Envision of Wireless Big Data Storage for Low-Earth-Orbit Satellite based Cloud

Article *in* IEEE Wireless Communications · February 2018

Huawei Huang
The University of Aizu

Song Guo
The Hong Kong Polytechnic University

Kun Wang
Nanjing University of Posts and Telecommun…

This is the pre-publication version of our paper that has been accepted by the Special Issue-Wireless Big Data: Technologies and Applications-February 2018, IEEE Wireless Communications Magazine. Content may be changed in the final version.

Envision of Wireless Big Data Storage for Low-Earth-Orbit Satellite based Cloud

Huawei Huang, Song Guo and Kun Wang

***Abstract*—Satellite-based communication technology regains much attention in the past few years, where the satellites mainly play the role as relay devices to the communication networks. In this article, we focus on the data storage leveraging the low-earth-orbit (LEO) satellite based datacenter (also called space based cloud), which has been considered as a very promising secure paradigm to data storage, especially in the big data era. We first propose a layered architecture for the space based cloud infrastructure. Then, a novel management framework is devised to enable the operability of the proposed architecture. Finally, some challenging open issues are revealed from both academic and industrial perspectives.**

I. Introduction

Big Data storage [1] relates to the storage and management of large-scale datasets, while the availability and reliability should be guaranteed at same time. Basically, a data storage system consists of infrastructure and management mechanism. The hardware based infrastructure should support elastic reconfiguration, and be adaptive to various application scenarios. To maintain the datasets, the management mechanism is equipped on the top of hardware infrastructure to provide interfaces for query, further programming intensions (such as analysis) and other interaction with the stored data.

A. Background

Big data paradigm incurs the explosive growth of data and requires efficient storage and management. Referring to [1], the existing state-of-the-art big data storage system types mainly include Distributed File Systems, NoSQL Databases, NewSQL Databases and Big Data Querying Platforms. Although these storage systems have been proposed to meet the stringent requirements on storage and management, it is widely admitted that those storage systems that handle cloud operations and cloud storage, are very leaky and prone to cyber attacks. Therefore, some pioneering companies are seeking new paradigms to provide real secure data storage and management. As a typical example, Cloud Constellation [2] intends to establish a LEO satellite based cloud infrastructure named SpaceBelt, which plans to offer secure data storage for internet service providers, large enterprises such as telecommunication businesses, and government organizations. To solve the widespread global crisis of data insecurity, this system will utilize a combination of LEO orbiting satellites and secure ground networks, as shown in Fig. 1, which allows customers to store vast amount of sensitive and mission-critical data securely in space. The advantages of such system are easily recognized [2], because it can secure the high-value data by: (a) insulating it completely from the Internet and terrestrial leased lines, (b) liberating it from cyber attacks and surreptitious activities, (c) protecting it from natural disasters and majeure events on the earth, (d) addressing all jurisdictional issues, and (e) avoiding risks of violating privacy regulations.

B. Our Motivation and Contribution

Recently, a number of studies and technologies leveraging the LEO satellite based communication networks have emerged in both academia and industry. For example, Jia et al. [3] studied the data transmission and downloading primarily using the inter-satellite links in the LEO satellite based networks. In industry, several representative LEO satellite based projects have been launched, such as the *OneWeb satellite constellation* has been proposed to provide global Internet broadband service to individual consumers as early as 2019. *SpaceX* has detailed ambitious plans [4] to bring fast Internet access to the entire world with a new satellite system that offers greater speeds and lower latency than existing satellite networks. *Boeing* will build a project named GiSAT [5] with 702 satellites. It will offer twice the capacity of previous digital payload designs for Cayman-Islands based global IP with a new digital payload.

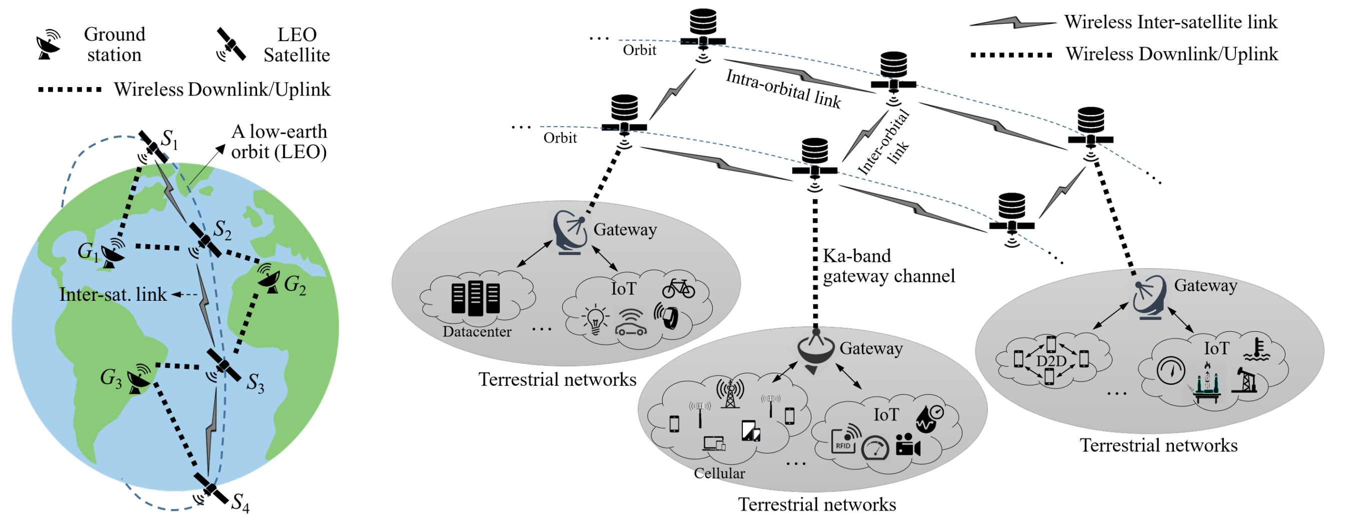
We notice that all the satellites mentioned in the existing studies and projects mentioned above essentially play the role as "relay devices" for the satellite based communication networks as auxiliary extensions of terrestrial core networks. However, inspired by the SpaceBelt project, we believe that in the upcoming *space based cloud* era, LEO satellites could undertake more missions than only performing as relay devices for communication networks. Therefore, we first propose a novel data storage architecture built beyond the LEO constellation. Then, aiming to stimulate the academic research, industrial development and the formulation of standards, we envision some crucial technique challenges and open issues on the big data storage under such LEO based architecture. To the best of our knowledge, this article is the first to study the space based cloud.

II. Architecture of LEO based Data Storage

In this section, we present a layered architecture of data storage infrastructure built on the space based cloud.

A. Explanation of Each Layer

We introduce each layer in terms of constitution and their functionalities from bottom to top as shown in Fig. 2.



(a) LEO illustration with only one orbit.

(b) A use case under the space based cloud.

Fig. 1. Illustration of LEO-satellite based system and a use case under space based cloud. The data originating from terrestrial networks such as datacenter clouds, cellular networks, Internet of Things (IoT) networks, and Device-to-Device (D2D) networks, can be stored in the storages located at LEO satellites.

1) Customer Layer: The Customer Layer contains various users including the Internet service providers, government entities and other multifaceted enterprises. Large volume of data is generated from the terrestrial networks such as datacenter clouds, cellular networks, Internet of Things (IoT) networks, and Device-to-Device (D2D) networks. Users can upload data or submit data-downloading requests to the upper layer.

2) Gateway Layer: All the requests of uploading or downloading received by Gateway Layer will be aggregated to an internal central Workload Scheduler, which then makes decision on workload allocation according to some specific policies such as load-balancing, first-in-first-out, best-fit and other priority based algorithms. Note that, a job request can be delivered to a designated ground station or even a satellite, which transmits data from the Data-Storage Layer to the ground stations.

3) Data-Storage Layer: In this layer, the datasets are stored in distributed server farms, which are deployed over LEO satellites. In each LEO satellite, at least one server should be embedded in the on-board unit. Then, each server hosts one or multiple Virtual Machines (VMs), in which some popular data storage technologies such as Hadoop DataBase (HBASE), Highly Immersive Visualization Environment (HIVE) and Hadoop Distributed File System (HDFS), can execute to provide data management functionalities. We call such servers the LEO servers for short hereafter.

B. Essential Basics

Channel State: In particular, each channel between a LEO satellite and a GS is periodically available because the satellite is periodical circling around the earth in a designated orbital plane. In the system where multiple LEO satellites exist in each orbital plane, as shown in Fig. 1(a), a GS can connect with multiple consecutive satellites simultaneously, and vice versa. As a matter of fact, the transmission rate of a satellite-GS channel is affected by two critical factors [6]: the power-supply from the solar-board of satellite and the time-varying weather conditions. Although the latter one is uncontrollable, the bandwidth (i.e., channel transmission rate) of a uplink or downlink could be still determined by manipulating the power-supply in the satellite based on the observed channel state.

Replication or Network-Coding: Giant volume of heterogenous datasets originated from users need to be stored in the LEO servers. Each original user dataset is with a unique size and should be stored in a manner of replication or coding. If the resilience and robustness of data recovery are considered, each dataset needs to be divided into multiple partitions, each of which should be with an appropriate size and stored in multiple replicas or in encoded chunks over the distributed LEO servers. When the network-coding mechanism is adopted in the data storage, a dataset can be partitioned and encoded using various coding schemes. As a typical one under the erasure coding category, the Maximum-Distance Separable (MDS) code has been widely adopted in [7], [8] and is described as follows. In an (N, K) MDS code, a data object with size M is divided into K pieces, and each has a size M/K. Then, encode them into $N(\geq K)$ coded chunks with same size, and store them over N (distributed) nodes. In such manner, the original file can be recovered from any set of K encoded chunks.

III. Characteristics of the Proposed Architecture

In this section, we envision the typical advantages and challenges of the proposed space based data storage cloud.

A. Advantages and Benefits

On one hand, the biggest advantage of space based data storage architecture is that it provides the definite security for the customer's data. We can see from both Fig. 1(b) and Fig.

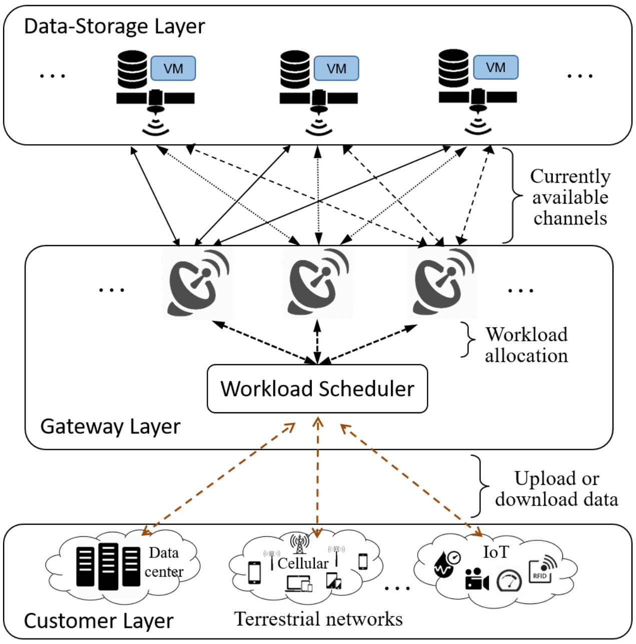


Fig. 2. The three-layer architecture for space based storage infrastructure.

2 that the data-storage layer is completely insulated from the terrestrial networks, cutting off all the threats (such as cyber attack, espionage and data theft) from terrestrial Internet and leased-line infrastructure. Besides, the proposed architecture would obviously benefit the big data storage with the complete immunity from natural disasters occurred on the earth.

On the other hand, as the jurisdictional controls are imposing ever growing restrictions on where data can go around the world today, our proposed space based cloud storage architecture will provide a safe haven for data storage without interruption to any unintended network jurisdictions.

Furthermore, this architecture can offer opportunities to some cloud networks designed to move data around the globe without considering geographic boundaries as well.

B. Application Scenarios and Data Flow Modes

Under the proposed architecture, several representative application scenarios can be summarized as follows.

In the viewpoint of customers:

- *Data upload/download service*: The most common application scenarios are data storage service and delay-sensitive access service.
- *Remote sharing*: Different sectors of a large organization/company located at different global sites can share big data via the global circulating space based cloud with a faster pace than that of the traditional terrestrial cloud networks.

In the infrastructure administrator's perspective: As an administrator of the proposed infrastructure, there are many data processing tasks should be conducted. For example,

- *Data migration*: To realize the distributed network coding for big-sized data, data migration is necessary to reducing the decoding latency when customers requiring their original data. But this must be based on the situation that the relay LEO satellites are being deployed.
- *Data evacuation*: In the space disaster context such as a meteor storm or comet strike, the data residing in some LEO servers is required to evacuate to safe sites in case of unrecoverable damages to data.

Via observation from different application scenarios, we can classify the data flow modes into the following three categories with respect to a pair of source-destination session: (a) GS-LEO-GS mode, where multi-hop inter-satellite links (ISLs) are potentially utilized and the LEO satellites play the role as relay devices for remote data delivery; (b) GS-LEO mode, in which users can upload data or deliver requests to LEO servers; (c) LEO-GS mode, where users access data from LEO servers. Any application scenarios are composed of such three data flow modes. Based on them, we then discuss some critical technique challenges under the space based data storage infrastructure in the next subsection.

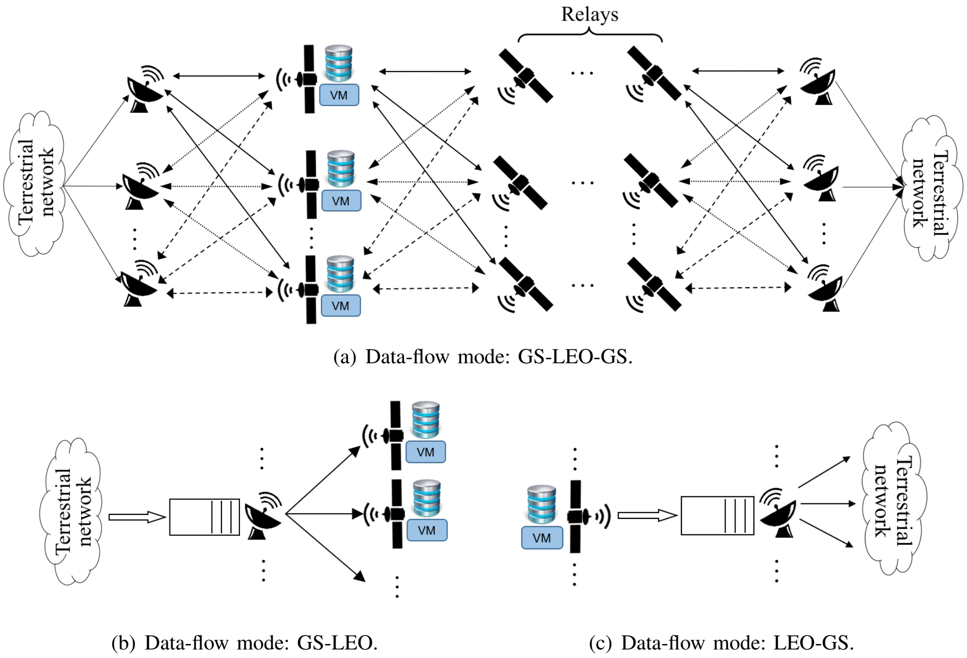
C. Technique Challenges

Since multiple satellites are assumed to exist in each LEO, and the cycle period of each LEO satellite varies between around 90 minutes to around 120 minutes, the connection between a GS and a LEO satellite periodically appears and lasts only 20 minutes around. Thus, under such periodically available connections, there are a lot of technique challenges summarized below to realize the proposed space based data storage infrastructure.

1) Data Storage Mechanism: The first question is how data is going to be stored in the data-storage layer: replication based or network coding based approach? If the former one is adopted, how each dataset is replicated, mirrored and deployed over the distributed LEO servers along an orbit? Otherwise, what network coding scheme will be used (linear random coding, erasure coding, etc.)? How to store all the encoded chunks such that the original dataset can be easily and quickly retrieved thus to be recovered? Therefore, when choosing the storage mechanism, the tradeoff between the access delay and the storage cost, which is measured in terms of the storage space occupied by replicas or encoded chunks, should be carefully taken into account.

2) Effects of Satellite Mobility: LEO satellites have high mobility speeds relative to earth, resulting in that the topology of a satellite network dynamically changes all the time. Even though the time-varying topology changes are periodic and thus can be predicted, the handovers between LEO satellites and GSs are inevitable. This leads to higher operating expenditure (OPEX) and higher latency to the data flow transmissions for sure in the space based cloud.

3) Power Allocation: Since each LEO satellite is working under the power-supply by its solar panel equipped, the total power level is limited in a satellite. Also, the transmission rate of a satellite-GS channel is reported positively related to the power-supply [6] and affected by the current channel conditions when a GS is downloading data from a satellite. Given the current channel states, the energy-efficient power



(a) Data-flow mode: GS-LEO-GS.

(b) Data-flow mode: GS-LEO.

(c) Data-flow mode: LEO-GS.

Fig. 3. Typical flow models under multiple LEO-satellite system.

allocation in each satellite is a great challenge for all concurrently connected satellite-GS sessions for a large number of downloading tasks at hand.

4) Bandwidth Scheduling: Similar to power allocation, the bandwidth scheduling between each GS and each LEO satellite is another challenging task such that all the delay-sensitive uploading/downloading requests can be completed within a tolerable latency.

5) Support the Big Data Stream Processing: The data volume grows exponentially in today's big data era, and the real-time applications urge the stream processing paradigm. This is not a big deal in terrestrial networks. For example, one can implement a stream processing platform in either a datacenter [9] or an edge cloud [10]–[12], which is able to handle the batch processing, parallel computing leveraging. However, achieving real-time data processing will become a huge challenge in the space based storage cloud, since datasets are maintained throughout distributed LEO servers. It is not hard to imagine that a stream flow will suffer larger latency in the space based cloud with GS-satellite channels than that in the conventional terrestrial networks.

6) Fault Tolerance: A data storage server residing in a satellite will be down in case of some disasters in outer space such as radiation, strike by space junk or meteors. Thus a fault-tolerance mechanism should be provided to ensure the recovery of a data once some failures occur. However, fault-tolerance is a hard nut to crack even in the terrestrial networks [13]. There is no 100-percent reliable fault tolerant software, especially for the space based server farm. Thus the main task is to reduce the probability of failures to an tolerant level when manufacturing the data storage hardware embedded in satellites.

7) Consistency Maintenance: Like in a traditional cloud storage, the data version management and consistency maintenance functionalities must be provided by the space based data storage infrastructure. Unfortunately, they are much more difficult than that in a terrestrial cloud, since a synchronisation operation towards data version update and consistency will incur a large scale of data fluctuation over all the involved LEO servers. This requires cost-efficient management strategies in the system controller.

8) Security and Privacy Challenges for Granular Access Control: In reality, big data infrastructure customers need to comply with diverse security regulations, e.g., privacy policies, legal restrictions, and other special corporate requirements. Consequently, the fine-grained access control mechanisms are needed in the Gateway Layer following the specific regulations.

D. A Framework of Management

Corresponding to the technique challenges, and meanwhile to meet the management requirement of a data storage system mentioned in the beginning of this article, we devise a novel management framework for the operability of the proposed space based cloud architecture.

As shown in Fig. 4, the Management Plane consists of seven major modules which are in charge of workload management and making decisions for both Gateway Layer and Data-Storage Layer. In detail, the *Routing Scheduling* delivers the traffic routing related parameter settings such as routing path finding and bandwidth computing, for both satellites and ground stations according to routing optimization results obtained according to a specified traffic engineering policy. The *Workload Scheduling* controls the *Packet switching module* in GSs for the packet forwarding according to a certain workload allocation strategy. The *Fault Tolerance* module is designed to provide failure recovery mechanism that can handle both node

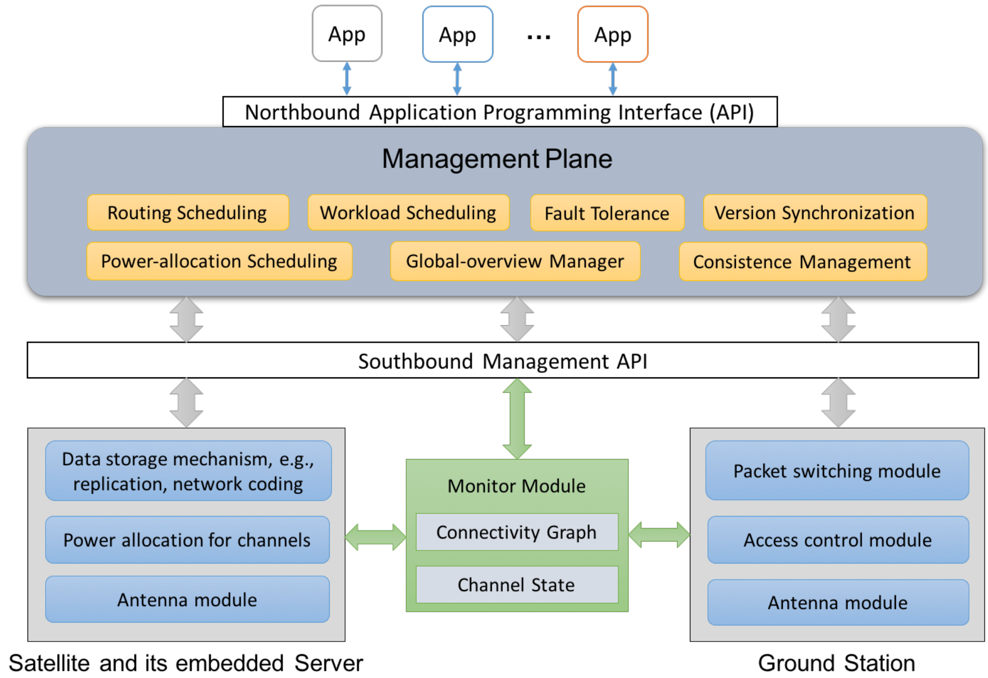


Fig. 4. The management framework for space based storage cloud.

and link failures occurred in the satellites, data storage servers, and the links between any pair of devices. Next, *Version Synchronization* and *Consistence Management* modules are collaboratively controlling the update of data versions when a user makes some changes to the data stored. Then, because the power-supply in satellites are strictly constrained by a energy-budget, which needs to keep satellites alive and ensures the communications. Thus, the *Power-allocation Scheduling* module is used to yield the power-supply parameters for the transmission channels in satellites. Finally, *Global-overview Manager* is designed to aggregate information from the global *Monitor Module* and feed the useful information to all the other centralized controlling modules. Specifically, the *Monitor Module* is an individual auxiliary component, which is used to collect the global data such as the time-varying connectivity graph and downlink/uplink channel conditions. Note that, it can be realized by dedicated hardware embedded in either satellites or terrestrial devices. The role of the monitor module is of great importance, because the collected status data of satellites and transmission channels is very critical for decision making during the mission execution in LEO based clouds.

In addition, the Northbound Application Programming Interfaces (APIs) are developed to interact with the various applications in user side. And the southbound APIs are used to exchange protocol messages between the Management plane and the underly hardware devices such as Monitor Module, ground stations, satellites and the servers embedded.

IV. Open Issues

In this section, some open issues will be addressed for academia, industry and standards organizations, based on the proposed LEO-satellite based cloud storage architecture.

A. Centralized or Decentralized Control Mechanism

The control mechanism is a fundamental issue for the LEO satellite based data storage cloud. In general, the control mechanisms can be classified into the following two categories:

Centralized Manner: As centralized control architectures, several Software-defined Networking based satellite networks [14] are emerging in recent years. Although the centralized control mechanism brings high efficiency to the network management, it also should be noticed that building such centralized control platform will incur a non-negligible cost. Because SDN controllers normally execute in servers located at a terrestrial network, thus the control channels between a controller and each node (satellite or ground station) will consume a considerable bandwidth resource in each link going through. This contributes the bandwidth allocation burden with no doubt.

Decentralized Manner: In this manner, each LEO satellite independently determines its operating parameters such as power allocation for each channel and topology management. Thus, this manner requires the energy-efficient and delay-sensitive distributed algorithms that could run in the on-board unit of satellites. The advantage is that only a limited number of or no messages need to exchange among satellites and their neighbors. Then, the disadvantage is that the global optimal control policy may not be achieved.

B. Consistent, Availability, Partition-tolerance (CAP) Issues

If the decentralized manner is adopted, the space based cloud becomes a distributed data storage system, thus the CAP issues should be taken into account.

- *Consistency:* To improve the availability (to be explained next) and recovery resilience in case of failures, each data piece (replicated or coded) should be stored in multiple LEO servers over the distributed storage system. However, the parallel storage brings the difficulty on maintaining the consistency of dataset when a new version needs to be synchronized, because consistency indicates that all the copies of a dataset must be identical.

- *Availability:* This property refers to the robustness that the entire distributed storage system will not be seriously affected in terms of reading and writing by the inevitable server failures.
- *Partition-tolerance:* When the network is partitioned because of congestions or failures by some unavoidable accidents, the partition-tolerance desires the distributed storage system still works well.

It is worth noting that it is impossible to build an asynchronous network model that all the CAP properties are satisfied simultaneously [15]. Therefore, we can implement the distributed space based data storage system by meeting at most two of the three requirements. That is, according to different design policies, we can establish a CA system by discarding the partition-tolerance, a CP system by neglecting availability, and an AP system that gives up on consistency.

C. Leverage Inter-Satellite Links

As shown in Fig. 1(b), the ISLs include intra-orbital links and inter-orbital links, which connect two consecutive satellites on the same orbit and on different orbits, respectively. From Fig. 3(a), the leveraging of both intra-orbital links and inter-orbital links provides the foundation to applying sophisticated routing algorithms. This is greatly helpful to some purposes on data storage such as: (a) to reduce data uploading/downloading delay; (b) to improve robustness/resilience of data storage system; (c) data migration or evacuation under disaster scenarios; (d) even the VM migration under the data processing scenario.

D. Dedicated Hardware Equipments

Space is an extraordinary harsh environment with strong radiation incurred by the cosmic rays, and extreme temperatures alternated between light and shaded sides around earth. To ensure that LEO servers could work well in orbit, we need to provide substantial shielding equipment preventing from cosmic rays, and new cooling system for the embedded server machines in satellites. These requirements of dedicated hardware equipments will stimulate industries to develop new technologies.

E. Standardized Interfaces

As designed in Fig. 4, the applications in the customer side could use the space based cloud infrastructure through the Northbound APIs, which offer the access to the Data-Storage Layer in terms of query, uploading and downloading operations. To realize these functionalities, the Management Plane needs to interact with the hardware based devices including monitor module, ground stations, satellites and the embedded servers, via the Southbound Management APIs. For example, the standardized query APIs are necessary to accessing the databases because heterogeneous data storage systems are allowed to exist in the proposed space based storage architecture. Obviously, the implementation of such two categories of APIs and their standardization are the indispensable efforts that should be devoted to by industries and standards organizations in the future.

F. Integration with Analytics Frameworks

As data-analytics is recognized as one of the most critical factors that are producing big business value, it has become the dominating development trend that the analytics components are integrated with the pure data storage systems in the terrestrial big data networks. Likewise, it is in the same situation under the space based storage cloud. Although the integration is going to increase both the capital expenditure and operating expenditure, it is believed that this will strongly impact the data storage solutions, thus further bring opportunities to the big data value chain under the paradigm of space based cloud.

V. Conclusion

In this article, we envisioned the challenges and opportunities for space based storage clouds. To meet the requirement of two crucial components of a data storage system, a layered architecture has been firstly proposed to establish the storage infrastructure beyond LEO satellites. Then, to enable the operability of the proposed architecture, a management framework has been devised. Finally, aiming to promote the proposed space based cloud from concept to reality, we summarized several open issues for academia, industry and standards organizations.

References

[1] M. Strohbach, J. Daubert, H. Ravkin, and M. Lischka, *Big Data Storage*. Springer International Publishing, 2016, pp. 119–141.

[2] “The information ultra-highway for enterprises and governments,” 2017. [Online]. Available: http://cloudconstellation.com/

[3] X. Jia, T. Lv, F. He, and H. Huang, “Collaborative data downloading by using inter-satellite links in leo satellite networks,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1523–1532, 2017.

[4] “Spacex satellite constellation.” [Online]. Available: https://en.wikipedia.org/wiki/SpaceX_satellite_constellation

[5] “Gisat-1.” [Online]. Available: https://en.wikipedia.org/wiki/GiSAT-1

[6] M. J. Neely, E. Modiano, and C. E. Rohrs, “Dynamic power allocation and routing for time-varying wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, 2005.

[7] X. Pei, Y. Wang, X. Ma, and F. Xu, “T-update: A tree-structured update scheme with top-down transmission in erasure-coded systems,” in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, 2016, pp. 1–9.

[8] B. Li, A. Ramamoorthy, and R. Srikant, “Mean-field-analysis of coding versus replication in cloud storage systems,” in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, 2016, pp. 1–9.

[9] K. Wang, Y. Wang, X. Hu, Y. Sun, D.-J. Deng, A. Vinel, and Y. Zhang, “Wireless big data computing in smart grid,” *IEEE Wireless Communications*, vol. 24, no. 2, pp. 58–64, 2017.

[10] H. Huang and S. Guo, “Adaptive service provisioning for mobile edge cloud,” *ZTE Communications*, vol. 15, no. 2, pp. 1–9, 2017.

[11] M. Du, K. Wang, X. Liu, S. Guo, and Y. Zhang, “A differential privacy-based query model for sustainable fog data centers,” *IEEE Transactions on Sustainable Computing*, vol. PP, no. 99, pp. 1–11, 2017.

[12] J. He, J. Wei, K. Chen, Z. Tang, Y. Zhou, and Y. Zhang, “Multi-tier fog computing with large-scale iot data analytics for smart cities,” *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–10, 2017.

[13] K. Wang, Y. Shao, L. Shu, C. Zhu, and Y. Zhang, "Mobile big data fault-tolerant processing for ehealth networks," *IEEE Network*, vol. 30, no. 1, pp. 36–42, 2016.
[14] J. Bao, B. Zhao, W. Yu, Z. Feng, C. Wu, and Z. Gong, "Opensan: a software-defined satellite network architecture," in *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, 2014, pp. 347–348.
[15] S. Gilbert and N. Lynch, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services," *Acm Sigact News*, vol. 33, no. 2, pp. 51–59, 2002.