Cyberinfrastructure: Applications and Challenges

Qiuhui Tong, Bo Yuan and Xiu Li
Intelligent Computing Lab, Division of Informatics
Graduate School at Shenzhen, Tsinghua University
Shenzhen 518055, P.R. China
tongqh@126.com, yuanb@sz.tsinghua.edu.cn, li.xiu@sz.tsinghua.edu.cn

***Abstract*—This paper presents a comprehensive review of Cyberinfrastructure (CI), an emerging collaborative research environment, including its representative applications in four science communities around the world. An in-depth analysis is also conducted to reveal the key functions and desired features that can be expected from modern CI systems.**

Keywords—Cyberinfrastructure; HPC; collaboration; research

I. INTRODUCTION

Cyberinfrastructure (CI) refers to a research environment built on technologies including distributed computing, data storage, information processing and communication, which integrates geographically dispersed researchers or institutes for coordinated research activities and knowledge discovery [1].

In the past, the research accomplishment of a certain institute is mostly enjoyed by itself due to the limitation of information technology and dispersed locations, leading to the lack of communication in the science communities. However, collaboration across institutes and researchers is essential in scientific research, such as information sharing with other entities or access to remote observing networks [2]. Without proper resource sharing mechanism, the research capability of even large institutes may be partly limited by the facilities that the institutes own, let alone independent researchers who may also act as an engine of innovative research at some times [3]. Meanwhile, the science community has witnessed the dramatic improvement of information technology in all aspects in the past decades, making collaborative research technically feasible. It was in this context that the term Cyberinfrastructure was coined in a report of the National Science Foundation (NSF) Blue-Ribbon advisory panel in 2003.

While the research activities in many disciplines are relying more and more on information technologies, the emergence of CI can provide the science and engineering communities including large institutes and independent researchers with access to critical computing resources and essential data repositories. Meanwhile, CI can get individual researchers involved in a large virtual research team to give them access to more and better information and facilities for discovery and learning, further boosting the output efficiency of research. After all, with CI, the focus of research activities can be shifted from data acquisition to data discovery.

Similar to the concept of CI, many other countries have also been working on similar scientific platforms under different names. For example, the e-Science program, launched in the United Kingdom, is dedicated to providing an advanced scientific infrastructure for international research corporation and the storage and curation of large volume data [4]. In China, there is an e-Science website named "Chinese Science and Technology Resource", which is the portal for the Platform of National Science and Technology Infrastructures [5]. The website provides users with scientific databases, the access to large instruments, high performance computing (HPC) resources and knowledge transform platforms. Its main functions include resource management, resource search and navigation and resource monitoring.

Building a CI system generally amounts to a large-scale and open-ended project, which requires significant investment. To reduce the cost and minimize the potential risk, the establishment of CI should fully capitalize on existing facilities in an effort to gradually upgrade them to a functionally complete research infrastructure. The core of a CI system is a series of supercomputer centers. For example, the NSF supported Extreme Science and Engineering Discovery Environment (Xsede, formerly Teragrid) brings together dozens of supercomputers and provides a powerful platform for big data analysis [6]. Nowadays, the rapid improvement of computing hardware, the increasingly ubiquitous networking, better interoperability of information formats, decreasing cost of HPC and data storage and the revolutionary computing and visualization technologies all provide a concrete foundation for CI and expedite the development of CI [4].

In the rest part of this paper, Section II presents the general architecture of CI while representative CI applications in four disciplines are introduced in Section III. In-depth analysis of the functions and features of CI are conducted in Section IV and this paper is concluded in Section V.

II. ARCHITECTURE OF CYBERINFRASTRUCTURE

A typical CI system can be divided into three layers according to the technologies involved (Fig. 1). The foundation layer consists of the well-established and standard supporting information technologies, including computation, data storage and communication. The outer layer contains customized software applications and services for a specific field. The core CI layer is located between these two layers, which capitalizes on mature technologies and provides users a platform to tailor the upper environment to their specific need [4]. To implement a functional CI system, the following aspects are generally

considered the most important components: Computing Facility, Data Storage, Data Analysis and Visualization, Virtual Organizations for Distributed Communities, Learning and Workforce Development [7].

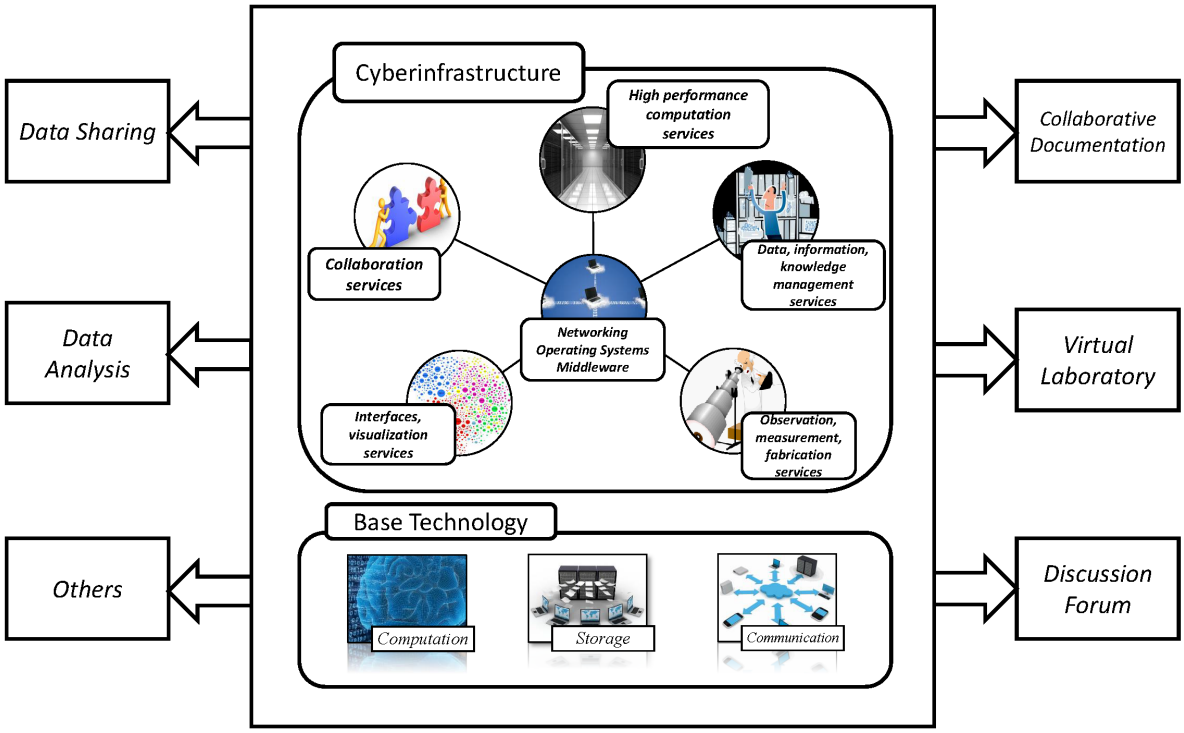

Fig. 1. The architecture of Cyberinfrastructure

A. High Performance Computing

Computing is at the central of all numerical simulations, from climate modeling to fluid dynamics research. The performance of electro-optical components has been increasing at exponential rates according to Moore's Law. Currently, China's Tianhe-2 supercomputer is the fastest supercomputer in the world, running at 33.86 PFLOPS [8]. With the access to state-of-the-art HPC facilities, researchers can now run simulations that were too computationally expensive to run in the past and/or execute many more simulations within the same time period to better explore the problem of interest, achieving unprecedented accomplishments. However, HPC resources are often limited for independent researchers or small institutes and HPC assets at separated locations seldom have active communication with each other. With the deployment of CI, different HPC nodes can actively communicate with each other, resulting in increased utilization rate. Furthermore, CI can provide users with easy access to different HPC nodes according to their specific research requirements to find the most suitable HPC facility.

Given the fact that, in the past decades, the HPC hardware has been consistently upgrading at fast pace, a supporting software layer, which can insulate the implementation in the application layer of CI from the fast evolution of HPC hardware, is strongly desired so that the user interface can be more likely to stay stable. In other words, the underlying HPC functionality should be transparent to the domain-specific applications. The collaboration and coordination among different HPC agencies are also needed to make more diverse and richer systems available to researchers and institutes.

B. Data

People now often find themselves living in a data-intensive world thanks to the proliferation of information technologies. Researchers in all areas produce, access, analyze, integrate and store terabytes of data on a daily basis through experimentation, observation and simulation [7]. To avoid the risk of being inundated in the large bulk of data or being compromised by low quality data, there is a need of proper data management technology. More specifically, effective techniques for integrating, analyzing and mining the large reservoir of data to catalyze the process of research are playing a key role in data-driven domains. Within CI, the volume, the variety as well as the velocity of data will rocket to a new level in the near future, which requires much more effort devoted to the management of data.

Instead of working on isolated data from very limited resources in certain proprietary format, in CI, data of various types can be collected from a wide range of experiment instruments or sources and a unified collection policy is needed. For example, data curation is applied to manage and validate the data and make the data reusable in the long term. Also, careful consideration should be paid to promote the interoperability of data, which eases the task of data sharing. Furthermore, the maintenance of high quality metadata (the data containing essential information of the research data including the content, context, structure, interrelationships and

provenance) is critical for large-scale data collection to ensure the efficiency of data processing and the data quality [7, 9].

Upon the completion of data collection, easy access and effective access control are needed as CI must guarantee that a large volume of data can be smoothly accessed by all legal researchers and the privacy of the users can be properly protected. While allowing researchers to freely obtain the very information or data they want, the privacy of certain types of data should also be enforced in light of ethical or legal issues. Due to the quantity of data in CI, efficient data mining techniques are required to help users extract the most useful contents they need without getting lost in the large bulk of data. In the meantime, visualization techniques can help users understand the data in an intuitive manner, helping researchers identify the overall pattern or trend of data so that to accelerate the knowledge discovery procedure.

III. Applications of Cyberinfrastructure

A. *Biology*

Biology is well known for dealing with large volume of data, from genomics, physiology, population genetics to imaging. With the assistance of computers, biologists can get access to informative datasets and deal with complicated computation. For example, the structure of proteins, which is the foundation for an array of functions within living organisms, can be estimated via simulations regarding the folding of polypeptide chains or the repelling and attraction among amino acids.

However, advanced information technologies are often inaccessible to all but a few scientists, which can give rise to the imbalance of development. Still, much more information and more complicated models are required by the biology community to keep up with the rapid development in this domain. With the deployment of CI, biologists are given the opportunities to benefit from advanced HPC entities to conduct more complex computation and modeling work and be more informed with the information from different researchers at distributed locations. With CI, even independent researchers with limited resources can get access to top research facilities and abundant data, which will promote the progress of the whole society. The significance of CI in biology science has been appreciated by many countries, which have devoted significant efforts to establishing CI platforms for the biology community. iPlant is one of the existing CI implementations in biology, which will be introduced in the next.

iPlant Collaborative [10], established by NSF in 2008, aims to equip researchers in life science with powerful computational infrastructure for handling huge datasets and complex analysis. iPlant provides users with state-of-the-art facilities including HPC, data storage, analysis tools and workflows, visualization, image analysis and educational resources [10]. iPlant supplies tools to assemble and annotate sequences, analyze phenotypes, and integrate environment data. iPlant can help biology researchers perform complicated analysis on large datasets without the need to master detailed command lines [10]. Since biologists may need a growing number of customized applications, iPlant is an open-source project, allowing users to develop applications catering to their specific need. In light of the cost and efficiency issue, iPlant fully takes advantage of existing tools and facilities and establishes a broader and more sophisticated platform [10]. iPlant leverages both its own hardware resources as well as the NSF Xsede hardware. iPlant also makes use of massive computational and data storage systems created by Texas Advanced Computing Center (TACC) and other Xsede service providers.

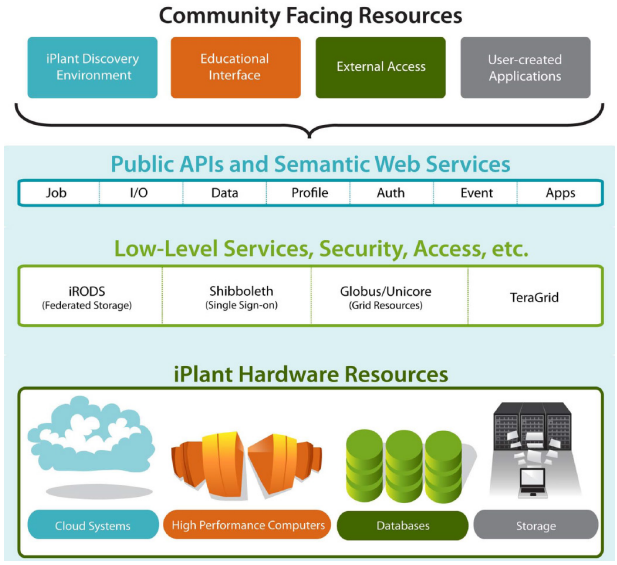

Fig. 2. The architecture of iPlant [10]

The architecture of iPlant is shown in Fig. 2. The request for HPC resources from a scientist can be handled through Discovery Environment (DE), the primary web portal to the powerful computing, storage, and analysis application resources of iPlant. For data management, the contributor of the data can determine the access authority of the data while metadata can be edited through DE or API using iCommands. Many data analysis tools are available in DE and Atmosphere (cloud service of iPlant), while a discussion forum is available for users to discuss problems and seek for assistance [10]. For security concerns, data is replicated among Texas, San Diego and the University of Arizona via iRODS (integrated Rule-Oriented Data Management System), which is adopted as the data management middleware [11].

B. *Geoscience*

Geoscience is consistently faced with a large volume of data and is closely related to a number of disciplines such as geology, physical geography, geophysics, geodesy, soil science, ecology, hydrology, glaciology and atmospheric science. Note that observation stations are often geographically dispersed and data collected from different stations would be of little value without proper data fusion and sharing facilities.

However, the refined data and instruments are often confined to a certain group of researchers or institutes, which can hamper the development of the community. Compared to biology, the data resources are located at more diverse and separated locations and reliable and high-speed network connections are required. Within CI, geo-scientists can get

access to a variety of observation data from different spots easily, which can produce a more comprehensive picture of the topic of interest. More analysis tools and HPC resources are also available for the community with the adoption of CI, providing equal opportunities for both the large institutes and individual geoscience researchers. Furthermore, CI can help to upgrade the data-rich geoscientific research environment to an analysis-rich environment [12], which may create profound impact on geoscience.

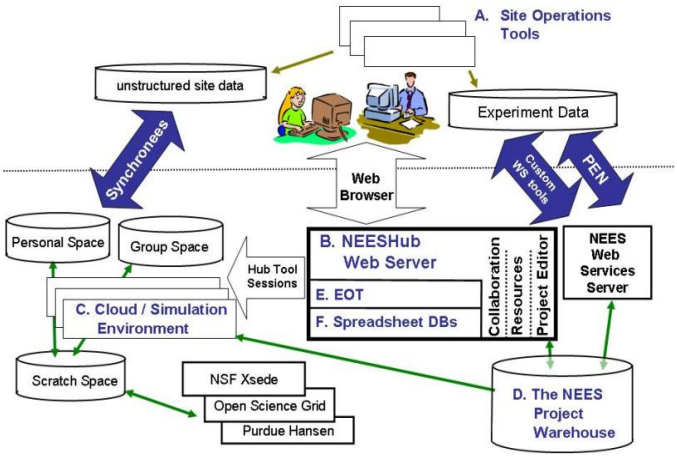

Fig. 3. The architecture of NEES [16]

Geo-Cyberinfrastructure can foster seamless integration of heterogeneous geo-information, web-based mapping and geo-analytical services across the Internet [13]. The Network for Earthquake Engineering Simulation (NEES) is one of the successful examples of CI in geoscience.

NEES was launched by NSF in 1999 with the vision of "shifting the emphasis of earthquake engineering research from current reliance on physical testing to integrated experimentation, computation, theory, databases, and model-based simulation" [14]. NEES provides a new paradigm where earthquake research and education become a collaborative effort among the community rather than a collection of loosely coordinated research and education projects by individuals [15]. NEES is pioneering the mission to reduce the impact of tsunamis and earthquakes on society through research, engineering, science and education.

Fig. 3 illustrates the workflow of NEES. In the heart of NEES is NEEShub, a website that provides functions such as managing experimental and simulation data, running simulations using data stored in NEEShub, providing contents and tools for learning and outreach as well as facilities for developing and sustaining a virtual community [16, 17]. The research of NEES can be shared through NEEShub with engineers, researchers and practitioners around the world. Some of the NEES servers are placed at individual sites to collect and process data from simulations, while the website servers and data servers are located at Purdue University [17]. Users can upload their own data and share them with the community through the website, which supports various simulation and visualization tools without the need to download or install any codes. Furthermore, for large computing work, users can get access to HPC platforms including the Open Science Grid and Xsede [17]. For security consideration, NEEShub deploys five classes of security controls: access control, user authentication, network intrusion detection, host intrusion detection and diligent maintenance of the software infrastructure [16].

C. *Oceanography*

Ocean plays an important role in our daily life. For example, it is one of the key regulators of climate change and a booster for economy for its abundant resources [18]. Ocean is also a sophisticated dynamic system, which calls for the introduction of consistent observation. With advanced instruments installed permanently in the ocean that can communicate with researchers on the land, data can be collected continuously with very short latency by observatory networks from different observation spots. Oceanography produces a vast volume of data in each second and CI can provide users with up-to-date data stream in a proper format to be fed into simulation or modeling tools.

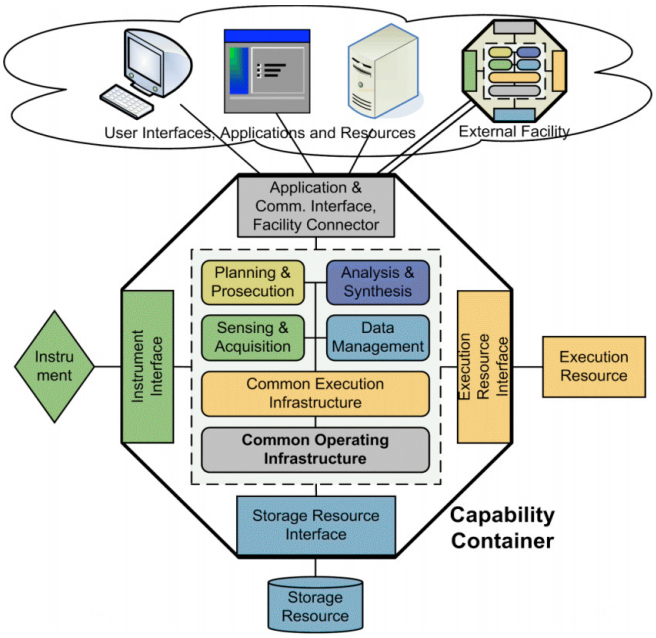

Fig. 4. The architecture of OOI [19]

The Ocean Observatories Initiative (OOI) , funded by NSF, embarks on a new era of ocean observation [19]. It is planned as an environmental observatory, covering a great diversity of oceanic environments, including the physical, chemical, geological and biological variables in the ocean and seafloor, to enable interactive observation [19, 20]. The core capabilities and principal objectives of OOI are collecting real-time data, analyzing and modeling data on multiple scales and enabling adaptive experimentation within the ocean [20]. The functions of OOI are grouped into six service networks (Fig. 4): Sensing and Acquisition, Data Management, Analysis and Synthesis, Planning and Prosecution, Common Execution Infrastructure and Common Operating Infrastructure. The first four service networks are viewed as the application-supporting service networks while the last two are the infrastructure ones.

Instrument management, tasks execution, and data collection are executed by the Sensing and Acquisition service network. Data Management service network is responsible for data ingestion, data transformation, as well as data discovery and access. Users can get access to the service of observation analysis and visualization according to their specifications through Analysis and Synthesis service network. Planning and Prosecution service network leverages the capabilities of the integrated network of sensing and modeling and supports resource nesting and autonomy. Common Execution Infrastructure service network supports software package decomposition, deployment, implementation and integration, and provides an environment for specific user-requested purposes. Common Operating Infrastructure service network can bind other service networks into a more coherent whole, and deal with the governance and security issues as well.

D. Social Science

Nowadays, researchers in social science also rely heavily on advanced information technologies to conduct experimental studies and other data-related work. Unlike natural science where most data is contributed by sensors or experiments that usually generate data in a predefined form, social scientists often have to deal with data of more complex nature, which can be multilingual, historically specific, geographically dispersed, and highly ambiguous in meaning [21]. The social science can take a giant step forward with the help of appropriate and usable CI [22]. For example, data from different areas around the world can be gathered via the Internet and a statistical application can be used to assist the data analysis. CI can provide the social science community with more complete data resources, more sophisticated tools as well as HPC facilities.

Compared to the adoption of CI in natural science, the development of CI for social science is still in its early stage. Project Bamboo, initiated in April, 2008, is a CI initiative for arts and humanities. The object of Project Bamboo is to explore the possibility for a distributed system of shared technology services to support the digital humanities [23]. The project is led by humanities computing experts from the University of Chicago and the University of California, Berkeley [23]. The major accomplishment achieved includes the following aspects: cataloging digital tools for humanities scholarship, interoperability of digital collections, proxied access to remotely hosted tools for scholarship, research environments to store, manipulate, and manage digital content as well as identity and access management [24].

Time-sharing Experiments for the Social Sciences (TESS), a testing ground funded by NSF [25], focuses on providing social scientists with new ways of data collection and innovative experiments, in an effort to increase the precision of data measured or understood, to promote the efficiency of the output and to reduce average cost per study. TESS offers researchers the opportunity to capture the internal validity of experiments while taking the benefits of working with large, diverse population of research participants. TESS conducts the general population experiments with an intention to combine the strengths of experimental and survey designs in supporting causal inferences in social science. Time-sharing is another critical feature of TESS, which improves efficiency in data collection by collecting demographic information that all investigators can share and thus reduces the cost [26].

The Inter-university Consortium for Political and Social Research (ICPSR) acts as a global leader in data stewardship and rich data resources provider for social science [27]. ICPSR, a membership-based organization with over 500 member colleges and universities around the world, hosts more than 500,000 files in social science [7, 27]. As for data analysis, Virginia Economics Laboratory (VeconLab) provides around 60 on-line programs for users to conduct experiments mainly regarding the game theory either for teaching purpose or research purpose [7, 28]. All the tools provided by VeconLab can be operated in the online mode via a web browser without the need of downloading.

IV. Challenges and Opportunities

In the above, it has been demonstrated that, for a variety of disciplines, CI can accelerate the progress of research at unprecedented speed with high-quality data, powerful tools and extensive collaboration among dispersed facilities and researchers. As a result, more opportunities can be expected from CI for the science and engineering community to pursue greater achievement. With the employment of CI, large volume of research data, sophisticated analysis tools and HPC resources are available for the entire science community including independent researchers. In this way, researchers can mainly focus on the very issue that they are investigating rather than gaining the essential data and tools.

However, opportunities always come with challenges. As discussed before, data-related activities run through all the key procedures of CI and data management is an essential issue deserving careful consideration. In the era of big data, it is estimated that around 2.5 EB data is created everyday [29]. As the scale of data escalates and the variety of data increases, more efforts must be devoted to the issues of big data for a sophisticated CI, including curation, storage, search, sharing, transformation, analysis and visualization. The privacy of data and close collaboration among dispersed facilities are also important for the healthy development of CI.

A. Big Data Aquisition

Within CI, more people and institutes are able to contribute to the data reservoir, increasing the quantity and complexity of data consistently. The raw data coming into CI can be dirty because the data is collected from different channels among distributed facilities with potential issues such as incompleteness, noise as well as inconsistence, which can result in the low effectiveness of data. In this situation, data preprocessing is important for subsequent analysis and interpretation. The main tasks of data preprocessing in CI include data cleaning, data integration, data transformation, and data redundancy elimination. For example, data coming from different channels may cause the issue of heterogeneity. Still, unstructured data is another problem facing the society. How to transform the unstructured data into a structured format is one of major tasks for CI. A proper representation is desired to combine the diverse data and present users with a unified view and make data meaningful for subsequent analysis [30]. Data

cleaning can identify problematic and abnormal data while redundancy elimination can significantly reduce the cost of storage and communication. However, most existing studies on data cleaning are based on the assumption that there are well-understood errors or constraints [31]. Finally, data needs to be transferred to data storage infrastructure and the larger scale of data requires much higher transmission bandwidth to reduce the latency.

B. Big Data Storage

The storage capability is a huge challenge for the CI system. The consistently incoming massive data puts a stringent requirement on the optimization of storage systems, including the performance of both storage hardware and the managing software. For example, the Sloan Digital Sky Survey (SDSS), one of the most ambitious and influential surveys in the history of astronomy, began collecting astronomical data in 2000 [32]. In the first few weeks, it amassed more data than all the data collected in the history of astronomy. Currently, SDSS has amassed more than 140 terabytes of data, and in the future its successor is anticipated to acquire that amount of data every five days [33].

For a comprehensive CI system, a large scale distributed storage system with multiple servers can reduce the workload for a single storage facility. However, new problems will be introduced by distributed storage. For instance, multiple servers are connected by a computer network for which failure is inevitable. In this case, effective mechanism for fault tolerance and data consistency among different copies of the same data in different storage nodes is urgently required for the distributed storage system [30].

C. Big Data Analysis

Different types of data come from diverse resources in CI: sensors, researchers or instruments. However, without human insight, data is just a bunch of numbers and analysis needs to be conducted to discover the hidden value of data. Unfortunately, researchers can easily be overwhelmed by explosively growing data without effective data analysis tools. The solution to helping researchers figure out the exact data that they need from the large data reservoir in CI relies on big data analysis. Data mining is the key procedure for data analysis whose main purpose is to extract useful information from a data set and transform it into an understandable structure for further use.

In CI, data often comes in the form of data stream and it is often not feasible to locally store a stream in its entirety [34]. For example, real-time data analysis is needed to annotate objects from the instant video captured by AUV (Autonomous Underwater Vehicle) or to identify abnormal conditions from data collected by the seafloor sensors with a short latency. Furthermore, current data mining models are mostly applied in a fully automatic manner with little human intervention. However, with advanced visualization techniques, better performance can be achieved by combining the efficiency of automation and the flexibility of human control. Interactive data mining can help researchers acquire deeper insights and understanding of the data while human feedbacks can be used to improve the accuracy of data mining models [35].

To handle the massive data that may be found in CI, there is a need of not only appropriate analysis techniques but also competent HPC facilities. Traditionally, data mining algorithms are designed without taking parallel computing into consideration. In recent years, the volume of data to be processed is well beyond the capability of single-CPU systems. Cluster computing, MapReduce and GPU computing are among the most promising HPC paradigms for handling massive data, featuring proven performance in various applications. In particular, GPU computing is the most cost effective and energy efficient HPC option.

D. Data Confidentiality

The extensive sharing of data in CI poses serious issues regarding the data privacy due to legal and ethical considerations [30]. Without sufficient privacy protection, users would be reluctant to share their data. Note that CI needs to ensure the protection for both the contributors of the data and the users of the data. Confidentiality must be supported and reinforced by relevant policy and legal frameworks and technical assistance [36]. However, data privacy is a multifaceted property according to the specific privacy being sought and different technologies are required for different objectives [37]. In recent years, a number of techniques have been proposed to preserve the privacy of data. Most methods apply some form of transformation on the data, which may lead to granularity reduction. However, the reduction in granularity can result in lower effectiveness of data management. In this case, there is a trade-off between the information completeness and the privacy preservation.

E. Virtual Organization

In a CI system, geographically distributed researchers or institutes can form a Virtual Organization (VO) to conduct research collaboratively. A VO can encompass entire communities of scientists instead of a single group focusing on a particular instrument or database [38]. Members in a VO can collaborate on experiment designs, execution and post-analysis, information sharing and data management using collaborative tools, just like working in the same laboratory. Appropriate visualization and virtualization techniques can also help researchers seamlessly work together on the same object or model. The domain-specific software applications and tools can also be made available for the legal members in the VO. In this case, the heterogeneity of VO requires an interoperable environment to make members effectively collaborate with each other.

Furthermore, seamless communication is needed to make sure that members working on the same topic are staying on the same page. Reliable Internet-based audio or video conference services, such as iChat and Skype, are helpful for members to catch up with each other. Documentation is another way for members to keep informed. Popular collaborative document editing applications include Google Docs, EtherPad and Microsoft Office with SharePoint. Physical instruments can also be shared in a VO as members of a VO can operate the instruments or sensors remotely through tele-observation and tele-operation tools. Finally, since a VO encompasses a large number of diverse members, a flexible user interface is needed to enable easy discovery and learning.

V. Conclusion

Cyberinfrastructure, a collaborative research environment linking researchers and institutes, can boost the efficiency and effectiveness of research activities with shared research data, sophisticated tools and close collaboration among institutes or researchers who were traditionally isolated from each other. In this paper, we give a comprehensive review of CI including the general picture of CI and its representative applications in four science communities. It is demonstrated that CI is becoming more and more essential for conducting large scale research as well as accelerating the development of the entire research community. It is not simply a technique for solving a specific research problem. Instead, it aims to revolutionize the way how research is conducted.

Note that CI itself is not an independent technical framework and is in fact closely related to many popular information technologies such as cloud computing, Internet of Things, distributed data storage, visualization, parallel computing and virtual reality. As a data-intensive system, there are also many data-related problems that need close scrutiny, especially in the dawn of big data. This paper summarizes some key technical issues including data acquisition, data storage, data analysis, data confidentiality as well as virtual organization, which will play a key role in the future development of CI.

Acknowledgment

This paper is supported by National Natural Science Foundation of China (NSFC, Project No.: 71171121).

References

[1] S. Wang, "A cyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis," Annals of the Association of American Geographers, vol. 100, pp. 535-557, 2010.

[2] D. J. Wright and S. Wang, "The emergence of spatial cyberinfrastructure," Proceedings of the National Academy of Sciences, vol. 108, pp. 5488-5491, 2011.

[3] K. H. Buetow, "Cyberinfrastructure: empowering a "third way" in biomedical research," Science, vol. 308, pp. 821-824, 2005.

[4] D. Atkins, K. Droegemeler, S. Feldman, H. Garcia-Molina, M. Klein, D. Messerschmitt, et al., "Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure," NSF, 2003.

[5] Chinese Science and Technology Resource. Available: http://www.escience.gov.cn/eng/general.htm

[6] R. LeDuc, M. Vaughn, J. M. Fonner, M. Sullivan, J. G. Williams, P. D. Blood, J. Taylor, W. Barnett, "Leveraging the national cyberinfrastructure for biomedical research," Journal of the American Medical Informatics Association, vol. 21, pp. 195-199, 2013.

[7] Cyberinfrastructure Council, "Cyberinfrastructure vision for 21st century discovery," NSF, 2007.

[8] Top 500 Supercomputers. Available: http://www.top500.org

[9] P. E. Carter and G. Green, "Networks of contextualized data: A framework for cyberinfrastructure data management," Communications of the ACM, vol. 52, pp. 105-109, 2009.

[10] iPlant. Available: http://www.iplantcollaborative.org

[11] S. A. Goff, M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, et al., "The iPlant collaborative: cyberinfrastructure for plant biology," Frontiers in Plant Science, vol. 2, 2011.

[12] P. Yue and L. He, "Geospatial data provenance in cyberinfrastructure," in 17th International Conference on Geospatial, pp. 1-4, 2009.

[13] T. Zhang, M.-H. Tsou, Q. Qiao, and L. Xu, "Building an intelligent geospatial cyberinfrastructure: An analytical problem solving approach," in 14th International Conference on Geoinformatics, pp. 64200A-64200A-14, 2006.

[14] B. Spencer, T. Finholt, I. Foster, C. Kesselman, C. Beldica, J. Futrelle, et al., "NEESgrid: a distributed collaboratory for advanced earthquake engineering experiment and simulation," in 13th World Conference on Earthquake Engineering, 2004.

[15] I. Buckle and R. Reitherman, "The consortium for the George E. Brown J. network for earthquake engineering simulation," in 13th World Conference on Earthquake Engineering, 2004.

[16] T. J. Hacker, R. Eigenmann, S. Bagchi, A. Irfanoglu, S. Pujol, A. Catlin, et al., "The neeshub cyberinfrastructure for earthquake engineering," Computing in Science & Engineering, vol. 13, pp. 67-78, 2011.

[17] G. P. Rodgers and R. Thyagarajan. NEES Cyberinfrastrucuture. Available: https://nees.org/collaborate/wiki/NEESCyberinfrastructure

[18] P. Favali and L. Beranzoli, "Seafloor observatory science: A review," Annals of Geophysics, vol. 49, pp. 515-567, 2006.

[19] OOI. Available: http://oceanobservatories.org/about/

[20] A. Chave, M. Arrott, C. Farcas, E. Farcas, I. Krueger, M. Meisinger, et al., "Cyberinfrastructure for the US Ocean Observatories Initiative: Enabling interactive observation in the ocean," in Oceans 2009-Europe, pp. 1-10, 2009.

[21] J. Unsworth, "Our Cultural Commonwealth: the report of the American Council of learned societies commission on cyberinfrastructure for the humanities and social sciences," Technical Report, ACLS, 2006.

[22] F. D. Berman and H. E. Brady, "Final report: NSF SBE-CISE workshop on cyberinfrastructure and the social sciences," NSF, 2005.

[23] S. Katz. Project Bamboo. Available: http://chronicle.com/blogPost/humanities-cyberinfrastructure-project-bamboo/6138

[24] Project Bamboo. Available: https://wikihub.berkeley.edu/display/pbamboo/About+Project+Bamboo

[25] The TEES Opportunity for the Social Sciences. Available: http://www.knowledgenetworks.com/accuracy/summer2009/Freese-summer09.html

[26] TESS. Available: http://www.tessexperiments.org

[27] ICPSR. Available: http://www.icpsr.umich.edu/icpsrweb/content/membership/about.html

[28] VeconLab. Available: http://veconlab.econ.virginia.edu

[29] Big Data. Available: http://www-01.ibm.com/software/au/data/bigdata/

[30] M. Chen, S. Mao, Y. Zhang, and V. Leung, Big Data: Related Technologies, Challenges and Future Prospects: Springer, 2014.

[31] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," Proceedings of the VLDB Endowment, vol. 5, pp. 2032-2033, 2012.

[32] The Sloan Digital Sky Survey. Available: http://www.sdss.org/

[33] Big Data. Available: http://en.wikipedia.org/wiki/BigData

[34] L. Golab and M. T. Özsu, "Issues in data stream management," ACM Sigmod Record, vol. 32, pp. 5-14, 2003.

[35] Y. Zhao, "Interactive data mining," Ph.D. Thesis, University of Regina, 2007.

[36] F. Bonchi, B. Malin, and Y. Saygin, "Recent advances in preserving privacy when mining data," Data & Knowledge Engineering, vol. 65, pp. 1-4, 2008.

[37] J. Domingo-Ferrer and Y. Saygın, "Recent progress in database privacy," Data & Knowledge Engineering, vol. 68, pp. 1157-1159, 2009.

[38] C. L. Borgman, G. C. Bowker, T. A. Finholt, and J. C. Wallis, "Towards a virtual organization for data cyberinfrastructure," in 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 353-356, 2009.