Hindawi
Security and Communication Networks
Volume 2018, Article ID 1746809, 14 pages
https://doi.org/10.1155/2018/1746809

Research Article

International Network Performance and Security Testing Based on Distributed Abyss Storage Cluster and Draft of Data Lake Framework

ByungRae Cha,[1] **Sun Park**,[1] **JongWon Kim**,[1] **SungBum Pan**,[2] **and JuHyun Shin**[3]

[1]*School of Electrical Engineering and Computer Science, GIST, Gwangju, Republic of Korea*
[2]*Department of Electronic Engineering, Chosun University, Gwangju, Republic of Korea*
[3]*Department of ICT Convergences, Chosun University, Gwangju, Republic of Korea*

Correspondence should be addressed to JuHyun Shin; jhshinkr@chosun.ac.kr

Received 18 September 2017; Revised 7 December 2017; Accepted 1 January 2018; Published 18 February 2018

Academic Editor: Ilsun You

The megatrends and Industry 4.0 in ICT (Information Communication & Technology) are concentrated in IoT (Internet of Things), BigData, CPS (Cyber Physical System), and AI (Artificial Intelligence). These megatrends do not operate independently, and mass storage technology is essential as large computing technology is needed in the background to support them. In order to evaluate the performance of high-capacity storage based on open source Ceph, we carry out the network performance test of Abyss storage with domestic and overseas sites using KOREN (Korea Advanced Research Network). And storage media and network bonding are tested to evaluate the performance of the storage itself. Additionally, the security test is demonstrated by Cuckoo sandbox and Yara malware detection among Abyss storage cluster and oversea sites. Lastly, we have proposed the draft design of Data Lake framework in order to solve garbage dump problem.

1. Introduction

Most new technologies improve product performance, and these technologies are called persistent technologies. Persistent technologies can be either disconnected or radical because of their nature, but many of them have a gradual character. Sometimes destructive technologies arise, and innovative technologies bring to market a value proposition that is quite different from what was used in the past. The trend of software field specially had been much changed in aspect of software development process, application architecture, deployment and package, and application infrastructure as shown in Figure 1. In this paper, we intend to develop high-capacity, distributed storage, and network acceleration technologies based on open source to gain the opportunity of transition and growth from existing storage technologies of existing leading companies to innovative storage technologies for emerging small and medium enterprises. For this, we are using open source Ceph [1–3]. Ceph is an open source project launched to implement large SDS (Software-Defined Storage) [4] using Intel processor-based general purpose H/W. SDS creates a virtualized network of storage resources by separating the control and management software from the underlying hardware infrastructure. This can be used to create storage networks that may tie together large pools of storage resources that can appear as one virtual entity. Thus, costs and limitations and storage services are possible without a monopoly.

For industrial IoT, the emerging Data Lake concept is proposing to turn things upside down for enterprise; instead of defining a database structure first and then populating it with data that fits into this structure, the Data Lake simply stores any and all kinds of data and then makes this data available when it is needed, in whatever format is needed. Also, we want to be able to keep this data for a longer time, in order to perform long-term pattern analysis. Data Lake repository can be queried on an ad hoc basis, along with a data refinery [5]. The concept of a Data Lake has evolved over time in enterprises, starting with concepts of data warehouse which contained data for long-term retention and

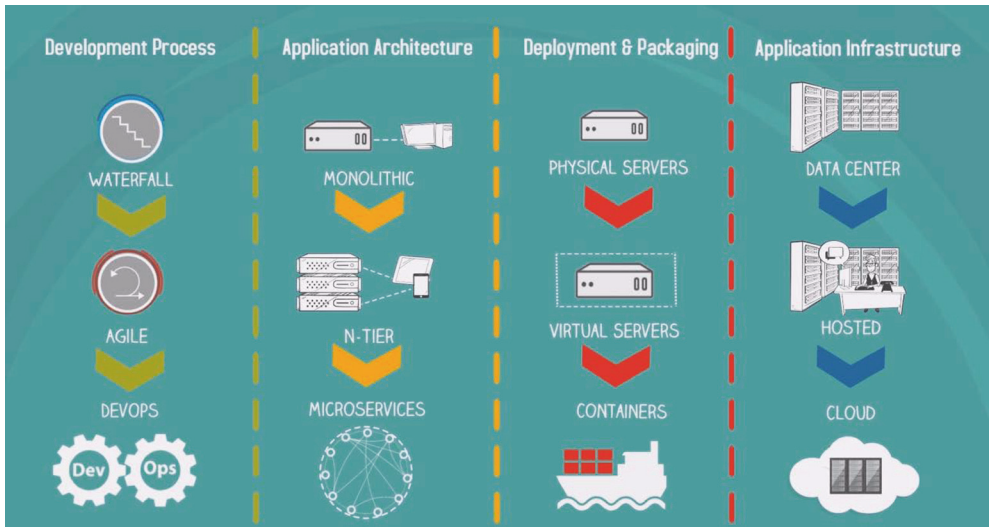


Figure 1: Trend changings in aspect of SW.

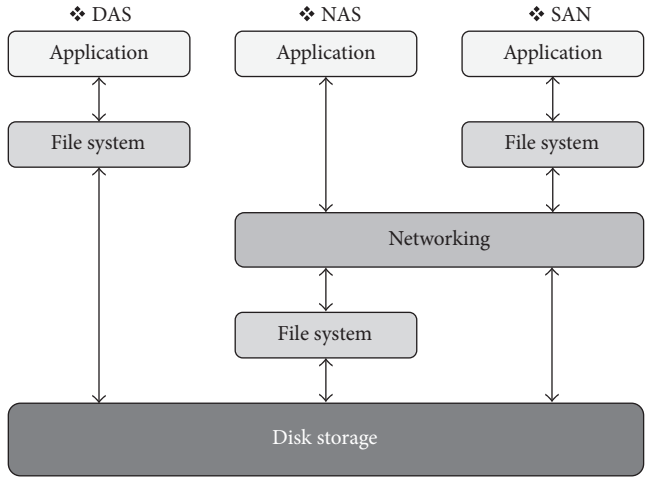


Figure 2: Classification of DAS, NAS, and SAN.

stored differently for reporting and historic needs. Data Lake evolved with these concepts as a central data repository for an enterprise that could capture data as us, produce processed data, and serve the most relevant enterprise information. Data Lake can be defined as a vast repository of a variety of enterprise-wide, raw information that can be acquired, processed, analyzed, and delivered. A Data Lake is expected to be able to derive enterprise-relevant meanings and insights from this information using various analysis and machine learning algorithms [6].

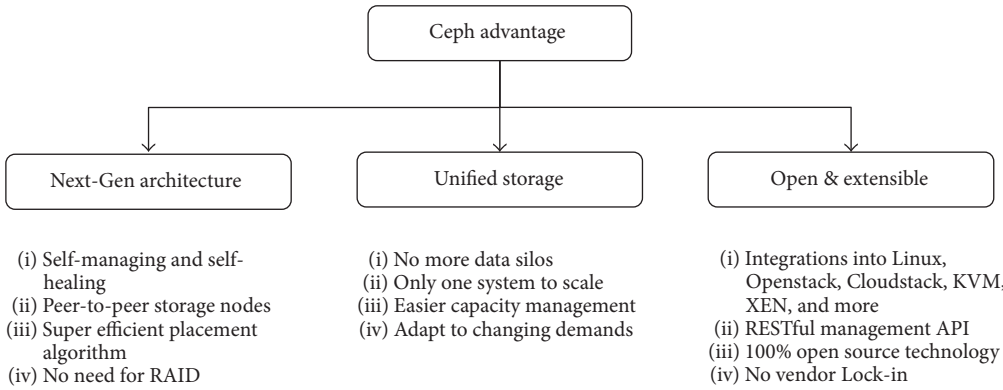
In this paper, using KOREN network, the disk media performance test, network bonding, and network traffic test and security tests of Cuckoo sandbox and Yara malware detection are performed to improve performance and security of mass distributed Abyss storage cluster based on open source Ceph. Lastly, the Data Lake framework using Abyss storage cluster has the potential to become a quite useful foundation for analytical processing. In order to solve the demerit of one-way Data Lake called garbage dump, we have proposed the applying topology and machine learning technology and the draft design of Data Lake framework.

2. Related Work

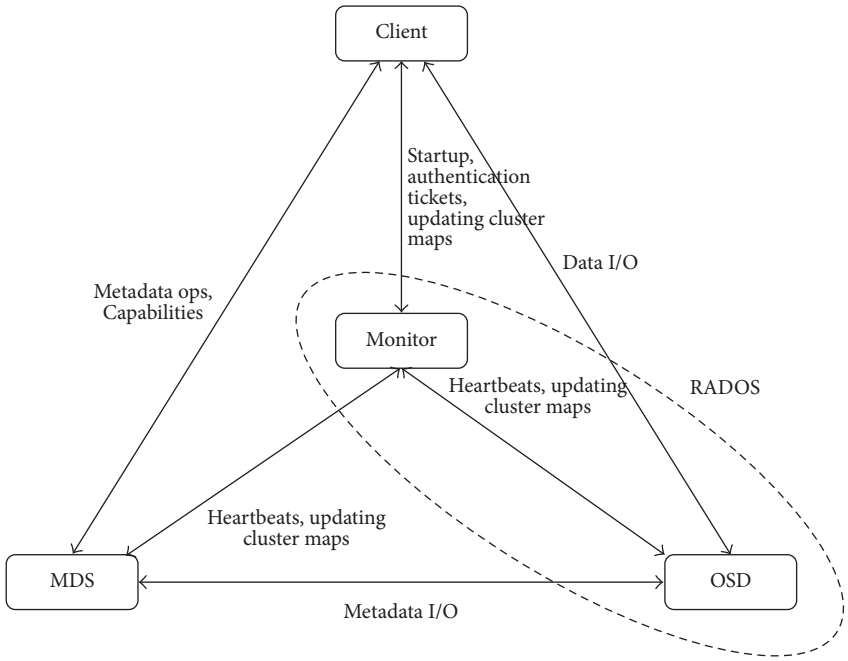
In this section, we briefly described the related terms of open source Ceph [1–3], Data Lake [5–9], and Docker-based security tools of Cuckoo sandbox [10] and Yara malware detection [11, 12].

First, before describing open source Ceph, the techniques of storage are divided into DAS (Direct Attached Storage), NAS (Network Attached Storage), and SAN (Storage Area Network) as shown in Figure 2. Open source Ceph for mass storage is a SAN and Linux distributed file system of a petabyte scale and started with a doctoral research project on Sage Weil's storage systems at UCSC (University of California, Santa Cruz). The advantages of open source Ceph include the next-generation architecture, integrated storage, and openness and scalability, as shown in Figure 3 (advantage of Ceph).

❖ Advantages of Ceph



❖ Component interactions of Ceph



❖ Triple-copy & self-healing by Ceph

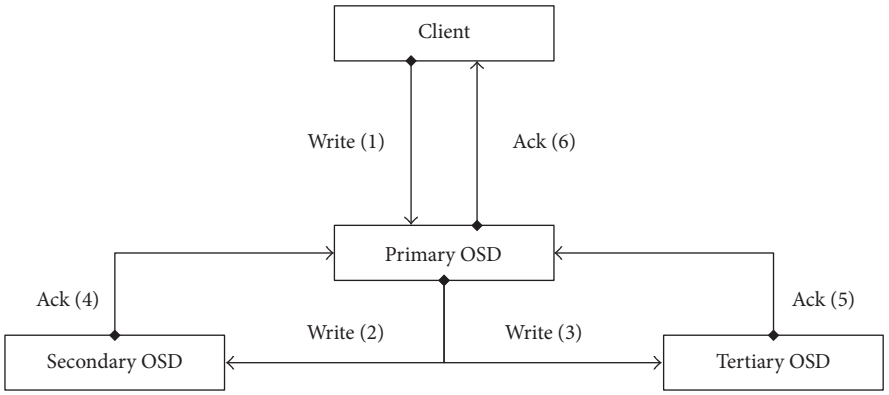


Figure 3: Advantages, architecture and interaction, and triplication of open source Ceph.

Table 1: Key differences between data warehouse and Data Lake.

Items	Data warehouse	Data Lake
Data	Structured, processed	Structured/semistructured/unstructured, raw
Processing	Schema-on-write	Schema-on-read
storage	Expensive for large data volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, configure and reconfigure as needed
Security	Mature	Maturing
users	Business professionals	Data scientists, etc.

The architecture and interaction with open source Ceph system can be expressed as shown in Figure 3 (component interaction of Ceph), and Figure 3 (triple-copy and self-healing by Ceph) briefly shows the Ceph self-healing and data triplication procedures. Thus, open source Ceph comes with a built-in benchmarking SW tool called the RADOS (Reliable Autonomic Distributed Object Store) bench, which can be used to measure the performance of Ceph clusters at the resource pool level [1–3].

Second, a Data Lake refers to a massively scalable storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a Data Lake uses a flat architecture to store data [5–9].

Lastly, Yara is a tool aimed at helping malware researchers to identify and classify malwares. With Yara you can create descriptions of malware families (or whatever you want to describe) based on textual or binary patterns. Each description, a.k.a. rule, consists of a set of strings and a Boolean expression which determine its logic. And Cuckoo sandbox is a malware analysis system. In other words, admin can throw any suspicious file at Cuckoo sandbox and in a matter of seconds Cuckoo will provide you back with some detailed results outlining what such file did when executed inside an isolated environment.

3. Development of Abyss Storage Cluster Prototype and Draft Design Data Lake Framework

In order to design and expand Data Lake framework, first, we designed and developed the prototype of Abyss storage cluster H/W using open source Ceph based on Ubuntu Server.

3.1. Development of Abyss Storage Cluster Prototype. Abyss storage cluster prototype has been designed and developed as shown in Figures 4 and 5 [13]. The logical components of the mass volume Abyss storage cluster for SMB (Small and Medium Business) and Docker-based security tools are shown in Figure 4. And there is H/W prototype of the Abyss storage, 3D rendering image, and the 3D printing material of the product case of Abyss storage as shown in Figure 5, actually.

3.2. Docker-Based Security Tools (Cuckoo and Yara) on Abyss Storage Cluster. Abyss storage cluster supports the Docker of virtualization technology and Cuckoo and Yara of security tools as shown in ① and ② of Figure 4. Docker is lightweight virtualization technology, an open platform for developers, and sysadmins to build, ship, and run distributed applications, whether on laptops, data center VMs (Virtual Machines), or the cloud [14, 15]. The Linux kernel's support for namespaces mostly isolates an application's view of the operating environment, including process trees, network, user IDs, and mounted file systems, while the cgroups of kernel provide resource limiting, including the CPU, memory, block I/O, and network. To use virtualization technology, Docker-based security tools based on Abyss storage cluster are Cuckoo sandbox and Yara malware detection. Cuckoo sandbox is an application that provides a virtual sandbox for the automatic analysis of malware specimens. Another powerful feature of Cuckoo sandbox is the ability to utilize the Yara framework. Yara provides a rule-based approach to create descriptions of malware families based on textual or binary patterns. A description is essentially a Yara rule name, where these rules consist of sets of strings and a Boolean expression. Docker-based security tools in Abyss storage cluster have been operated and verified among domestic sites and oversea sites.

3.3. Draft Design of Data Lake Framework. A Data Lake is a scalable storage repository that is an advanced version of the data warehouse and data silo in the BigData era. There are some differences between Data Lake and data warehouse in aspect of data, processing, storage, agility, security, and users as shown in Table 1 [16]. It is important to recognize that while both the data warehouse and Data Lake are storage repositories, the Data Lake is not data warehouse 2.0 nor is it a replacement for the data warehouse.

And a data silo is a repository of fixed data that remains under the control of one department and is isolated from the rest of the organization; much like grain in a farm silo is closed off from outside elements. Data silos can have technical or cultural roots. Data silos tend to arise naturally in large organizations because each organizational unit has different goals, priorities, and responsibilities. Data silos can also occur when departments compete with each other instead of working with each other towards common business goals. Information silos are generally viewed as a hindrance to effective business operations and organizations are increasingly trying to break down silos that are a barrier to collaboration, accessibility, and efficiency [17].

There are many reasons for users to be frustrated with the information pooling in their Data Lakes. The core issue was that the larger the information lake grew, the more difficult analyzing the data became. A Data Lake of any significant

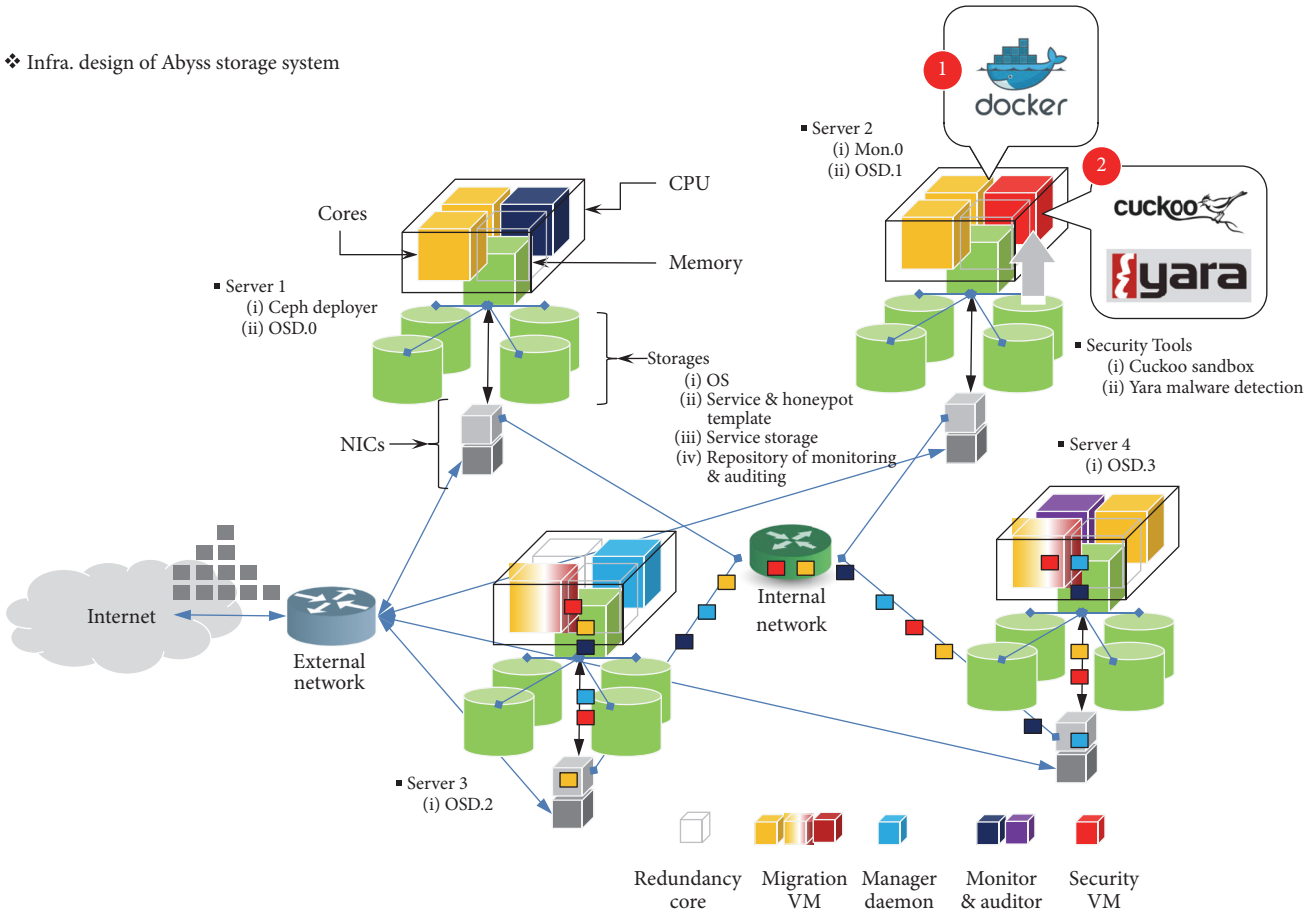


Figure 4: Concept diagram of mass distributed Abyss storage cluster and Docker-based security tools.

Figure 5: H/W prototype and 3D printing product of Abyss storage cluster.

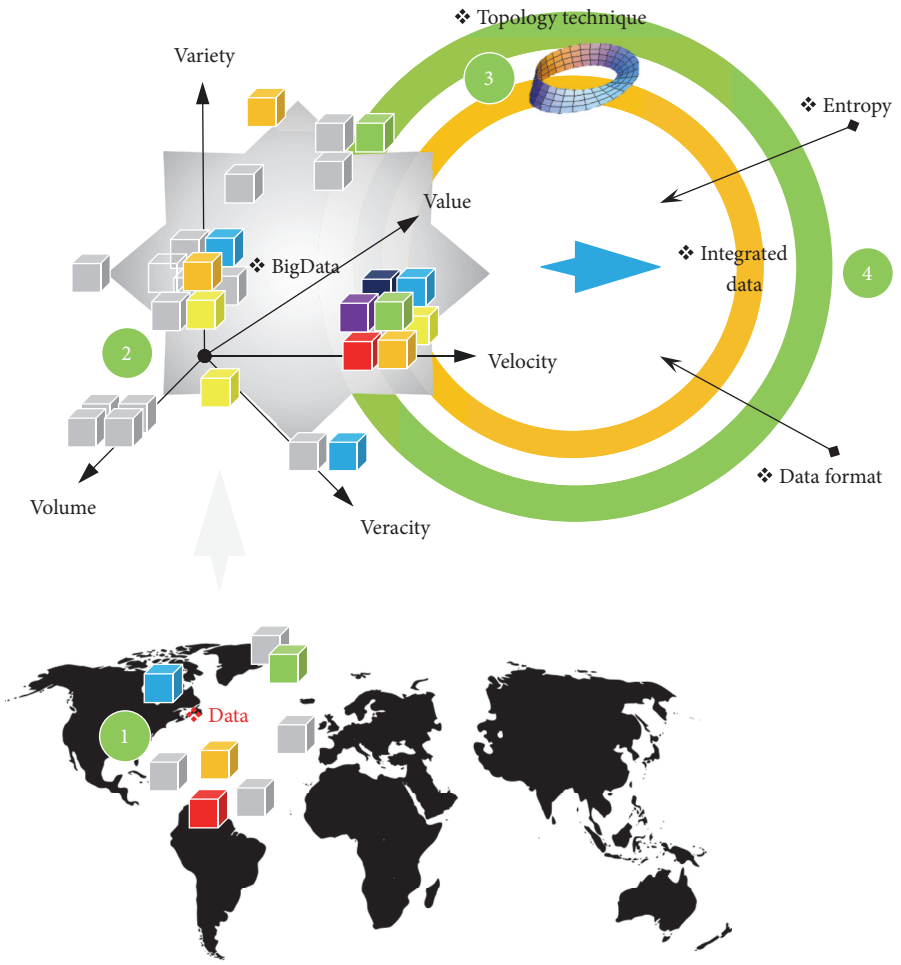


Figure 6: Concept of applying mathematics topology for BigData.

size was often dubbed a "one-way lake," since data is eternally pouring in, but data and any analysis are never taken out, or even accessed once the data is placed inside the Data Lake.

There is one reason why the Data Lake turns into a "one-way" Data Lake. But this issue traces its roots to how data was placed into the Data Lake in the first place: the intent was never to organize the data for future usage. Instead the Data Lake became a place just to "dump" data. So much effort was spent on gathering data from every possible source that few engineers and companies gave much thought to organizing the data for future usage. The Data Lake has the potential to become a quite useful foundation for analytical processing. In order to solve the demerit of one-way Data Lake called garbage dump, we have proposed the applying mathematics topology and machine learning technology and the draft design of Data Lake framework in Figures 6 and 7. We specially proposed applying the mathematics topology and machine learning technology in Data Lake framework as shown in ① and ② of Figure 7.

Mathematics topology [18–21] is concerned with the properties of space that are preserved under continuous deformations. This can be studied by considering a collection of subsets, called open sets, which satisfy certain properties, turning the given set into what is known as a topological space. Important topological properties include connectedness and compactness. The initial motivation is to study the shape of data. TDA (Topology Data Analysis) has combined algebraic topology and other tools from pure mathematics to allow mathematically rigorous study of "shape." The main tool is persistent homology, an adaptation of homology to point cloud data. Persistent homology has been applied to many types of data across many fields. Moreover, its mathematical foundation is also of theoretical importance. The unique features of TDA make it a promising bridge between topology and geometry.

Furthermore, ML (machine learning) is the subfield of computer science that evolved from the study of pattern recognition and computational learning theory in AI and ML explores the study and construction of algorithms that can learn from and make predictions on data; such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. ML is employed in a range of computing tasks where designing and programming explicit algorithms with good performance are difficult or infeasible. Thus, iML is closely related to computational statistics,

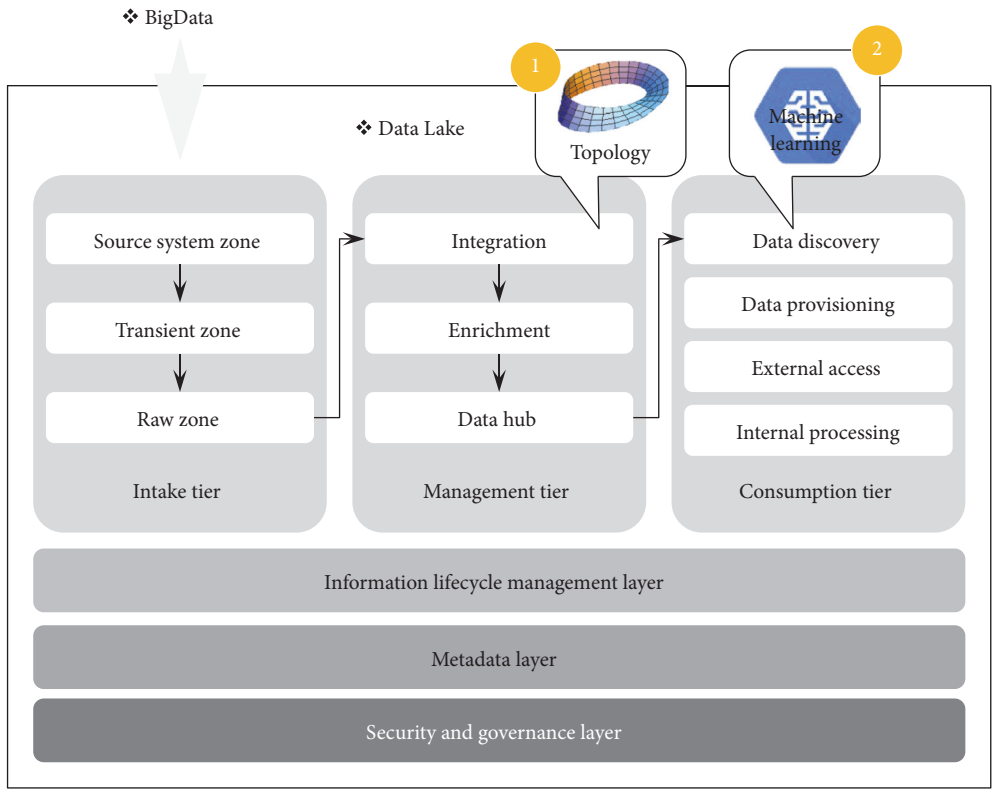


Figure 7: Draft design of Data Lake framework.

which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory, and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. ML can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies [22]. And recently, variety of open source based ML tools (TensorFlow [23], Caffe [24], Torch7 [25], Cuda-convert [26], Chainer [27], and MXNet [28]) have been provided due to the influence of Deep Learning [29] neural network.

4. Performance and Security Testing of Abyss Storage Cluster Prototype

In order to improve performance of Abyss storage, we have tested performance of storage media by disk types, network bonding for acceleration of internal network of Abyss storage cluster, and international network performance test using KOREN [30].

4.1. Disk Media Test of Abyss Storage. The disk media tests of Abyss storage servers were performed by disk media types (HDD (Hard Disk Drive), SSHD (Solid State Hybrid Drive), and SSD (Solid State Drive)). Using RADOS Bench S/W, we performed the read and write operation of disk media types for 10 seconds and recorded the average of IOPS (Input/Output Operations per Second) using RADOS Bench SW as shown in Figure 8 [13].

4.2. Network Acceleration by Bonding. For network acceleration, the internal network of Abyss storage cluster is bonded with two 1 GB switches into VLAN (Virtual Local Area Network) as shown in Figure 9. We have performed the tests of write, sequence read, and random read operation between general network and network bonding, and Figure 10 shows the average IOPS comparison of test results. Test results of the system with network bonding for network acceleration were improved by at least 170% more than general network system [13].

5. Network Performance and Docker-Based Security Test of Abyss Storage Cluster Prototype

5.1. Environments of Network Performance and Security Test. We performed network performance test through uploading and downloading of multimedia data among GIST in Korea, Myren in Malaysia, Thailand, and Philippine sites. Figure 11 presents the test-bed environment using KOREN for real network performance and security test of Cuckoo sandbox and Yara malware detection among domestic sites and oversea sites.

5.2. Domestic Network Performance Test. The domestic network performance test has been performed between Abyss

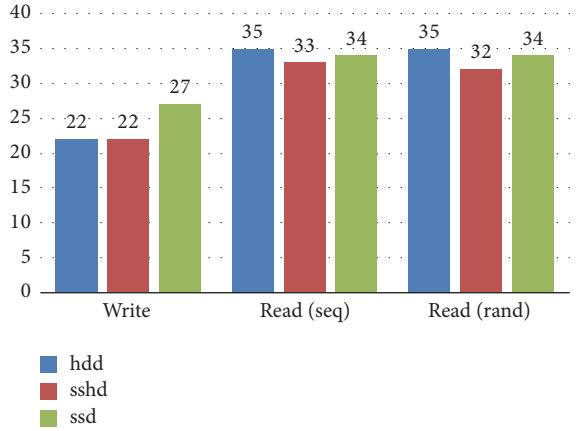


Figure 8: IOPS comparison of Abyss storage by disk media types.

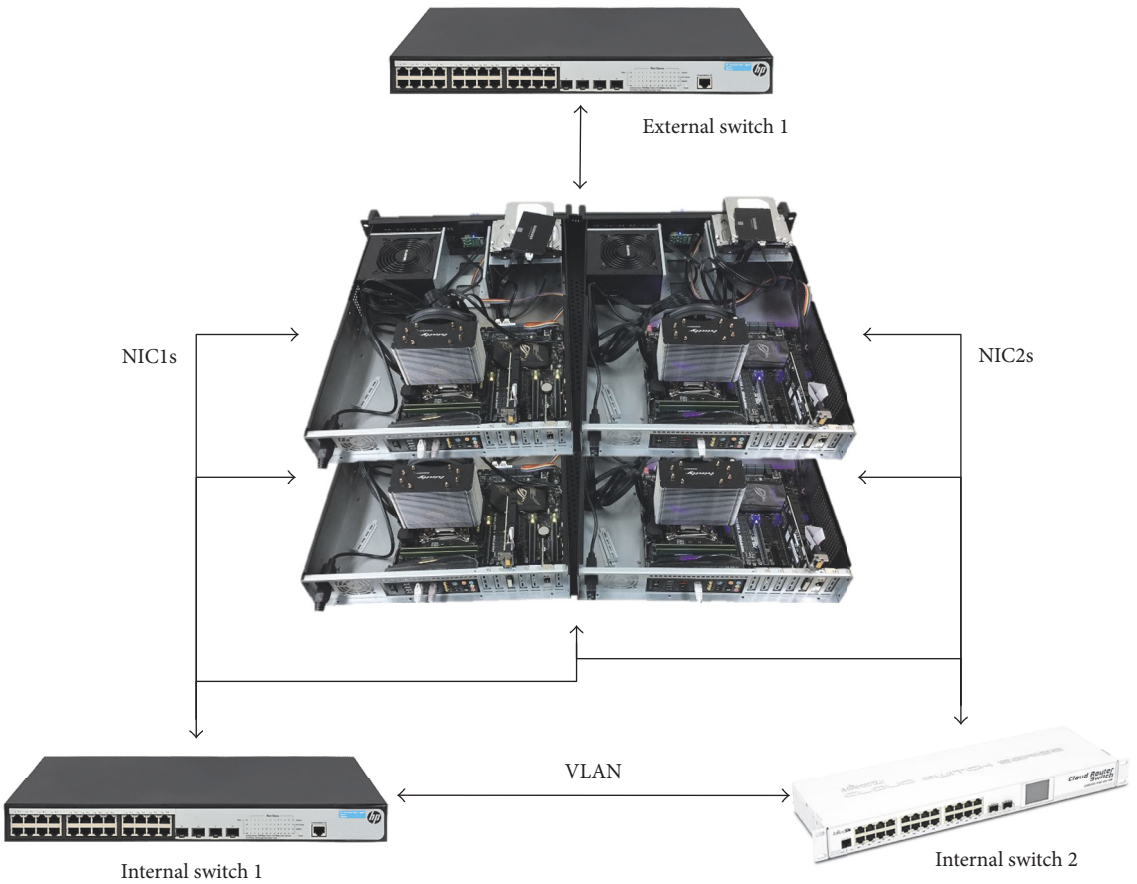


Figure 9: Network bonding of Abyss storage cluster.

storage cluster in GIST and another GIST site. And we measured the external network traffic of Abyss storage and the internal network traffic inside Abyss storage cluster using SpeedoMeter [31], a network real-time monitoring tool. Figure 12 shows test results of file uploading speed, and Figure 13 presents external and internal traffics on Abyss storage cluster in GIST during file uploading. Conversely, Figure 14 shows test results of file downloading speed, and Figure 15 presents external and internal traffics on Abyss storage cluster during file downloading. The comparison of uploading and downloading network traffic between domestic sites is depicted in Figure 16. Although it will be recognized later, the domestic network performance is higher than network performances among oversea sites.

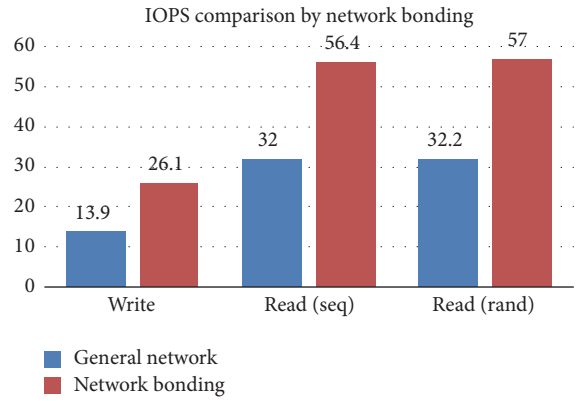


Figure 10: IOPS comparison between general network and bonding network.

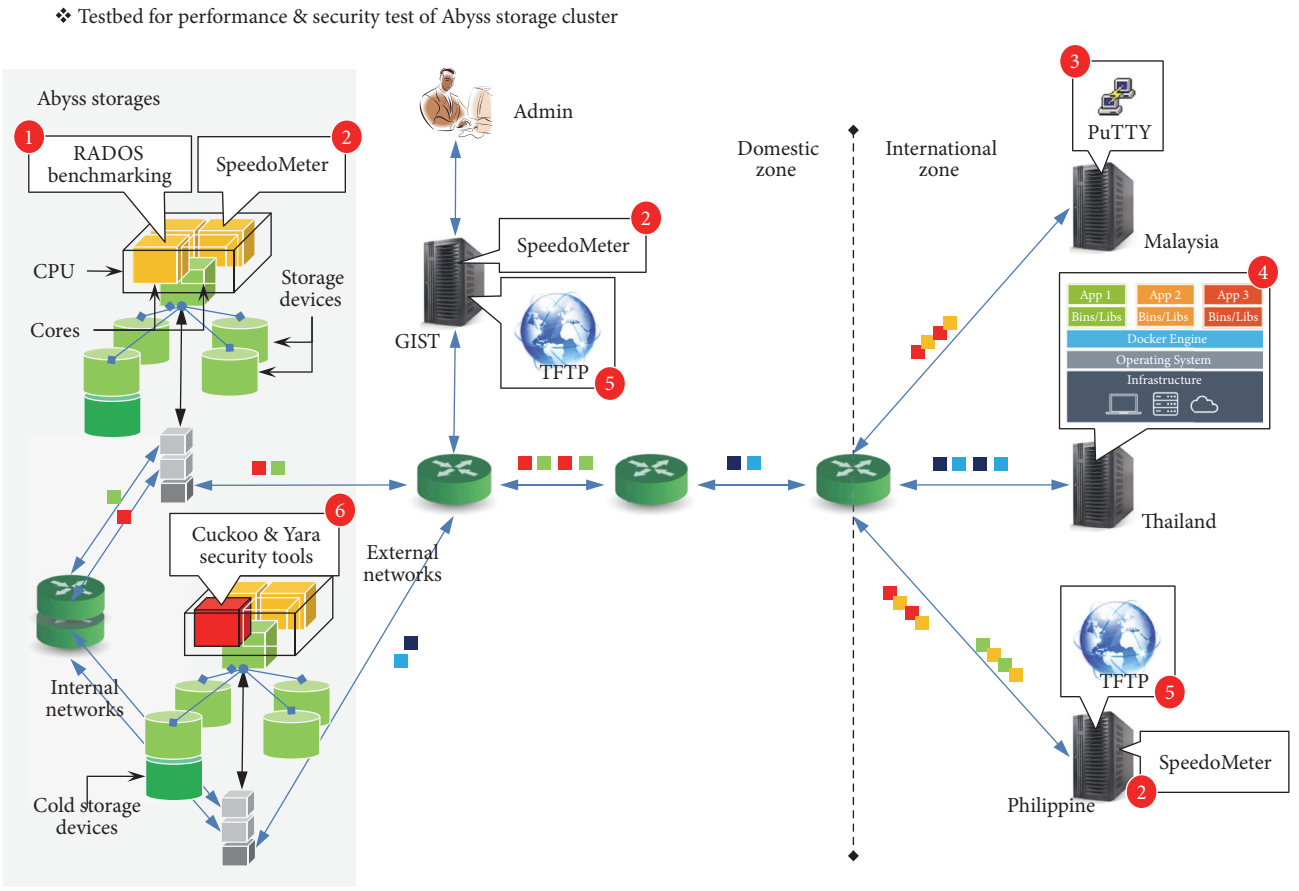


Figure 11: Testbed of Abyss storage cluster prototype for network performance test and security tools test of Cuckoo and Yara using KOREN network.

5.3. International Network Performance and Docker-Based Security Test. In this subsection, using KOREN, the international network performance test has been performed among Abyss storage cluster in GIST, Myren in Malaysia, Thailand, and Philippine sites as shown in Figure 11. Figure 17 shows the speeds of uploading and downloading for each file capacity between Myren in Malaysia and the developed Abyss storage cluster in GIST. Comparing the figures specially shows that the variance among download speeds is much higher than the upload speed. Additionally, the video test between Abyss storage cluster in GIST and Myren has been performed as shown in Figure 18. Figures 19 and 20 present speeds of uploading and downloading among GIST, Thailand, and Philippine sites. In international network performance test, the peculiar cases are more unsafe states of network traffic and high variances of network speed on all sites in

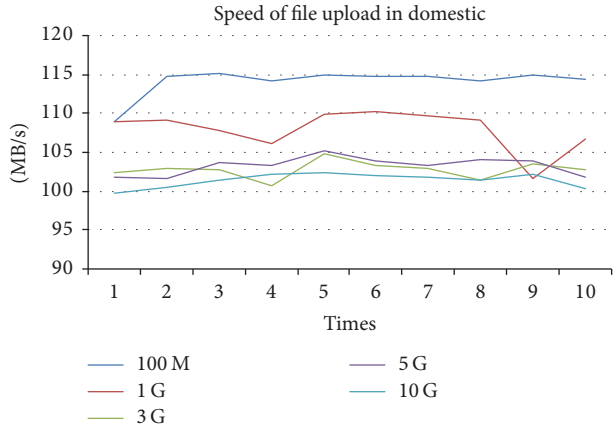
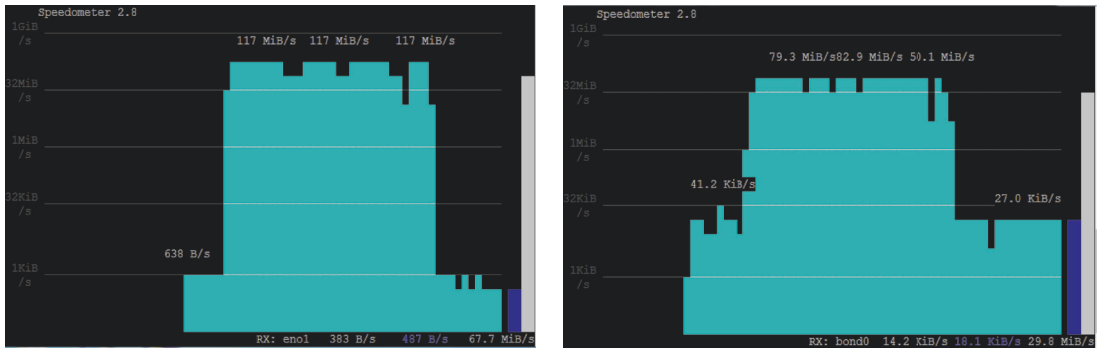


Figure 12: Test result of file uploading speed between Abyss storage cluster and another GIST site.



(a) External traffic of Abyss storage

(b) Internal traffic of Abyss storage cluster

Figure 13: External and internal traffics of Abyss storage cluster during files uploading.

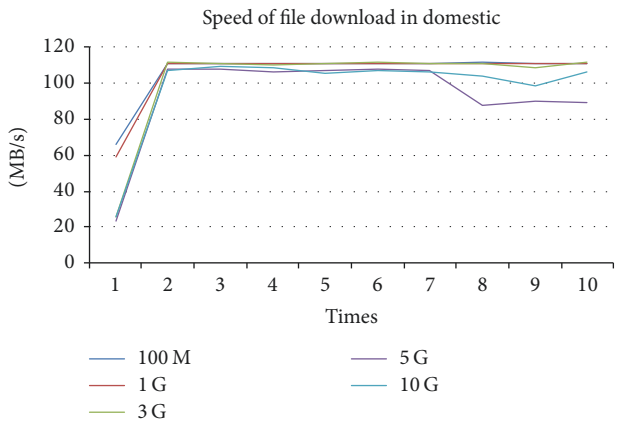
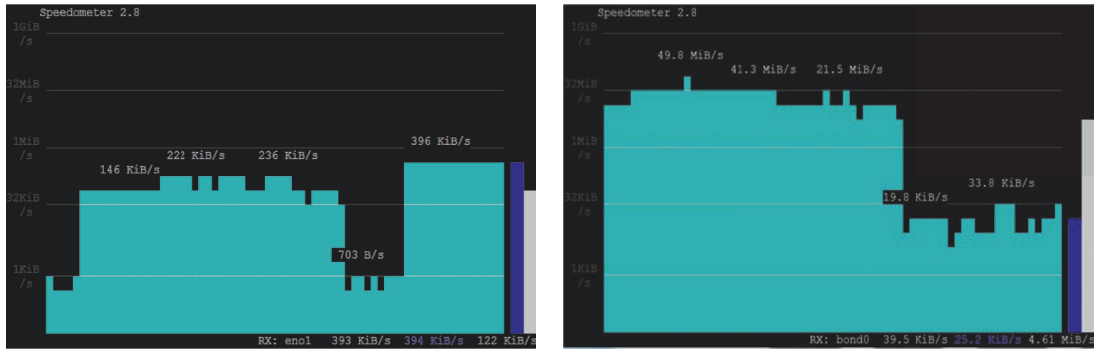


Figure 14: Test result of file downloading speed between Abyss storage cluster and another GIST site.

South-East Asia area than domestic site test. Moreover, the comparisons of upload and download network performance among GIST, Malaysia, Thailand, and Philippine are depicted in Figure 21. Lastly, we have performed the Docker-based security test using Yara malware detection tool among Abyss storage cluster in GIST and oversea sites. This has examined and verified the possibilities of virtualization-based security functions.



(a) External traffic of Abyss storage (b) Internal traffic of Abyss storage cluster

Figure 15: External and internal traffics of Abyss storage cluster during files downloading.

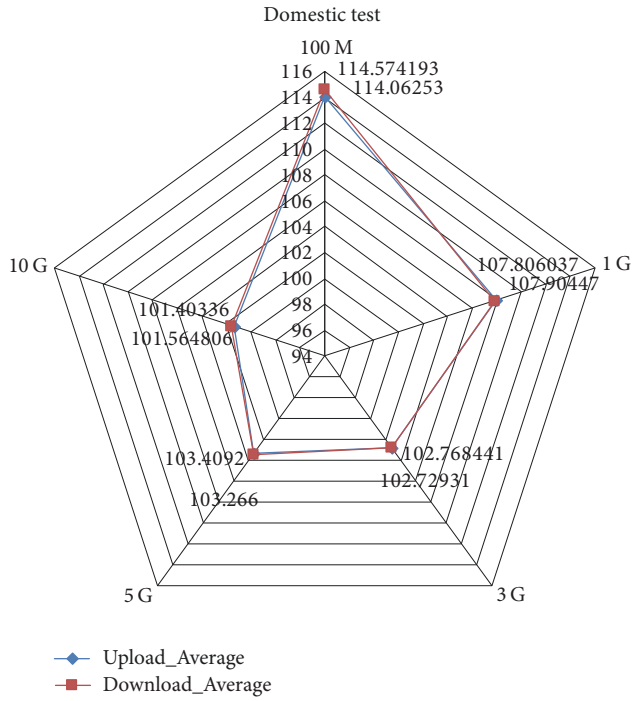


Figure 16: Comparison of uploading and downloading network traffics in domestic sites.

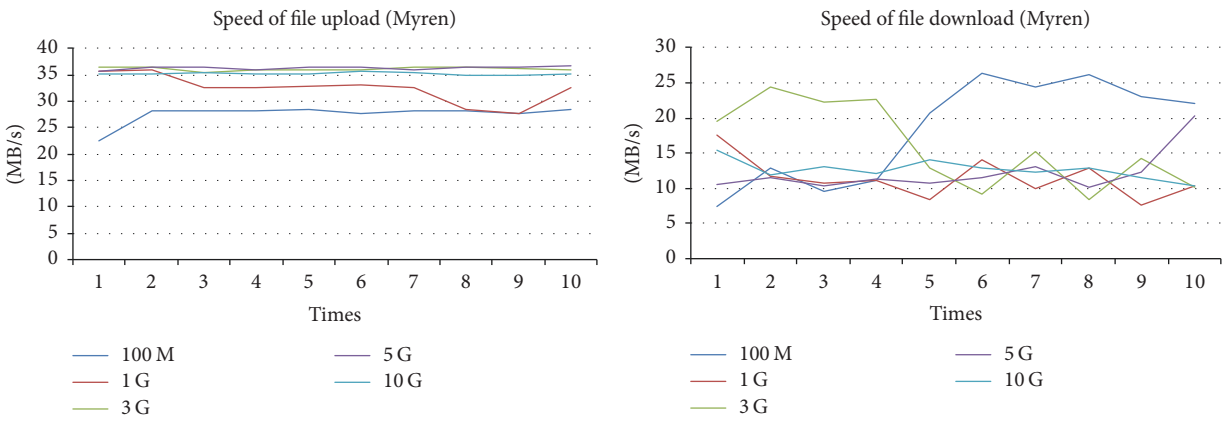


Figure 17: Test result of uploading and downloading speed between Abyss storage cluster in GIST and Myren in Malaysia.

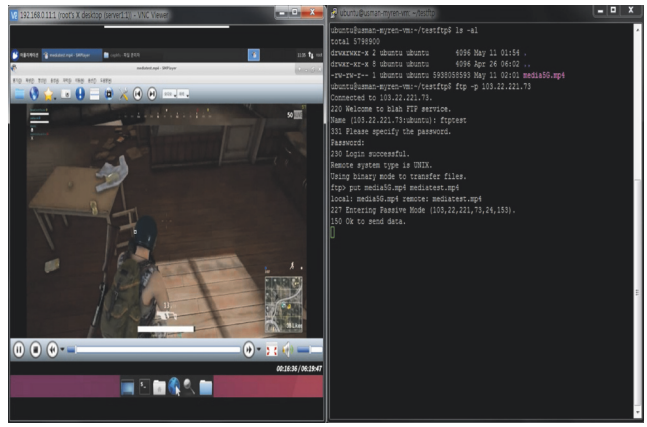


Figure 18: Video test between Abyss storage cluster in GIST and Myren in Malaysia.

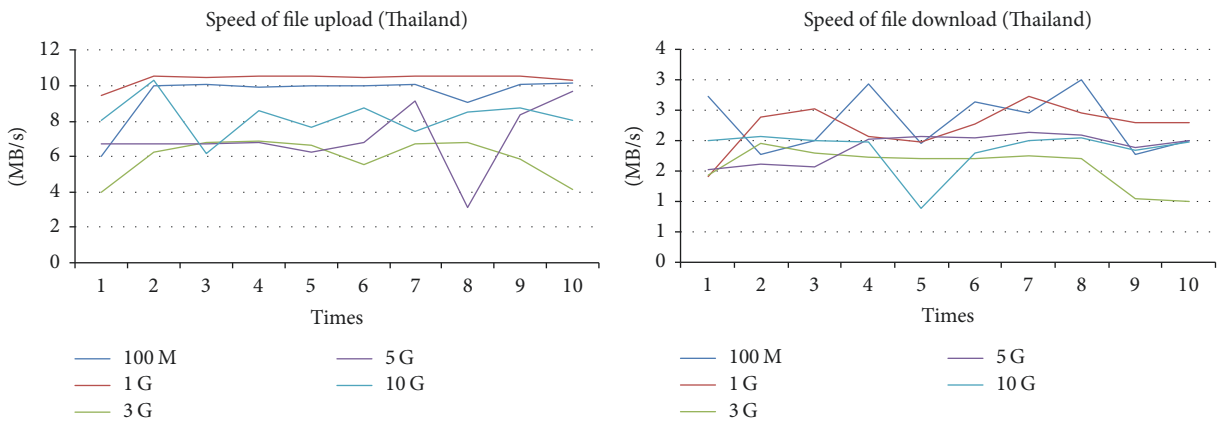


Figure 19: Test result of uploading and downloading speed between Abyss storage cluster in GIST and Thailand.

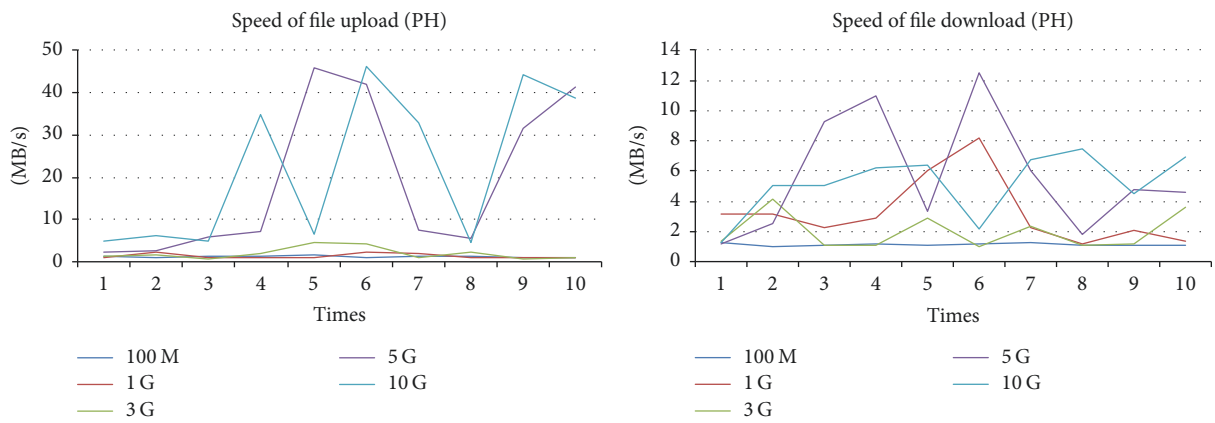


Figure 20: Test results of uploading and downloading speed between Abyss storage cluster in GIST and Philippine.

6. Conclusion

In this paper, the performance tests of the developed mass volume distributed Abyss storage cluster and real-world network performance tests using KOREN have been carried out. Based on this, we intend to explore ways to improve performance of Abyss storage cluster. Detailed tests to improve performance include performance testing of read and write operations for each disk media types (HHD, SSHD, SSD) of Abyss storage servers, internal network testing inside Abyss storage cluster by network bonding, and testing of network performance among domestic and oversea sites. Additionally, we have performed the Docker-based security test using Cuckoo sandbox and Yara malware detection tools among

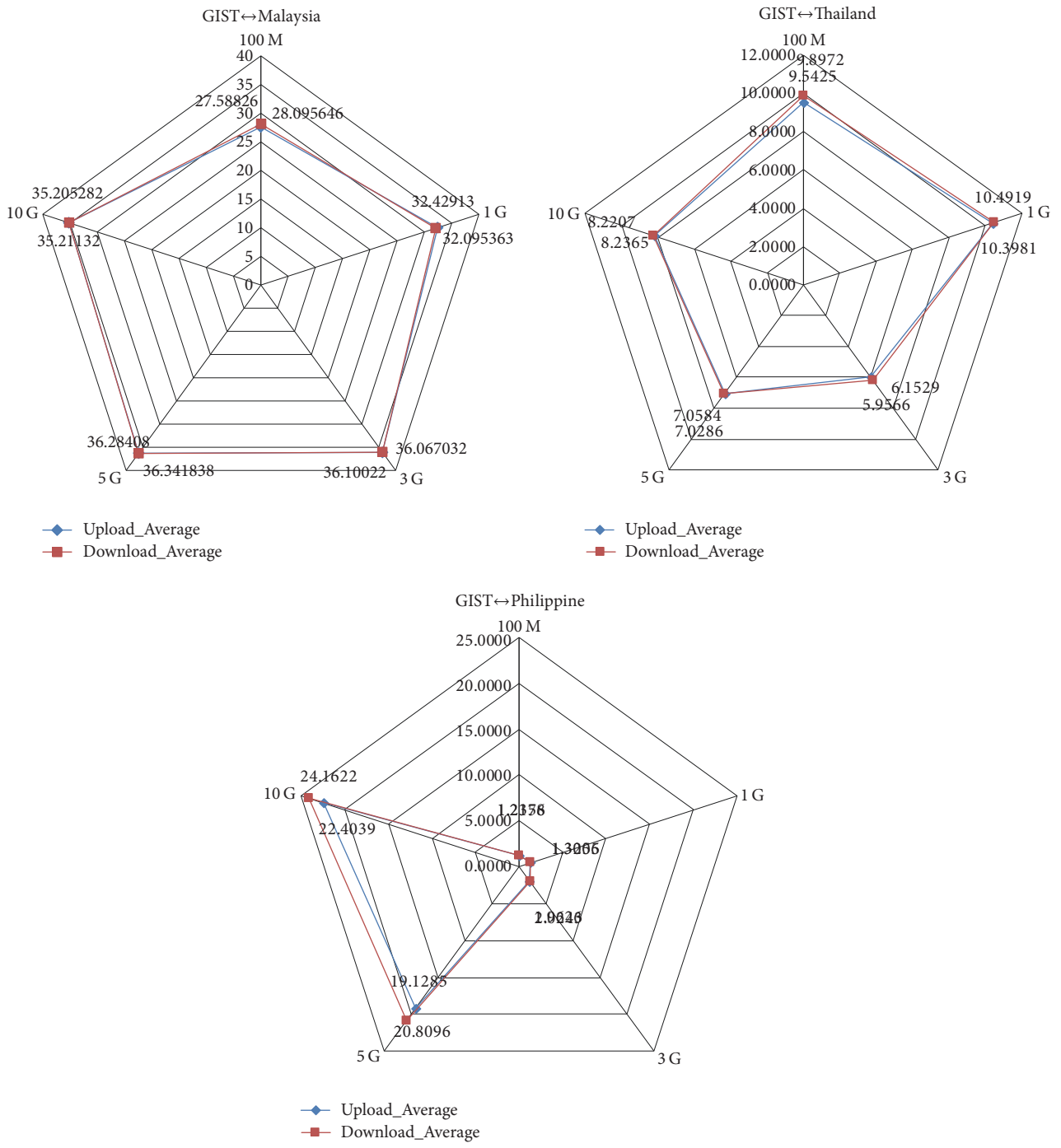


Figure 21: Comparisons of uploading and downloading network traffics among GIST, Malaysia, Thailand, and Philippine.

Abyss storage cluster in GIST and oversea sites. Lastly, we have proposed the draft design of Data Lake framework with mathematics topology and machine learning in order to solve garbage dump problem. In future research of this study, we will focus on MPTCP (MultiPath TCP) [32] for efficient operation of network and Searchable Encryption [33] for data retrieval and security enhancement.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Human Resource Training Program for Regional Innovation and Creativity through

the Ministry of Education and National Research Foundation of Korea (2015H1C1A1035823). And this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2017R1E1A1A03070059).

References

[1] Ceph, http://ceph.com/.

[2] Karan Singh, Learning Ceph,.

[3] K. Singh, *Ceph Cookbook*, 2016.

[4] Software-Defined Storage, https://www.sdxcentral.com/cloud/definitions/what-is-software-defined-storage/.

[5] D. Slama, F. Puhlmann, J. Morrish, and R. M. Bhatnagar, *Enterprise IoT - Strategies & Best Practices for Connected Products Services*, O'Reilly, 2016.

[6] T. John and P. Misra, *Data Lake for Enterprises Leveraging Lambda Architecture for Building Enterpeise Data Lake*, Packt Publishing, 2017.

[7] N. Miloslavskaya and A. Tolstoy, "Big Data, Fast Data and Data Lake Concepts," in *Proceedings of the 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016*, pp. 300–305, USA, July 2016.

[8] P. Pasupuleti, PACKT Publishing, Beulah Salome Purra, Data Lake Development with Big Data, PACKT Publishing, 2015.

[9] B. Inmon, Data Lake Architecture - Designing the Data Lake and Avoiding the Garbage Dump, Technics Publications, 2016.

[10] Cuckoo Sandbox, https://www.cuckoosandbox.org/.

[11] Yara, https://virustotal.github.io/yara/.

[12] D. Ricardo, *Intelligence-Driven Incident Response with YARA*, SANS.

[13] B. Cha, Y. Cha, S. Park, and J. Kim, "Performance testing of mass distributed abyss storage prototype for SMB," *Advances in Intelligent Systems and Computing*, vol. 611, pp. 762–767, 2018.

[14] Docker, https://www.docker.com/.

[15] Docker, https://en.wikipedia.org/wiki/Docker_(software).

[16] Tamara Dull, *Data Lakes vs Data Warehouse: Key Differences*, 2015, http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html.

[17] Data Silo, http://searchcloudapplications.techtarget.com/definition/data-silo.

[18] Topology, https://en.wikipedia.org/wiki/Topology.

[19] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[20] Gunnar Calsoon, Why Topological Data Analysis Works, https://www.ayasdi.com/blog/bigdata/why-topological-data-analysis-works/.

[21] Topological Data Analysis (TDA), https://en.wikipedia.org/wiki/Topological_data_analysis.

[22] Machine Learning, https://en.wikipedia.org/wiki/Machine_learning.

[23] TensorFlow, https://www.tensorflow.org/.

[24] Caffe, http://caffe.berkeleyvision.org/.

[25] Torch7, http://torch.ch/.

[26] Cuda-convert, http://mdtux89.github.io//12/11/torch-tutorial.html.

[27] Chainer, https://chainer.org/.

[28] MXNet, https://mxnet.incubator.apache.org/.

[29] Deep Learning, http://deeplearning.net/.

[30] KOREN, http://www.koren.kr/koren/eng/index.html.

[31] SpeedoMeter, https://hub.docker.com/r/opennsm/speedometer/, ~/ dockerfile/.

[32] Z. Jiaxin, K. Fenfen, Y. Zuo, L. Qinghua, H. Minghe, and C. Yuanlong, "Multi-attribute Aware Path Selection Approach for Efficient MPTCP-based Data Delivery," *Journal of Internet Services and Information Security*, vol. 7, no. 1, 2017.

[33] J. Xiuxiu, G. Xinrui, Y. Jia, K. Fanyu, C. Xiangguo, and H. Rong, "An efficient symmetric searchable encryption scheme for cloud storage," *Journal of Internet Services and Information Security*, vol. 7, no. 2, May 2017.