

A Review on Big Data Movement with Various Approaches

¹Nay Myo Sandar, ²Surekha Lanka, and ³Shuvra Tripura

Faculty of Business and Technology, Information Technology Program, Stamford International University

E-mail: naymyo.sandar@stamford.edu, surekha.lanka@stamford.edu, shuvra.tripura2000@gmail.com

Received February 9, 2024; **Revised** March 30, 2024; **Accepted** April 21, 2024

Abstract

With the growth of technologies and applications, a large amount of data has been produced at an increasing rate from various resources such as social media networks, sensor devices, and other information serving devices. This large collection of massive, complex, and exponential growth of dataset is called big data. As we know very well, almost every business relies heavily on big data to drive decision making and improve operations. The traditional database systems cannot store and process such data due to large and complexity. Therefore, many enterprises have already adopted cloud computing for better storage and processing since it can provide a pool of resources for servers and storage. However, it is significant issue for moving large amounts of data to and from the cloud can indeed present several challenges including bandwidth limitation, latency, and network congestion. To optimize the performance of big data transfer to cloud computing, this paper extensively reviews the previous research works, discusses research issues, summarizes key findings and approaches for dealing with big data movement. From the literature, researchers proposed various network techniques which can be applicable for future use that can help improve the transfer of big data to and from the cloud and ensure for efficient storage and processing of large datasets.

Keywords: big data; cloud computing; big data movement; network techniques

Introduction

In recent years, there has been a continually growing a large amount of data being generated from different resources such as social media networks, sensory devices, mobile, tracking devices etc. Such a collection of large, complex and explosive growth of data is called

big data. The research issue of big data encompasses a variety of challenges in storing, processing, and analyzing large and complex data. Several researchers also addressed the problem of big data issues. For example, this article discusses about the big data analytics techniques such as social media analytics, text analytics, video, audio analytics, and the various issues and challenges related to the data, processing, and management (Rawat & Yadav, 2021). The researchers in (Upadhyay et al.) provide a survey of various research works to handle the inefficient storage problem of Big Data. To handle big data, almost every business adopted cloud computing since it can provide access to a variety of computing resources including storage, processing power, and analytical tools without the need for large upfront investments in infrastructure.

However, careful consideration must be given to the significant challenge of network bandwidth for distributing a large amount of data to the cloud since it can incur a long latency because of bandwidth limitation and network congestion. In the context of big data movement problems, previous researchers have proposed various methods and approaches to address the challenges associated with transferring large volumes of data efficiently and reliably.

Since we consider big data movement problem as significant issue, we mainly explore the previous research works which techniques can be applicable and how they can efficiently transfer large amount of data to the cloud. Therefore, this paper presents the extensive review of big data movement by different frameworks and network technologies such as DTN, CDN, SDN, and data grid. These proposed frameworks and network technologies offer various benefits for big data movement including improved network performance, reliability, and scalability.

The structure of paper is organized as follows. Section 2 presents the overview of big data. Section 3 describes the concept of different types of networks. Section 4 discusses research issues and approaches for addressing big data movement problems. Section 5 concludes the paper and provides further directions for future work.

Research Objectives

Investigate techniques to improve the performance and reliability of allocating big data in cloud environment.

Explore interoperability challenges in network latency in cloud environment.

Explore techniques for better performance between two platforms.

Literature Review

This section provides an overview of big data including big data definition, evolution, challenges, and various approaches for big data movement.

What is Big Data?

Firstly, we studied about the concept about big data. The term “big data” is a collection of datasets which are so large, complex, and grows too fast. Specifically, big data is defined by three main parameters such as Volume, Variety, and Velocity as shown in Fig 1.

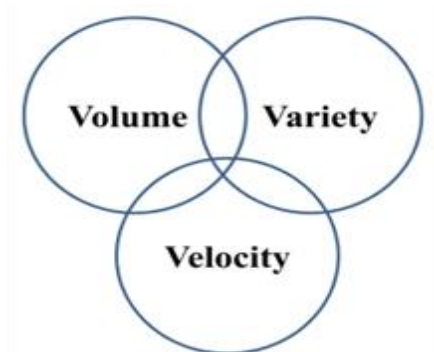


Figure 1 Big Data Parameters

In Fig 1, data volume indicates large dataset which is properly defined in terabytes to even yottabytes. Data variety represents a variety of data from different sources with different types (e.g., text, audio, video, files, sensor data, etc) that make the complexity to data. Data velocity refers to the speed of growth of data. These three parameters are basically measured for big data.

Big Data Evolution

Second, we explored how big data is developed. There are many examples of big data evolution as discussed below.

- Google’s self-driving car generates nearly 1 gigabyte of data per second since it scans and analyzes its environment (Garrett, 2013).
- Well-known social media (e.g., Facebook) produces 500 terabytes of data per day.
- More than 100, 000 videos per day are uploaded to YouTube, equivalent to 360 terabytes every day (Song, 2012).
- Google processes about 1 petabyte of data every hour (Antony & Paul, 2014).
- IBM (i.e., International Business Machines) creates 2.5 exabytes of data every day (Data Analysis Challenge, 2008).

- According to International Data Corporation, the total amount of global data was expected to reach to 2.7 zettabytes during 2012.
- As the amount of data captured by sensors grows, it could potentially increase to the level of yottabytes in 2015 (Soi, 2013).

Big Data Challenges

After that, several challenges addressed to big data are investigated. Due to the size of big data, it exceeds the storage capacity of traditional database architectures. In addition, some big companies (e.g., Facebook, Google, Yahoo, etc) need to process big data analytics to extract useful information. Since the nature of big data is so large and complex, it becomes difficult to process by using traditional data processing applications. As a result, cloud computing has become a possible solution for data storage and processing since it provides on-demand scaling resources of servers and storage for supporting big data.

However, moving large amounts of data to and from the cloud is a challenging task since it could entail a long latency, bandwidth limitation, and network congestion due to data size. Regarding big data movement problems, this paper surveys the earlier research papers in which different frameworks are proposed as available solutions. In addition, some previous papers have applied various network technologies to move big data such as delay tolerant network (DTN), content delivery network (CDN), software defined network (SDN), and data grid.

Issues and Approaches for Big Data Movement

This section reviews the previous papers in which several research issues are analyzed dealing with big data movement problem and proposed by different approaches including frameworks and several network techniques such as delay tolerant network (DTN), content delivery network (CDN), software defined network (SDN), and data grid. A summary of each paper is also presented.

1. Bulk Data Movement to Climate Dataset: Efficient Data Transfer Management with Dynamic Transfer Adjustment (Sim, 2010): The Earth System Grid (ESG) is one of the communities which have the challenge of managing for the distribution of massive datasets to thousands of scientists around the world. As a result, the ESG has implemented a data transfer management tool called the Bulk Data Mover (BDM) to provide efficient data delivery. The BDM achieves high performance with a variety of techniques including multi-threaded concurrent transfer connections, data channel caching, load balancing across multiple servers, and storage I/O pre-fetching. The BDM can transfer multiple files concurrently by using parallel TCP connections.

However, defining the level of concurrency is a non-trivial task. The high level of concurrency becomes increased congestion whereas the low level of concurrency will not utilize the available bandwidth. Hence, dynamic transfer adjustment is further complemented to enhance BDM data transfer. Fig 2 shows the transfer and concurrency management in BDM with dynamic transfer adjustment.

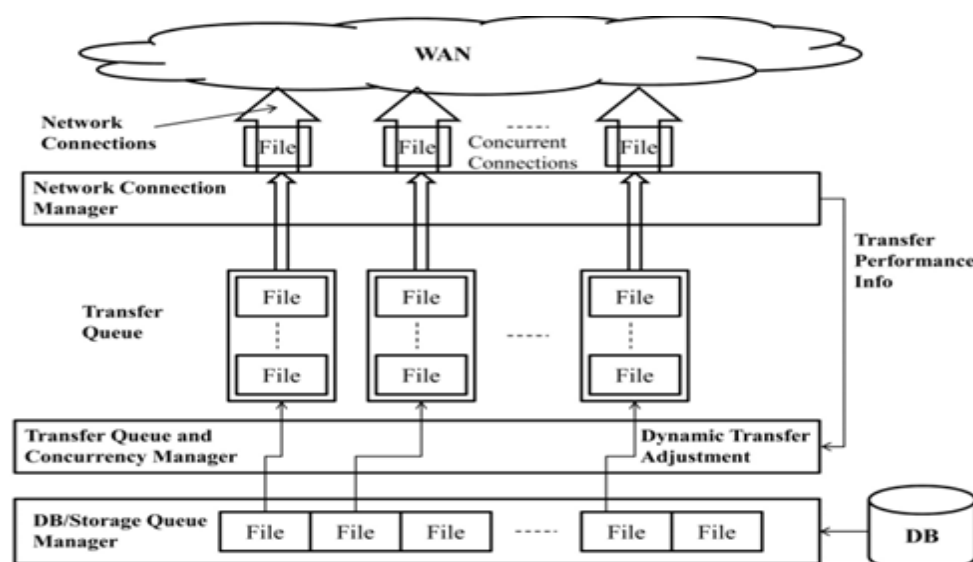


Figure 2 Transfer and concurrency management in BDM, showing dynamic transfer adjustment

As depicted in Fig. 2, BDM manages a DB queue and transfer queues for concurrent file transfers. Each transfer queue checks the threshold for the queued total file sizes and accepts more files from the DB queue if the queued total file size is below the threshold. In this case, the default threshold is set to 200MB according to file size. The performance of dynamic transfer adjustment can help in minimizing congestion of data transfers so that the BDM can reduce the number of concurrent transfers to optimize the connection. In addition, storage I/O prefetching in BDM performs inode creation to write files at the destination. By creating inodes at the destination paths while files are moving on the transfer queue, BDM achieves faster storage I/O during the transfers.

Then, the authors in this paper conducted experiments based on climate datasets which are characterized by large volume of files with different file sizes. The results show that the contribution of dynamic transfer adjustment in the management of BDM can obtain the optimized or controlled transfer performance. However, the proposed BDM tool is lack of fault tolerance technique since it plays a critical role in continuing the operation when the system is failure due to any reasons.

2. Exploring the Optimal Strategy for Large-scale Data Movement in High-Performance Networks (Brown et al., 2012): Next-generation collaborative eScience applications often generate large amounts of data which must be transferred to geographic locations with high-bandwidth network connection. To deal with this issue, the authors proposed a new Workflow-based Intelligent Network Data Movement Advisor (WINDMA) framework to improve performance of bulk data transfer.

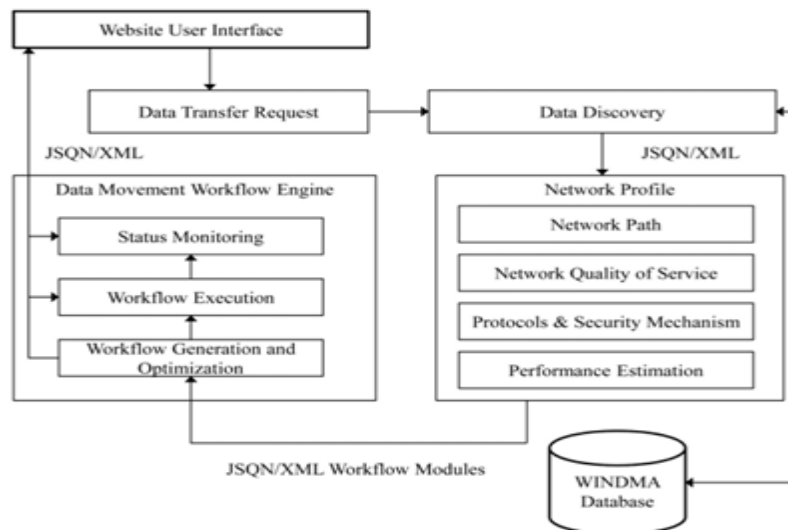


Figure 3 WINDMA functional components and control flow

Fig. 3 depicts the WINDMA framework with several components such as data discovery, network profiling, workflow generation, and optimization. First, the user needs to submit a data transfer request including service start/end time, data source/destination nodes, bandwidth, and financial cost limit on the deployment and utility expenses of WINDMA framework. Then, the component of data discovery performs to retrieve data from a database. The function of network profile searches the network path, quality of service, protocols, and security mechanisms for the request. Finally, WINDMA forwards the data request to a Data Movement Workflow Engine which constructs, optimizes, executes, and monitors a data movement workflow to satisfy the request.

The authors conducted experiments to evaluate the WINDMA framework, and the results demonstrate that their proposed framework improves the performance of bulk data transfer in wide-area networks. However, the authors in this paper ignored the implementation cost and maintenance cost of WINDMA framework. Despite the benefits of the framework, its associated costs of implementation and maintenance are major factors to be considered.

3. Delay Tolerant Bulk Data Transfers on the Internet (Laoutaris, 2009): In this paper, the authors proposed the problem of transmitting Tbyte-sized Delay Tolerant Bulk (DTB) data through commercial ISPs by taking advantage of already-paid-for bandwidth to avoid increasing transit costs. They first considered the performance comparison of End-to-End scheduling policy (E2E-Sched) and Store-and-Forward policy (SnF) utilizing storage inside transit ISPs.

End-to-End (E2E) transfer means DTB data flowing from a sender directly to a receiver. A more prudent E2E approach is to perform scheduling at the sender to avoid or slow transmissions during peak hours. An E2E-Sched policy can take advantage during the off-peak hours of the sender and transmit DTB traffic without impacting the sending access ISP. However, the problem with this policy is that the off-peak hours of the sending ISP do not coincide in time with the off-peak hours of the receiving ISP due to time-zone or traffic profile differences. When such non-coinciding valleys occur, end-to-end transfers are unable to fully utilize the free capacity of both ends.

An alternative approach to solve this problem is to perform Store-and-Forward (SnF) applying an assisting storage node inside the transit ISP. Having transit storage allows a SnF transfer policy to buffer DTB traffic inside the network and allow it to ride on top of multiple load valleys one by one, even if they do not coincide in time.

In the network model, they considered a sender of DTB traffic v connected to an access ISP (i.e., $ISP(v)$), and a receiver u connected to an access ISP (i.e., $ISP(u)$). The two access ISPs are communicated through a transit provider TR as shown in Fig 4.

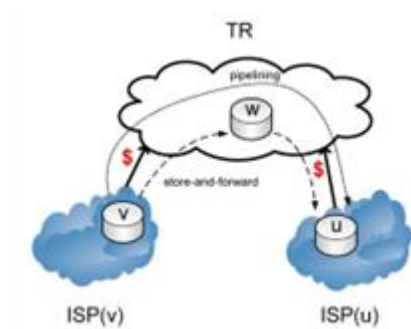


Figure 4 Sender v at $ISP(v)$ and receiver u at $ISP(u)$. 95-percentile pricing on links ISP

(v) $\longleftrightarrow$ TR and $ISP(v)$ $\longleftrightarrow$ TR. A transit storage node w inside the bottleneck-free transit provider TR to be used for store-and-forward of DTB flows from v to u .

In this figure, they assumed that there aren't any bottlenecks inside TR. Hence, they claimed that they don't need to consider placement issues of w , or multiple transit storage nodes since TR is bottleneck free. Next, they presented their main contribution towards the improvement of the understanding performance comparison of between E2E-Sched and SnF. Let $F(P)$ denote the maximum volume of DTB data which can be delivered for free by policy P between nodes v and u within a time allowance of T . Then if an application has to send a volume of B , their strategy is described as follows.

- If $B < F(\text{E2E-Sched})$, E2E-Sched can send them for free. In this case, there is no need to apply storage inside the network.
- If $F(\text{E2E-Sched}) < B < F(\text{SnF})$, SnF can utilize relatively cheap network storage to send the data at zero transit cost.
- If $B > F(\text{SnF})$, SnF can utilize network storage to send the data at the smallest possible transit cost.

Based on the extensive performance evaluation, there exists significant potential for using commercial ISPs to perform low cost DTB transfers.

4. Vehicular Carriers for Big Data Transfers (Gorcitz et al., 2012): In this paper, the authors proposed to exploit the delay-tolerant nature of bulk transfers to deliver data over the existing road infrastructures. Their work is motivated by using conventional vehicles as the communication medium for big data migration in an opportunistic manner and miles traveled in the world. The proposed idea is illustrated in Fig. 5.

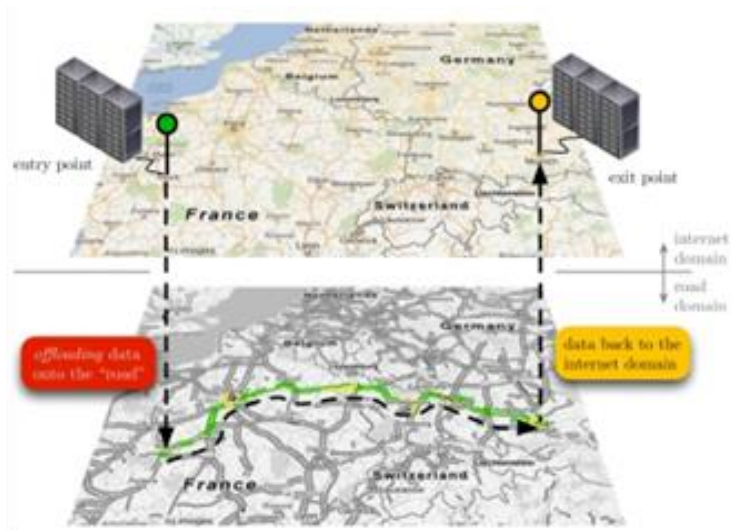


Figure 5 At some point in space, data is offloaded from the Internet domain onto the road domain and is carried by vehicles up to some destination address, saving significant Internet resources.

As shown in Fig 5, a novel technique for future smart vehicles is proposed for big data transfers by offloading data from the Internet onto a network composed of conventional vehicles. Vehicles are equipped with one or more removable memory storage devices such as magnetic disks or non-volatile solid-state storage devices to overcome the limitations of capacity and design of the Internet. The term “vehicle” refers to both passenger and commercial vehicles.

Then, they described briefly their system operation including vehicles and their users namely a service provider and a content provider. A content provider distributes the data to be piggybacked onto the vehicles through a wide area network such as the Internet. The service provider charges the content provider based on the amount of data transferred along the road infrastructure. A network of “offloading spots” provides the data to be piggybacked on the memory of vehicles. The term “offloading spots” refers to locations which provide the data to be transferred to the memory of devices of the vehicles or where on-board memory devices can be exchanged for pre-loaded memory devices that match the destination of the vehicle. The offloading spots can be placed at locations where vehicles may be parked as a shopping center parking lot, a street parking spot, or at the user’s home place.

The service provider selects the offloading spots consisting of the loading stations which transfer the data to the memory of the vehicle and the memory swap stations which replace the on-board memory of the vehicle. Vehicle memories can be loaded with data while parking at the offloading spot or exchanged for ready-to-ship memory devices so that the users can continue their travels without waiting for the data to be loaded. The selection of the offloading spots is based on the geographic location of the vehicle and its planned destination. The service provider also monitors the status of the offloading spots including the available parking space, memory exchange for the destination of the data available for shipment. The service provider also periodically queries the vehicles over the data network to determine the current geographic location and destination of the vehicles. The vehicle's positioning system includes a navigation system which generates routes and guidance between a geographic location and a destination. The historical locations and addresses are stored in a geographic location database managed at the service provider’s control center. The service provider also keeps a record of the status of the offloading spots in a specific database. The service provider matches the destination of the vehicles to the selected offloading spot based on parking space availability. If preloaded memory devices are ready to be shipped, the service provider checks for exchange at the offloading spots or contacts the content provider to transfer the data to be loaded on the memory devices.

5. A DTN-Based Sensor Data Gathering for Agricultural Applications (Ochiai,

2011): Agricultural researchers and farmers deploy sensors at their remote agricultural-fields to obtain data of temperature, humidity, soil moisture and so on. Automatic collection of data extremely helps their analytical work. Currently, network providers such as cellular phone networks can achieve such automatic collection from remote sites. However, this approach is not feasible for most of the farmers due to operational cost.

In this paper, the authors have explored a delay-disruption tolerant networking (DTN)-based system which collects sensor data from remote sites without relying on network providers. With the DTN approach, it uses the mobility of tractors and farmers, and allows the collection of data from those sparsely deployed sensors. They assumed that sensors, vehicles (e.g., tractors, farmers) and the homes of farmers have short-range radio communication devices. Farmers and researchers use their tractor or their foot to visit their remote sites regularly: e.g., every day or every week. The radio device on such vehicle's contacts with sensors at the remote sites and returns to their home position. There is a data server at their home position that collects the data from their remote sites. For example, when the tractor arrives at the remote sites, it gets the history record of the sensor data and saves it into its local storage. After return, it transfers the stored data to the data server. Fig 6 depicts the overview of the data collection mechanism from remote agricultural fields.

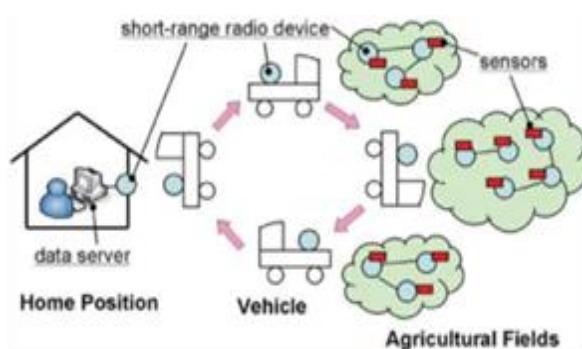


Figure 6 By using the movement of vehicle, DTN allows data collection from remote agricultural field sensors

In this work, they applied potential-based routing (PBR) approach to enable autonomous data transfer. Fig 6 shows how PBR allows autonomous data delivery from sensors to the data server by using vehicle. In PBR, they introduced potential (i.e., a scalar value which represents a heuristic proximity from the destination). By setting higher potential to the sensors and lower

potential to the server, the vehicle receives data from the sensors at the remote sites and provides data to the server when returned. In this way, sensor data is gathered at the data server with the use of movement of the vehicle.

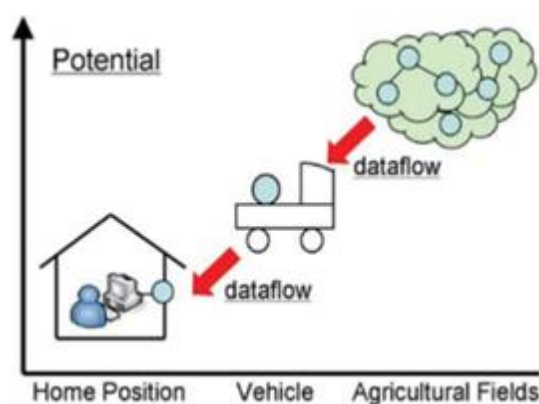


Figure 7 By setting higher potential to remote sensors and lower potential to data server, sensor data autonomously gathers at lower potential nodes, just as water flows from higher places to lower places.

They performed an experiment with 39 DTN nodes from UTMesh (wireless mesh networking testbed at the University of Tokyo) nodes. They installed their experimental software for disruption tolerant networking into these nodes and deployed them into the campus of the University of Tokyo. The software implemented in PEAR (Potential-based Entropy Adaptive Routing) for inheriting the concept of PBR is about 3000 lines in C source code. The footprint was about 51 Kbyte, which indicates it can run on embedded computers. They confirmed that data from sensors were transferred effectively by using the movement of vehicles. In their experiment settings, almost all data (99.8%) were delivered after a 10–75 min delay.

6. A Social Content Delivery Network for Scientific Cooperation: Vision, Design, and Architecture (Chard, 2012): Due to the rapid growth of scientific data, collaborative researchers struggle to manage and organize big data. More precisely, in contexts of collaborative research environment, data must be carefully distributed, analyzed, stored, and accessed by collaborators within their communities.

In this paper, the authors presented a possible solution of Social Content Delivery Network (S-CDN) in which researchers can collaboratively store and access research datasets. The so-called S-CDN implements upon scientific computing and eScience; and acts like a CDN model by replicating scientific data to edge nodes in order to provide faster access, process, and share for researchers.

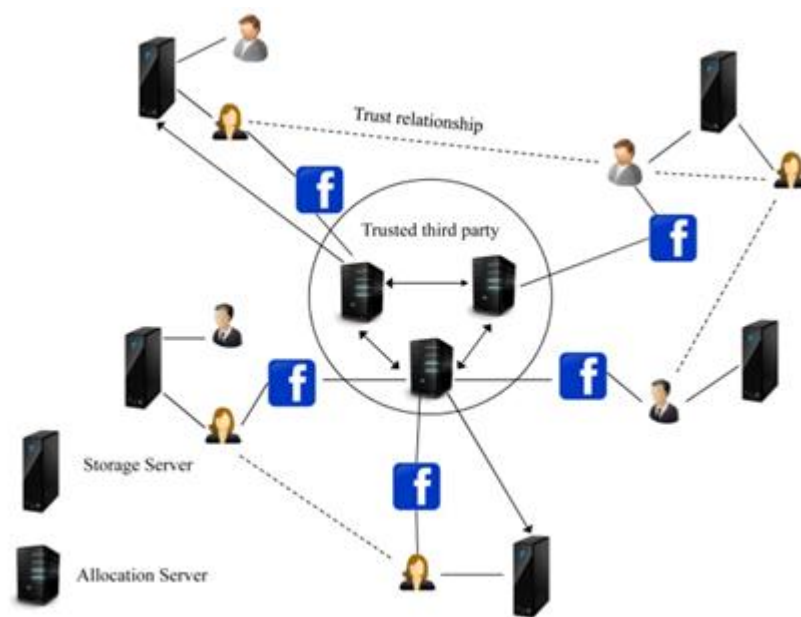


Figure 8 Architecture of S-CDN

Fig. 8 illustrates the architecture of S-CDN with four main components namely a Social Network Platform, Allocation Servers, Individual Storage Repositories, and Social Middleware. In S-CDN model, members of a scientific community must contribute one or more storage services through their social network platform. Then, the storage repositories act as CDN edge nodes where research datasets are cached. After, one or more allocation servers acts as catalogs in which a list of replicas are maintained, updated, and distributed. In addition, social middleware is used in terms of providing authentication and authorization for the platform between users and S-CDN (i.e., allocation servers and storage services).

The authors evaluate the feasibility of S-CDN architecture with the use case based on medical image analysis. However, the authors in this paper need to take the defect of allocation servers into account. Since the allocation servers are operated by third party, it can encounter any outages or malfunctions. This in turn takes a long time to troubleshoot for this problem.

7. Software Defined Network based Adaptive Routing for Data Replication in Data Centers (Kanagavelu, 2013): Recently, data volume has been increasing exponentially in today's Data Centers which has a variety of services such as Web search, e-commerce, storage backup, large scientific applications, video streaming, data analytics and social networking. In such Data Centers, data availability is a major challenge due to failures of server, network, and storage. Hence, data replication is an effective approach to achieve high data availability. Data replication is a technique designed to replicate data at two or more storage nodes in a Data Center.

However, the traffic load must be balanced among the paths in a Data Center network in terms of achieving low latency and high throughput for data replication.

In this paper, the authors addressed the problem of data replication in Data Center networks. Accordingly, they developed an SDN-based framework to minimize congestion in effective data replication. Fig 9 depicts the SDN-based data replication framework in which multi-routed data center network is considered.

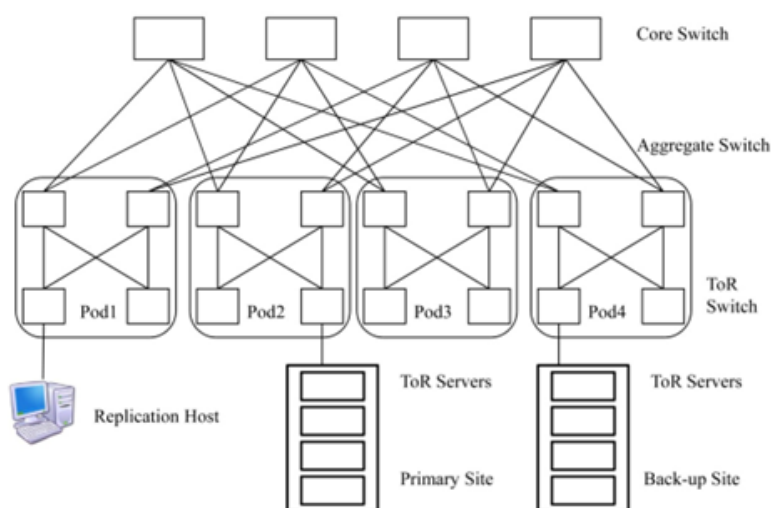


Figure 9 SDN based Data Replication Architecture

In Fig. 9, the network consists of switches interconnected in three tiers such as Top of Rack (ToR), aggregate, and core. In addition, other entities are also shown in Fig 8 such as pods, replication host (which begins the replication operation), primary site (which stores primary copy of data in a server), and backup site (which stores backup copy of data in a server). In this framework, a replication operation generates based on two flows; one flow is between the host and primary server and the other flow is between the host and backup server.

Then, the authors also implemented an adaptive routing method which is a key component in the data replication framework. With this method, they monitor the load statistics at the switches and select the appropriate routes based on the current network state. As a regard to the appropriate routes, they choose the least-loaded path with minimum load among other paths. The shortest paths are also used as candidate paths. They also select the path with minimum cost. In this sense, if there are one or more paths with the same minimum cost, the path with maximum capacity is considered as first priority. By considering the load in network, traffic load can be balanced, and network congestion can be reduced which incurs to high throughput.

Next, the authors designed a prototype by using OpenFlow-enabled switches and performed experiments to examine the effectiveness of the proposed framework. However, although SDN technique enables optimizing network utilization, it can sometimes have the risks of software bugs and result in unreliable network links. Regarding this issue, SDN controllers must be frequently upgraded in order to fix bugs and improve performance.

8. Towards A Scalable Scientific Data Grid Model and Services (Abdullah, 2008):

In Scientific Data Grid environment, it provides geographically distributed resources for scientists in order to access large scientific datasets. However, since the number of scientists is considerably increasing in such an environment, it tends to be a scalability problem.

The researchers in this paper addressed the scalability problem in Scientific Data Grid. To deal with this problem, they presented a P2P Scientific Data Grid model in terms of providing scalable service for Scientific Data Grid environment.

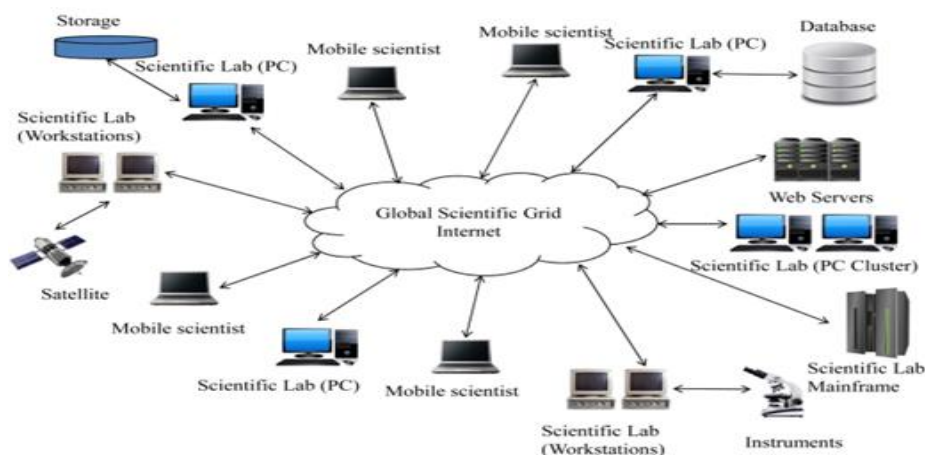


Figure 10 P2P Model for Scientific Data Grid

Fig. 10 demonstrates the proposed model for a Scientific Data Grid which supports scientific collaboration based on P2P services. In Fig 9, peers could be any network devices such as PCs, servers, and even supercomputers. Then, each peer can be self-organized to form a peer group. Once a peer joins a group and it can utilize all the services provided by the group. In this model, peers can access their required datasets without concerning which peers deliver the datasets.

The authors then evaluated the proposed model by using their own simulation written in PARSEC. PARSEC is a C-based discrete event simulation. Through the simulation results, the P2P model can provide a scalable service for the Scientific Data Grid Environment.

Discussions

As the emergence of technologies and applications, a large amount of data being generated exponentially from different resources such as social media network, sensor devices, and other information serving devices is leading to become big data. The traditional database systems are not capable of storing and processing such a large, complex, and growth of data. As a result, cloud computing can provide elastic scaling of computing power and storage resources for handling big data. However, moving large volumes of data to and from is a major challenge since it can be encountered undesirable latency. Dealing with big data movement problems, this paper extensively reviews the recent research works to collect possible suggestions and solutions. According to the findings from review, many researchers proposed various methods including frameworks and network techniques which support efficient methodologies to be applied in big data movement problems.

Knowledge from Research

Through the extensive review, some papers exploited various network techniques such as DTN, CDN, SDN, and data grid for big data movement. The concept of DTN can be taken advantage of in the case of disconnection of communication network occurring in disaster zone, rural areas, and developing countries. In DTN, nodes in the network relay data using a hop-by-hop store and forward method to reach to the destination. By using CDN, it can provide faster access to large datasets by a group of users. With the use of SDN, it can optimize the network performance by controlling programmatically for big data delivery. The architecture of data grid is also designed to provide geographically distributed resources and large datasets accessed by communities. The proposed methods and technologies from the review can be applicable for moving big data to and from the cloud for storing and processing.

Conclusion and Suggestions

Big data is the result of the rapid generation of vast amounts of data from various sources, including social media networks, sensor devices, and other information-serving devices, as a result of the introduction of new technologies and applications. The volume, complexity, and expansion of data that is being stored and processed is beyond the capacity of traditional database systems. Therefore, for managing large amounts of data, cloud computing can offer elastic scaling of

processing power and storage resources. However, there may be undesired latency, transporting massive amounts of data to and from can be quite difficult. To address issues with the transfer of huge data, this paper reviews various network techniques. As per the broad review, numerous researchers proposed diverse approaches such as DTN, CDN, SDN, and data grid to facilitate the application of effective methodologies in large data movement difficulties. By implementing these suggested network techniques, it can improve the performance and reliability of transferring large amounts of data between two platforms.

However, the research papers did not consider the uncertainty issue of data size. Due to the nature of big data, the growth of data size come from various sources is exponentially increased over time. If the data size is increased, it can be challenging to have optimal control of servers at the remote site for allocating data in it and also provisioning bandwidth for moving data. Therefore, the direction towards future work will be explored the research challenge of big data movement under uncertainty scenarios such as size of data and network conditions.

References

- Abdullah, A. O. (2008). *Towards a scalable Scientific Data Grid model and services*. International Conference on Computer and Communication Engineering.
- Antony, T., & Paul, S. (2014, February). Addressing big data with Hadoop. *International Journal of Computer Science and Mobile Computing*, 3(2), 459–462.
- Big Data. (n.d.). Retrieved from Big Data: http://en.wikipedia.org/wiki/Big_data
- Brown, P., Zhu, M., Wu, Q., Yun, D., & Zurawski, J. (2012). *Exploring the optimal strategy for large-scale data movement in high-performance networks*. In Performance Computing and Communications Conference (IPCCC), (pp. 181–182).
- Chard, K. C., Caton, S.J., Rana, O.F., & Katz, D.S. (2012). A social content delivery network for scientific cooperation: Vision, design, and architecture. In Conference: Third International Workshop on Data Intensive Computing in the Clouds (DataCloud 2012) (in conjunction with SC12). DOI:10.1109/SC.Companion.2012.128
- Data Analysis Challenge*. (2008). The MITRE Corporation.
- Fall, K. (2003). A delay-tolerant network architecture for challenged internets. In Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '03), (pp. 27–34). <https://doi.org/10.1145/863955.863960>

- Garrett, B. (2013). *Big data is changing your world...more than you know*. Atlantic Council.
Retrieved from https://econvue.com/sites/default/files/Big_Data_is_Changing_Your_World.pdf
- Gorcitz, R. A., Jarma, Y., Spathis, P., de Amorim, M. D., Wakikawa, R., Whitbeck, J., Conan, V., & Fdida, S. (2012). *Vehicular carriers for big data transfers*. In Vehicular Networking Conference (VNC). Retrieved from <https://hal.science/hal-00739361/document>
- Kanagavelu, R. L. (2013). *Software defined network based adaptive routing for data replication in Data Centers*. In 19th IEEE International Conference on Networks (ICON).
- Laoutaris, N. S. (2009). Delay tolerant bulk data transfers on the internet. In *ACM SIGMETRICS Performance Evaluation Review*, 37(1).
- Ochiai, H. I., Ishizuka, H., Kawakami, Y., & Esaki, H. (2011). A DTN-based sensor data gathering for agricultural applications. *IEEE Sensors Journal*, 11(11), 2861–2868.
- Rawat, R., & Yadav, R. (2021). *Big data: Big data analysis, issues and challenges and technologies*. In IOP Conference Series: Materials Science and Engineering, 1022(1), 012014. DOI:10.1088/1757-899X/1022/1/012014
- Sim, A. (2010). *Bulk data movement for climate dataset: Efficient data transfer management with dynamic transfer adjustment*. Lawrence Berkeley National Laboratory.
- Soi, S. (2013, October). Big data and methodology—a review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(10), 991–995.
- Song, Y.-S. (2012, December). *Storing big data—the rise of the storage cloud*. Retrieved from www.seamicro.com
- Upadhyay, S., Manwani, R., Varshney, S., & Jain, S. (2021). *Analytics and storage of big data*. In Presented at the International Semantic Intelligence Conference (ISIC 2021).