Fault Range Analysis of Information and Communications Assets in Power Grid: A Graph Data Perspective

Bo Chai[*], Siyan Liu, Jiangpeng Dai, Ting Zhao, Aihua Zhou and Kunlun Gao

Advanced Computing and Big Data Laboratory, Global Energy Interconnection Research Institute
State Grid Corporation of China, Beijing, 102209, China
[*]Corresponding author

Abstract—**In this paper, information and communication assets in the power grid are modeled as graph data. A graph database, Neo4j, is chosen to capturing the connections that could help to analyze fault ranges. Based on breadth-first search, recursion terminations are designed to obtain the fault range from the failed nodes. Two cases are given in the paper, and more cases will be further considered in the future work. In final, we provide the analysis results and shows the advantage of graph databases**.

Keywords-graph data; fault range analysis; Neo4j; breadth-first search

I. INTRODUCTION

With the penetration of smart grid infrastructure (e.g., distributed energy resources, smart transportation station, ElectricalVehicle charging stations), an increasing number of information and communication assets are deployed to realize high performance control and automated management in power system [1], [2]. Information and communication asset management has been determined as a main challenge brought by the aforementioned trend of power system development. The topological relations among the communication and information devices become dramatically complex, when a huge number of information and communication devices are taken into account [3]. However, traditional relational databases are closely coupled with information and communication asset management system. Thus, application functions related to topological queries (e.g., shortest path query, reachability query, breadth-first search) have been a bottleneck of the asset management system [4].

Compared with traditional relational databases, graph databases are more suitable to storage, manage, visualize and analyze highly connected data [3], [5]. Graph databases are suitable for a variety of data applications with graph structure, e.g., social networks, biological networks, transportation networks [6], [7]. Inspired by the graph theory, a graph model in the database consists of three types of basic elements: node (or vertex), edge, and property (or attribute). A node in a graph database is utilized to describe an entity, e.g., person, web page, IT device. The node is equivalent to one column in a sheet in the dual relationship database of the graph database. An edge is utilized to describe the relationship between two nodes. For instance, a "friend" edge connects two nodes

"James" and "Tom". That is to say, James and Tom are good friends. It is noticed that an edge can be either directed or undirected, and different kinds of edges can represent different meanings. Edge is the key concept in the graph model. The relationships make graph databases outperform other databases in some aspects, since it is not straightforward to define connections in relational databases. A property is an attribute associated with a node or an edge. For instance, if customers are defined as nodes, properties would be defined as age, gender, interests, or residence place.

Fault range analysis is mainly important in the information and communication system monitoring implement [8], [9]. During failure diagnosis, administrators concern about how the other devices are affected by the failed one. With the support of the chosen graph database Neo4j, a breadth-first search based fault range analysis can be realized to tackle the aforementioned challenge. In this paper, we firstly introduce the graph database Neo4j into the related application. The Neo4j graph database has great potentials in the application of cyber system monitoring and managing. Secondly, topological data is modeled for information and communication assets. All of nodes, edges, and properties are explicitly defined. Then, we obtain a property graph for the information and communication assets. Thirdly, we design a breadth-first search based fault range analysis algorithm and illustrate the analysis framework in detail. The algorithm can be further extended with more fault manners.

The remainder of this paper is organized as follows. Section II presents related work in the area. Section III illustrates the graph model for the information and communication devices. In Section IV, we introduce the breadth-first-search based fault range analysis algorithm. Finally, the conclusion is drawn in Section V.

II. RELATED WORK

Based on graph processing platforms, many large online system with deep links between records are developed, e.g., PageRank algorithm for Google, social graph analysis tool for Facebook. To meet the market requirements of graph data management and analysis, a lot of graph database management systems emerge, e.g., Neo4j, GraphX, GraphSQL. Among all these graph processing systems, Neo4j is one of the most

popular graph databases [10]. Neo4j supports property graph modeling [11]. The community develops Cypher language as graph query language for Neo4j. It provides powerful APIs developed by various programming languages, e.g., Java, Python, Ruby, Javascript. As a flagship graph database company, Neo4j has provided sophisticated supports for customers successfully, e.g., Walmart, eBay, and LinkedIn. Those applications for the enterprises include master data management, network and IT operations and fraud detection.

Neo4j has been successfully used to solve problems about network and data center management. System designers manage critical system infrastructures for high availability by capturing how devices are connected. It is noticed that graph database is an excellent format for capturing and modeling the connections, which can help to diagnose failures. The functions supported by Neo4j consist of impact analysis, network planning, root-cause analysis, routing and quality-of-service (QoS) mapping, as well as IT infrastructure management.

III. SYSTEM MODELING

A three-layer graph model of these information and communication devices is proposed. Firstly, the physical model of a simple system is introduced in this chapter. Then, virtual layers about business logics are described in the section. On the basis of the graph model, fault range analysis strategy is derived in Section IV.

The three-layer graph model is composed of a physical layer, which represents the facilities and their physical connections, a cluster layer, which represents clusters, and a system layer, which represents application systems and their mediate connections.

A. Physical Layer

The physical layer of the graph model shows the topology of the information and communication network. The definitions of vertex and edge are discussed below:

The layer is a hierarchical directed graph. In the physical layer, vertices stand for the information and communication facilities in the application systems. In order to make the example simple, only three types of devices are modeled, servers, storage devices, and network switches.

There are two types of properties for the vertices: universal property and particular property. On the one hand, no matter type of a vertex, it has three universal properties: Device ID, Model No., and Location. On the other hand, the particular properties vary according to the types of facility vertices. Servers have Subcategory, IP Address, Operation system, Configuration, and Type of Cluster properties. Storage devices have IP Address property. Network switches have Ports property. Besides, properties of vertices are scalable in the graph model.

All server vertices and storage devices in the physical layer are terminals, which means they are in the bottom level of the hierarchical graph. Generally, there are several levels of network switches above terminals.

The connections among devices are characterized by undirected edges. An edge represents a certain network cable.

For instance, Figure 1. displays a simple physical-layer graph model of an electric power geographical information system(GIS). As shown in the figure, vertices with ID S_i are terminal server vertices. Besides, NS_i stands for network switch, and SD_i stands for storage device.
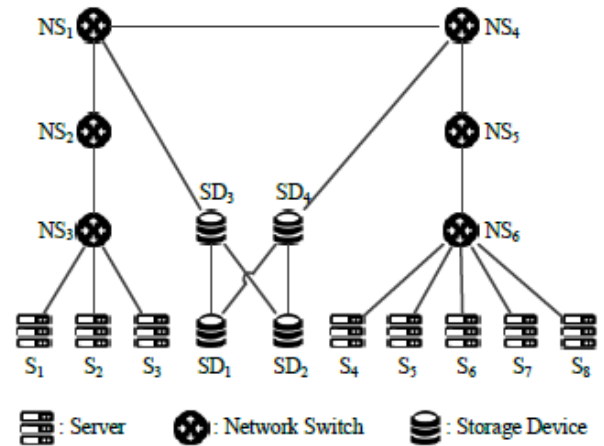

FIGURE I. THE PHYSICAL LAYER GRAPH MODEL OF AN ELECTRIC POWER GIS

Though some parts of the cyber system are simplified, the graph model shows the framework of electric power GIS. According tasks undertaken, servers can be classified into various types. In the example, S_1 and S_2 are data servers. S_3 and S_4 are load-balancing equipments. S_5 is a map server and S_6 is an agent server. In addition, S_7 and S_8 are application servers. It is significant to point out the classification, because it is related to the definition of virtual vertices.

B. Virtual Layers

In order to express business logic among information and communication devices, two virtual layers are adopted in the whole graph model. One is cluster layer, the other is system layer.
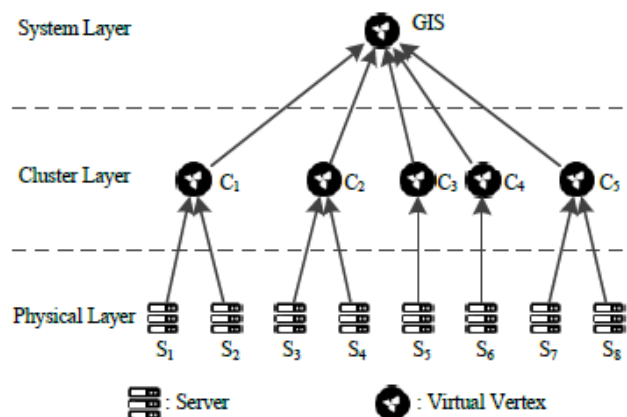

FIGURE II. A CROSS-SECTION OF THE THREE-LAYER GRAPH MODEL

The modeling method introduces the concept of supernodes, which is the aggregation of a group of vertex. A group of devices carrying out a certain function is considered as a vertex in the virtual layers. In other words, a virtual vertex is a functional abstract of a bunch of servers. In the cluster layer, a vertex is on behalf of a certain cluster of terminal servers in the physical layer. For system layer, a vertex is on behalf of an application system which consists of one or more clusters. In fact, virtual vertices are highly related to the classification of servers. Besides, network switches have no correspondence in the virtual layers, because there is no top-level business function running on network switch.

Between cluster layer and physical layer, as well as between system layer and cluster layer, there are edges indicating belongingness. If a terminal server is a part of a certain cluster, there will be a cross-layer edge between the terminal server vertex and the cluster vertex.

Figure. 2 shows a cross-section of the three-layer graph model. Virtual vertices in the cluster layer represent the application server cluster, the map server cluster and etc, while the whole GIS system is described as a vertex in the system layer.

Among virtual vertices in the same layer, working flows are shown by heteroid edges. An edge shows a precedure in a certain working flow. It is possible for the virtual layer to contain more than one working flows.
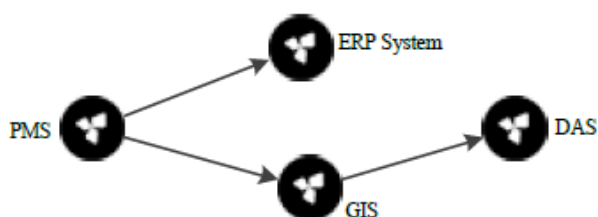

FIGURE III. DEPENDENCE AMONG SYSTEM VERTICES

Figure 3. shows system dependence. The edges indicate business logics, which help the system make an inference about fault range. For simplicity, only power production management system (PMS), electric power geographical information system, enterprise resource planning system (ERP), and distribution automation system (DAS) are taken into account.

IV.   FAULT RANGE ANALYSIS BASED ON GRAPH DATA

The administrators need to know if one device fails to work effectively, how the other devices and systems are affected. In addition, they care about the redundancy of the cyber system. In advance of any system service interruption, the administrators are required to predict and warn system failures. Fault range analysis is implemented to achieve the aforementioned goal. Searching algorithms can be further utilized to analyze fault range caused by any device failure.

Especially, breadth-first search is the algorithm for traversal or searching on the graph data structure. It starts from one node in a graph and explores the neighbor nodes first, and then move to neighbors [12]. Recursion termination design is extremely important for BFS algorithm, since the specific requirement needs to be satisfied. The impacts bought by different devices should be modeled in different fault manners. Fault manners should be carefully categorized and recursion terminations need to be designed corresponding to the fault manners. In current version, we only consider two kinds of fault manner: 1) the servers or the storage device fails to work; 2) the network switch fails to work. More fault manners may be included in the future work.

As is mentioned above, we obtain a graph-based fault range analysis framework. After graph data modeling, the information and communication assets are stored in the graph databases Neo4j. Neo4j provides powerful graph searching techniques, e.g., BFS, Depth-first Search (DFS). With the help of BFS, we can design recursion terminations to decide which nodes will be traversed from the failed device, that is to say, a large variety of fault manners can be considered to analyze the fault range. After introducing the analysis framework, we present the two cases in detail.

Case 1: If a server or a storage device fails to work, the network switch connected directly will be affected firstly. Due to the failure of the terminal device, the associated data cannot be correctly collected by the network switch. In addition, the cluster vertex, as well as the system vertex, connected will be affected. Because of the failure of the server or the storage device, the system may fail to operate regularly. It is noticed that the spare server or the storage device will be added to ensure system redundancy. For instance, two data servers are supporting the operation of the power gird GIS system. If only one data server fails to work regularly, the power grid GIS system maybe operate irregularly. However, the system node will be definitely affected in the current analysis framework.
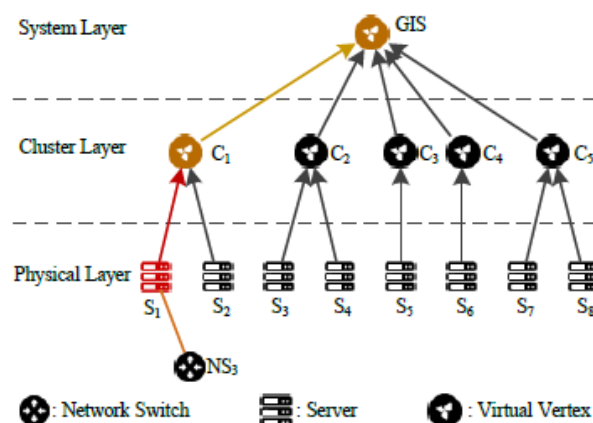

FIGURE IV. SERVER FAILURE IN ELECTRIC POWER GIS

Figure. 4 displays server failure case in an electric power GIS. The red server stands for the server failed, and the brown virtual vertices indicates the affected parts. When S_1 crushes, all edges of S_1 fails. On the one hand, data cannot transmit to NS_3, but NS_3 does not lose its function. On the other hand, the functions of cluster C_1 and system vertex *GIS* are affected. However, they can work in an ill condition.

Case 2: If the network switch fails to work, various parts in the graph model might be affected. Firstly, all the servers

and the storage devices connected directly or mediately will be affected. Due to the failure of the network switch, all the associated servers and the storage devices switch cannot exchange and upload data regularly. Secondly, the system supported by all those devices will operate irregularly. It is worth noticing that the network switch will affect the spare devices. Thirdly, the network switch connected by the failed network switch will also be affected. Lastly, since the system node cannot operate regularly and systems can effect with each other, the connected system node should also be considered in the fault range.
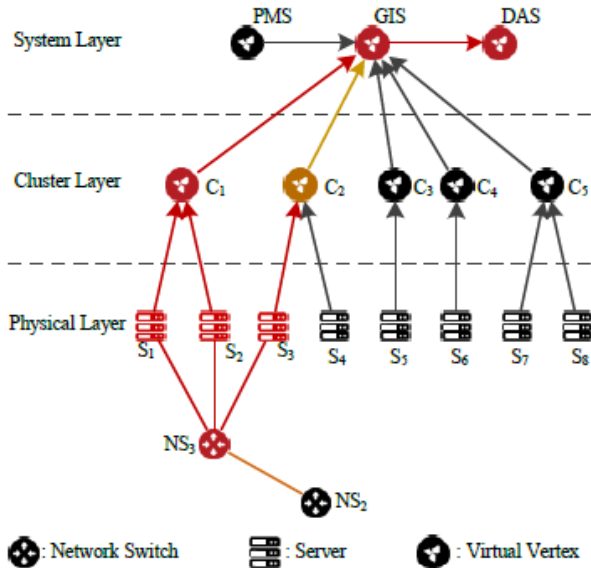

FIGURE V. NETWORK SWITCH FAILURE IN THE THREE-LAYER GRAPH MODEL

According to Figure 5. a failure on network switch NS_3 causes S_1, S_2, and S_3 disconnected. Cluster C_1, as well as GIS, loses its function. Finally, the failure of GIS spreads to DAS. Thus, the failure on NS_3 will result in unreliable operation condition of DAS system. It is noticed that the priority of network switches is superior to that of common servers and storage devices.

V. CONCLUSION

In this paper, we choose graph databases Neo4j to model information and communication assets in the power grid. Especially, we introduce the graph database Neo4j and show the potential of its application in IT device management. Then, we present the graph model in detail. In addition, two cases of fault manners are considered, and a breadth-first search based fault analysis algorithm corresponding to the two cases is presented. Finally, two examples are dissected to show the results.

REFERENCES

[1] J. Momoh. Smart grid: Fundamentals of design and analysis. Wiley, 2012.

[2] Bo Chai, Alberto Costa, Selin Damla Ahipasaoglu, Chau Yuen, and Zaiyue Yang. Optimal meeting scheduling in smart commercial building for energy cost reduction. IEEE Transactions on Smart Grid, 2016.

[3] A ke J Holmgren. Using graph models to analyze the vulnerability of electric power networks. Risk analysis, 26(4):955–969, 2006.

[4] Ian Robinson, Jim Webber, and Emil Eifrem. Graph databases: new opportunities for connected data. " O'Reilly Media, Inc.", 2015.

[5] Yunkai Liu and Theresa M Vitolo. Graph data warehouse: Steps to integrating graph databases into the traditional conceptual structure of a data warehouse. In Big Data (BigData Congress), 2013 IEEE International Congress on, pages 433–434. IEEE, 2013.

[6] Sotirios Beis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Benchmarking graph databases on the problem of community detection. In New Trends in Database and Information Systems II, pages 3–14. Springer, 2015.

[7] Mark Blair, Yunkai Liu, and Theresa M Vitolo. Extending automated intelligence systems via graph database: A case study of the" meth hunter". In Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on, pages 261–268. IEEE, 2014.

[8] Jun Yan, Haibo He, and Yan Sun. Integrated security analysis on cascading failure in complex networks. IEEE Transactions on Information Forensics and Security, 9(3):451–463, 2014.

[9] Youwei Jia, Zhao Xu, Siu-Lau Ho, Zheng Xiong Feng, and Loi Lei Lai. Security analysis of smart grids-a complex network perspective. In 9th IET International Conference on Advances in Power System Control, Operation and Management (APSCOM 2012). IET, 2012.

[10] Justin J Miller. Graph database applications and concepts with neo4j. In Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA, volume 2324, page 36, 2013.

[11] Hongcheng Huang and Ziyu Dong. Research on architecture and query performance based on distributed graph database neo4j. In Consumer Electronics, Communications and Networks (CECNet), 2013 3rd International Conference on, pages 533–536. IEEE, 2013.

[12] Haichuan Shang and Masaru Kitsuregawa. Efficient breadth-first search on large graphs with skewed degree distributions. In Proceedings of the 16th International Conference on Extending Database Technology, pages 311–322. ACM, 2013.