Design of an Efficient Scientific Research Farm Management System

Duseok Jin, Il-Yeon Yeo, and Jeong-Heon Kim

Abstract—Many modern scientific discoveries would not be possible without the aid of computers and specialized experimental instruments. A computing farm used as an experimental data-based research platform can accelerate scientific research with integrated computing infra-structure and data management tools. This paradigm shift is driving the demand for computing farms commonly used in scientific experiments such as physics, chemistry, biology, and medicine. Computing farm services are becoming more complex and larger as a complementary effort to meet the increasing demands of these diverse research areas, but maintaining sophisticated infra-structures and services with limited staff and budget squeeze is a major pain point for data centers that provide computing farm. Therefore, we propose architecture for intelligent management and automation system that not only allows scientists to easily manage, share and analyze large data sets, but also reduces the complexity of the system and the burden on the computing system administrator.

Research Keywords—Computational science research, Data-intensive experiment, Fast data transmission, Scientific computing farm

1 INTRODUCTION

Without the help of computers and special experimental tools, many modern scientific discoveries are not possible. As a representative example, the European Organization for Nuclear Research [1] (CERN), use the world's largest and most complex scientific instruments and computing resources to study the basic constituents of matter. The Large Hadron Collider [2] (LHC) collision data at CERN was being produced at approximately 25 petabytes per year and then it was being handled thousands of computers and storage systems in over 170 centers across 41 countries. In general, advanced computing facilities and a variety of distributed computing technologies are used to process large amounts of data.

This paper proposes a dynamic platform that supports a variety of experimental farm services, especially in a single data center. The platform provides elastic computing provisioning, advanced monitoring, rapid data transfer, on-demand software distribution, and efficient data management through an integrated scientific research farm management system. Applying the proposed system to the Global Scientific Experimental Data Hub Center (GSDC), which maintains a global data grid computing system for experimental data from advanced research equipment, demonstrates the benefits of efficiency in the management of scientific research farm services.

- *Duseok Jin is with the Korea Institute of Science and Technology Information, Daejeon, 34141, Korea. E-mail: dsjin@kisti.re.kr*
- *Il-Yeon Yeo is with the Korea Institute of Science and Technology Information, Daejeon, 34141, Korea. E-mail: ilyeon9@kisti.re.kr*
- *Jeong-Heon Kim (corresponding author) is with the Korea Institute of Science and Technology Information, Daejeon, 34141, Korea. E-mail: jh.kim@kisti.re.kr*

2 SYSTEM ARCHITECTURE

This section discusses the design of a scientific research farm management system that not only supports efficient experimental farm operations for system managers, but also enhances data transfer performance for scientific researchers. The conceptual architecture of the proposed system is shown in Fig. 1. An efficient platform should be able to easily deploy and customize a research farm to meet user's specific needs. We implement an elastic cluster provisioning system with puppet [3] to easily build a research farm services. It automatically deploys all required essential software and configurations faster than ever, while maintaining quality, security, and compliance.

Our elastic provisioning system consist of five modules: a hardware management script and four puppet profiles (Linux management profile, service management profile, policy management profile, and monitoring management profile).

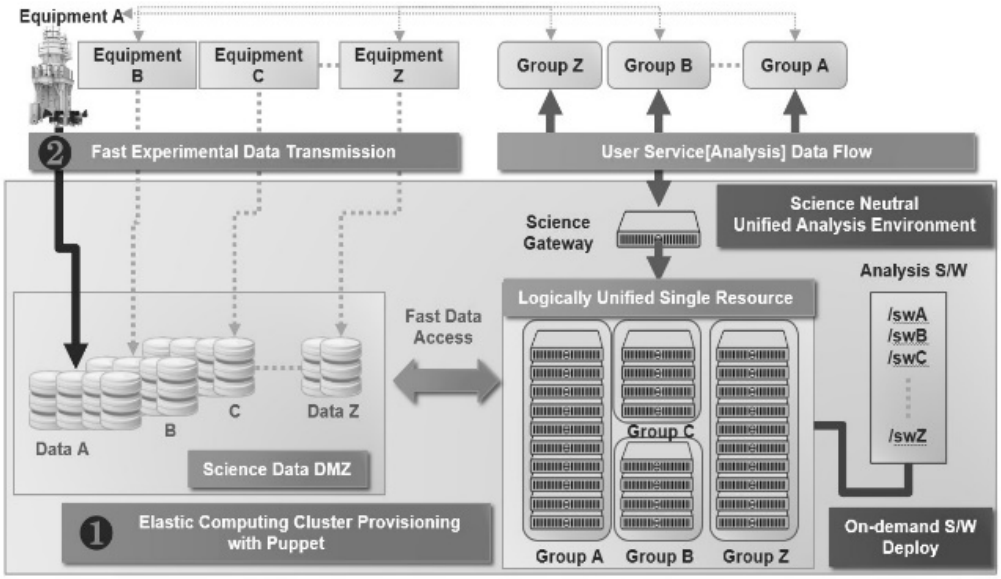

Fig. 1. A conceptual illustration describing the proposed scientific research farm management system that GSDC has developed to support scientific researches.

Each module specifies specific tasks at each step for provisioning and monitoring, such as powering on, installing the operating system, network configuration, deploying middleware software, managing authentication, configuring security polices, monitoring system and services, and automatically recovering from well-known system errors.

We also design a fast data transmission system through science DMZ [4], a scalable network design model for optimizing scientific data transmission. This model is designed to handle the transmission of large scientific data by creating a specific science DMZ

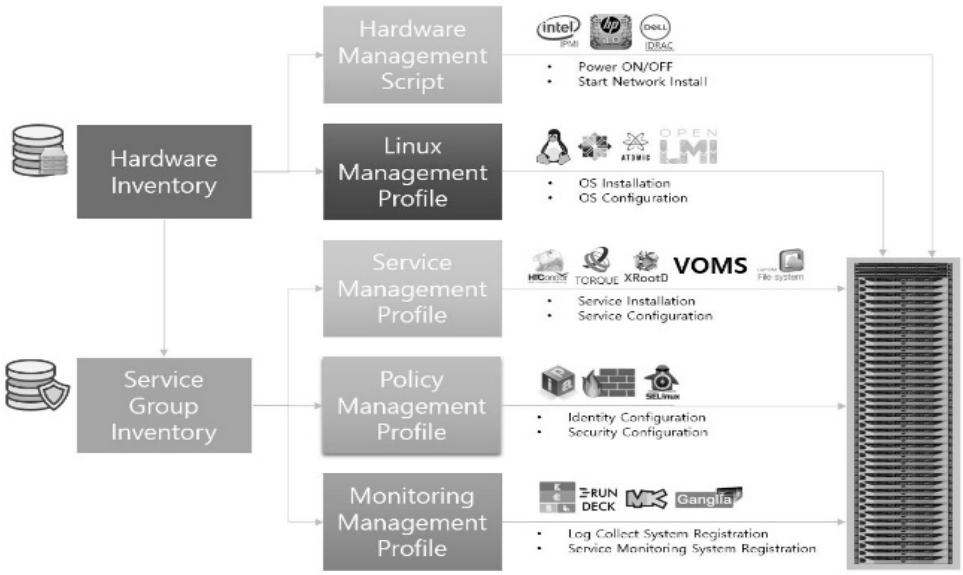

Fig. 2. A structure of the modules of elastic computing cluster provisioning system with puppet.

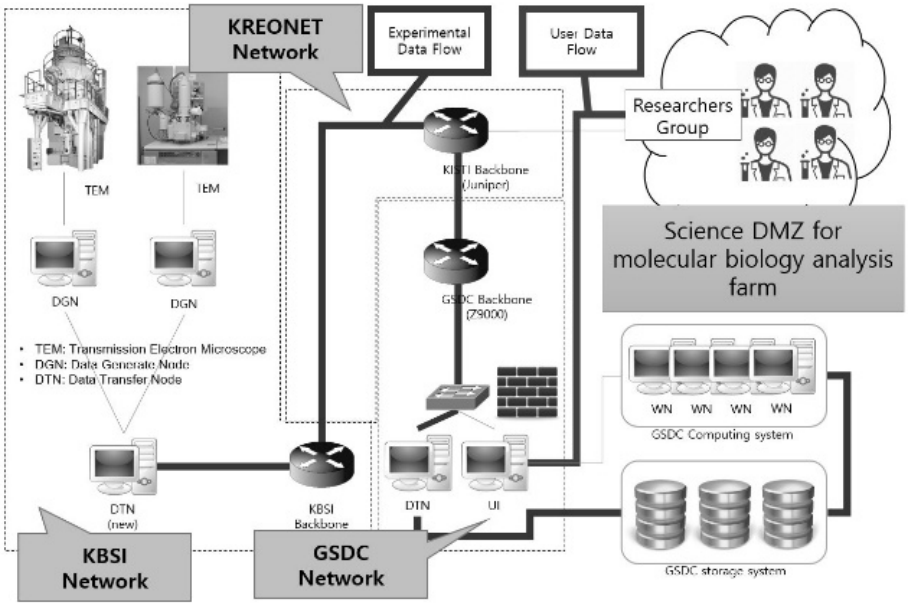

Fig. 3. An example of science DMZ model: The research farm is connected through separated network to the experimental facilities for scientific discoveries.

to accommodate the transmission. By separating the science DMZ network from the general-purpose network, each can be optimized without interfering with the other. The primary component of our system for data transfer is GridFTP [5]. It is a high-performance, secure, reliable data transfer protocol optimized for high bandwidth wide area networks. It is used extensively within large science projects such as the LHC and by many scientific facilities.

3 Sections and Paper Organization

The previous sections described the efficient scientific research farm management system. To evaluate the usefulness and efficiency of proposed platform, we implemented a scientific research farm for molecular biology analysis using Transmission Electron Microscope [6] (TEM). This farm is used for protein structure determination through cryo-electron microscopy (cryo-EM) [7]. Starting the trial service since September 2016, the TEM research farm currently provides about 300 terabytes of data storage on disk, 308 processing cores on 11 nodes and 1Gbps dedicated network connection between the DTN servers, and the farm has been utilized by molecular biology researchers at Korea Advanced Institute of Science and Technology (KAIST) and Korea Basic Science Institute (KBSI). Key technologies and the overall system of the trial service are not limited to molecular biology analysis as the independent structure enables diverse research farms. Therefore, we expect our results to be applied in physics, genomics, astrophysics and domains dependent on computers and specialized experimental instruments.

4 Conclusions

This paper described the architecture for intelligent management and automation system that not only allows scientists to easily manage, share and analyze large data sets, but also reduces the complexity of the system and the burden on the computing system administrator. We have also applied key technologies of the proposed system to a molecular biology analysis farm and are well utilized by users.

We hope this system will be gradually expanded from molecular biology to other domains such as physics, genomics, etc. Through the establishment of a virtuous system with convergence between scientific research and IT technology, we expect this study to lay the foundation for enhancing the S&T competitiveness.

Acknowledgment

This work was supported by the program of the Construction and Operation for Large-scale Science Data Center, 2017, funded by the KISTI (K-17-L01-C05) and by the program of the Global hub for Experiment Data of Basic Science, 2017, funded by the NRF (N-17-NM-CR01).

REFERENCES

[1] CERN | Accelerating science, available at https://home.cern/, Feb. 2017.
[2] LHC, available at https://home.cern/topics/large-hadron-collider, Feb. 2017.
[3] Loope, James. Managing Infrastructure with Puppet: Configuration Management at Scale. "O'Reilly Media, Inc.", 2011.
[4] Dart, Eli, et al. "The science dmz: A network design pattern for data-intensive science." Scientific Programming 22.2 (2014): 173-185.
[5] Allcock, William, et al. "The Globus striped GridFTP framework and server." Proceedings of the 2005 ACM/IEEE conference on Supercomputing. IEEE Computer Society, 2005.
[6] Williams, David B., et al., "The transmission electron microscope," Transmission electron microscopy, Springer, Us, pp.3-17, 1996.
[7] cryo-EM, https://en.wikipedia.org/wiki/Cryo-electron_microscopy, Feb. 2017.

Duseok Jin is a principal researcher at National Institute of Supercomputing & Networking, KISTI. He received his M.S in Computer Science from Univ. of Chonbuk, Korea and his Ph.D. in Computer Science from Univ. of Paichai, Korea in 2001 and 2011 respectively. His Research interests are Information Retrieval system, Cloud Storage and parallel/distributed file system.

Il-Yeon Yeo received the B.S. and M.S. degrees from the School of Electronic Engineering at Kyungpook National University, in 2000 and 2002, respectively. He has been serving as a Senior Researcher of Global Science experimental Data hub Center at Korea Institute of Science and Technology Information (KISTI) since 2012. He served as a Senior Researcher of Knowledge Information Center and National Science and Technology Information Service at KISTI since 2002. His research interests include Parallel Computing, Information Retrieval, Database, Grid Computing, and Security.

Jeong-Heon Kim received his B.Sc., Engineer diploma from Chung Ang University (Seoul, Korea) in 2003, and his Ph.D. in computer science and engineering from Chung Ang University in 2013. Since 2013, he has been worked as a researcher in the Korea Institute of Science and Technology Information. He is currently interested in Image Processing and System management.