project titled, "Scalable Cloud Resources Management for Geo-distributed Data Analytics in a Multi-cloud Environment." The proposed *workload prediction* techniques will significantly impact the overall performance and cost ($), important metrics for GDA systems running on a multi-cloud environment. To validate proposed approaches, he will conduct rigorous experiments on both simulated and real cloud environments. His research will allow GDA systems to determine optimal cloud configurations that exploit diverse and heterogeneous cloud compute resources to maximize benefits from a multi-cloud environment.

To finish the proposed project successfully, we will meet weekly during 2022 the summer and fall months. I will provide technical guidance for building our system, deploying the system on both simulated and real cloud environments, and conducting extensive experiments. For Anshuman's MS thesis, I will advise him to build a prediction model and integrate the model into our GDA system.

In conclusion, my endorsement for Anshuman's GRACA grant application is strong and without any reservations. The proposed research will be an essential part of his MS thesis and future research as a Ph.D. student. He will apply to the doctoral program to pursue the PhD degree under my supervision at UNO from Fall/2022. I will provide financial support to him for at least two years and further based on the availability of funds. This grant is essential for him as the financial aid will allow him to work on his research during the summer of 2022 when he is transitioning from the MS program to the Ph.D. program.

Anshuman is very open-minded, creative, energetic, and optimistic. I am grateful for the opportunity to help him pursue this research. I strongly urge you to consider awarding him this GRACA grant for the proposed research. Thank you for your time and consideration of his proposal. I look forward to the exciting opportunities this funding will bring to him. If you have any questions, please feel free to contact me in person.

Sincerely,
Kwangsung Oh, Ph.D.
Assistant Professor, Department of Computer Science
College of Information Science and Technology