Contents

Vol. 70, No. 2, 2015

Simultaneous English language translation of the journal is available from Pleiades Publishing, Ltd.
Distributed worldwide by Springer. *Astrophysical Bulletin* ISSN 1990-3413.

ISSN 1990-3413, Astrophysical Bulletin, 2015, Vol. 70, No. 2, pp. 123–127. © Pleiades Publishing, Ltd., 2015.
Published in Russian in Astrofizicheskij Byulleten, 2015, Vol. 70, No. 2, pp. –.

Initial-Stage Examination of a Testbed for the Big Data Transfer Over Parallel Links. The SDN Approach*

S. Khoruzhnikov[1], V. Grudinin[1], O. Sadov[1], A. Shevel[1,2], V. Titov[1,3], and A. Kairkanov[1]**

[1] *ITMO University, St. Petersburg, 197101 Russia*
[2] *Petersburg Nuclear Physics Institute, St. Petersburg, 188300 Russia*
[3] *St. Petersburg State University, St. Petersburg, 199034 Russia*

Received August 20, 2014; in final form, December 17, 2014

Abstract—The transfer of Big Data over a computer network is an important and unavoidable operation in the past, present, and in any feasible future. A large variety of astronomical projects produces the Big Data. There are a number of methods to transfer the data over a global computer network (Internet) with a range of tools. In this paper we consider the transfer of one piece of Big Data from one point in the Internet to another, in general over a long-range distance: many thousand kilometers. Several free of charge systems to transfer the Big Data are analyzed here. The most important architecture features are emphasized, and the idea is discussed to add the SDN OpenFlow protocol technique for fine-grain tuning of the data transfer process over several parallel data links.

DOI: 10.1134/S1990341315020017

Keywords: *virtual observatory tools*

1. INTRODUCTION

The Big Data[1] has been a known problem for many years. In each period the term "Big Data" meant a different volume and character of the data. Keeping in mind the "triple V": Velocity, Volume, Variety, we can note that all these features are relative to the current state of technology. For example, in the 1980s the volume of 1 TB was considered a huge volume. In 1824 Charles Babbage won the Gold Medal of the Royal Astronomical Society "for his invention of an engine for calculating mathematical and astronomical tables" with unprecedented accuracy and calculating speed corresponding to the Big Data of that time. In the '70s–'80s, the development of the FITS format standardized the storage, transmission, and processing of scientific and other images, and gave the opportunity to exchange data between different institutes. Nowadays, systems to process Big Data in the modern understanding are being developed and used. The standards for access to big amounts of astronomical data (metadata standards, formats, query languages, etc.), technologies for proceeding Big Data, all these are developed and supported by IVOA (International Virtual Observatory Alliance). The IVOA was created "to facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems, and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory." All astronomical projects dealing with Big Data follow IVOA standards. Such projects include ESO VLT, NOAO CTIO, NASA Kepler, NASA HMS, etc.

There is a range of aspects of the problem: storage, analysis, transfer, etc. In this paper we discuss one of the important aspects of the Big Data: the transfer over a global computer network.

2. SOURCES OF THE BIG DATA

There is a long list of human activities (scientific and business) which are the generators of large volumes of data [1–3]; see the projects SKA,[2] LSST,[3] FAIR,[4] ITER,[5] and also the web sites of CERN[6] and CLDS.[7]

*The article was translated by the authors.
**E-mail: tit@astro.spbu.ru
[1] http://en.wikipedia.org/wiki/Big_data
[2] http://skatelescope.org/
[3] http://www.lsst.org/lsst/
[4] http://www.fair-center.eu/
[5] http://www.iter.org/
[6] http://www.cern.ch/
[7] http://clds.sdsc.edu/

According to [1], the total volume of business emails in the world in the year 2012 was around 3000 PB (3×10^{18} B). The generally accepted estimate for the total volume of stored data is that it grows 1.5–2.0 times each year starting from 2000. In this paper (and for our testings), we will assume that the data volume around 100 TB (10^{14} B) and larger could be labelled as Big Data.

Another aspect of Big Data is the preservation of the data for long periods of time: several tens or more years. Many aspects of our personal, social, or business life, and technical data are now held in digital form. Large volumes of these data need to be stored and preserved. For example, the results of medical tests, data generated by important engines of various kinds (airplane engines, power station generators, etc.), and other data have to be archived for a long time. The same is true for scientific data obtained from experimental measurements on unique experimental installations. The data might be reanalyzed (with new approaches and/or ideas) after the experiment is completed. The preserved data will be kept in distributed (locally and globally) storage. It is assumed that replicas of the preserved data have to be stored in several places on different continents to avoid data loss due to technical, natural, or social disasters.

Historically one of the first fields where Big Data appeared was high energy physics. A number of aspects of data transfer were analyzed and a range of problems was solved. Now more and more scientific and business sectors are dealing (or plan to) with the Big Data [4]. Here is a list of astrophysical/physical projects under development or in operation [5–9]:

- Hipparcos, 1989–1992, 300 GB total volume of data;
- ESO VLT, 1999, the total volume of data is 65 TB and increases by 15 TB per year;
- NASA Kepler,[8] since 2009, 100 GB per month;
- LOFAR (LOw Frequency ARray),[9] 2012, up to 1 PB per day;
- Gaia (Global Astrometric Interferometer for Astrophysics), 1 PB per year;
- Pushchino Radio Astronomy Observatory, all projects, 10–100 GB per day;
- RadioAstron, 1.28 TB per day;
- CERN, all projects, several tens of PB per year;
- LSST (Large Synoptic Survey Telescope), 2020, 10 PB per year;
- ITER (International Thermonuclear Experimental Reactor), 2020, 1 PB per day;
- CTA (Cherenkov Telescope Array), 2015–2020, 20 PB per year;
- SKA (Square Kilometer Array), 2019–2024, 1500 PB per year.

Four of these projects are still under design, and the longer the lead time, the more data are planned to be obtained. The Gaia mission started operating in December 2013. The total volume of the data is planned to be 1 PB per year.

Often the observers have to store only a selected fraction of the obtained data [10]. For a deep analysis, quite often there is a need to distribute the obtained data among the collaborators around the world. That means that a good fraction of experimental data needs to be transferred over the Internet.

3. FREELY AVAILABLE UTILITIES FOR DATA TRANSFER OVER THE NETWORK

The time of transfer over a global computer network (Internet) depends on the real data link bandwidth and the volume of the data. Taking into account that we are talking about a volume of 100 TB and more, we can estimate the minimum required time for the data to be copied over a network link with 1 Gbit capacity. It will give us about 100 MB s^{-1}, hence 100 TB/100 MB s^{-1} = 1 000 000 s = $277^{h}.8 = 11^{d}.6$. During this time the parameters of the network link might change. For example, the percentage of lost network packages can vary significantly. The data link might suffer operation interruptions for different periods: seconds, hours, days.

Now let us look at the Linux kernel network parameters. In the directory `/proc` on Scientific Linux (clone of RedHat) version 6.5, there are about half a thousand parameters describing the network link in the kernel. Not all of them are equally sensitive or influencing on the data transfer process. The most important of them are the TCP window size, MTU, the congestion control algorithm, etc. Of course, the number of independent network links which could be used in parallel is quite important. Of importance also are the network parameters such as the round trip delay time (RTT) and percentage of lost network packages. It is quite obvious that in each large-volume data transfer (running for a long time), we need to be able to tune (to set) a different number of threads, a different size of the TCP window, etc. during the data transfer process to achieve the maximum data transfer speed.

Let us now consider the freely available data transfer utilities which might be used to transfer Big Data over the network and what they permit us to tune.

[8] `http://www.nasa.gov/mission_pages/kepler/`
[9] `http://www.lofar.org/`

3.1. Concepts of Comparison of Data Transfer Utilities

Let us list briefly the parameters to be compared for the data transfer utilities which might help to transfer Big Data.

- A multi-stream data transfer mode, the ability to use several TCP streams in parallel.
- A multi-link data transfer mode, the ability to use more than one data link in parallel; an important feature, especially considering the fact that the available network links are not equal in bandwidth and in conditions (reliability, price, real status, etc.).
- The possibility to set low level parameters, e.g., the TCP window size, etc.
- The method to bypass network problems (errors, timeouts, etc.). In other words, in case of failure of the data transfer, is it possible to continue the transfer after restart?

Ultimately, the data transfer consists of many steps: reading the data from the storage, transferring the data over a network, writing the received data to the storage on a remote computer system. In this paper our attention is concentrated more on the network transfer process.

3.2. Low-Level Data Transfer Utilities

Let us mention several utilities for data transfer over a network (at least some of them are known for many years).

- One of the low level protocols to transfer data over a network is UDT.[10] UDT is a library which implements the data transfer protocol that permits using `udp` but not `tcp`. In some cases the library can help to improve data link usage, i.e., to reduce the data transfer time.
- The RDMA over Converged Ethernet (RoCE) protocol has been studied in [4], and it was found that in many cases RoCE shows better results than UDP, UDT, and the conventional TCP.
- MP TCP[11,12] is an interesting protocol which permits using several data links in parallel for one data transfer. The protocol is implemented as a Linux kernel driver.
- The (Open)SSH family,[13] the well known data transfer utilities, which deliver rigid authentication and a number of data encryption algorithms. Data compression before encryption in order to reduce the data volume to be transferred is possible as well. There are two well known SSH versions: a patched SSH version[14] which can use an increased size of the buffers, and the SSH with the GSI authentication. No real restart after failure. No parallel data transfer streams.
- bbcp,[15] a utility for bulk data transfer. It is assumed that bbcp is running on both sides, i.e., the transmitter as a client, and the receiver as a server.
- bbftp,[16] a utility for bulk data transfer. It implements its own transfer protocol, which is optimized for large files (larger than 2 GB) and is secure, as it does not read the password in a file and encrypts the connection information.
- Xdd [11], a utility developed to optimize data transfer and I/O processes for storage systems.
- fdp,[17] a Java utility for multi-stream data transfer.
- gridFTP,[18] an advanced data transfer utility for the grid security infrastructure (GSI).

Many of them are quite effective for data transfer in terms of link capacity usage. However, Big Data transfer implies a significant transmission time (which may be several hours, days or longer). For long time intervals, it is not easy to rely on such simple transfer procedures. As we mention above, the network link might change the capacity and the percentage of lost network packages, and so on.

3.3. Middle-Level File Transfer Services

The FTS3[19,20] is a relatively new and advanced tool for the transfer of large volumes of data over a network. It has many of the features mentioned above, and more. There is the advanced data transfer tracking (log) feature, the ability to use the http, restful, and CLI interfaces to control the process of the data transfer.

Another interesting development is SHIFT,[21] which is dedicated to a reliable transfer of data over a LAN and WAN. Much attention was paid to the reliability, advanced tracking, performance of the data transfer, and the usage of parallel data transfer between the so-called equivalent hosts (between computer clusters).

[10] `http://udt.sourceforge.net/`
[11] `http://mptcp.info.ucl.ac.be/`
[12] `http://multipath-tcp.org/`
[13] `http://www.openssh.org/`
[14] `http://sourceforge.net/projects/hpnssh/`
[15] `http://www.slac.stanford.edu/~abh/bbcp/`
[16] `http://doc.in2p3.fr/bbftp/`
[17] `http://monalisa.cern.ch/FDT/`
[18] Ibid.
[19] `http://www.eu-emi.eu/products/-/asset_publisher/1gkD/content/fts3`
[20] `https://svnweb.cern.ch/trac/fts3`
[21] `http://fasterdata.es.net/data-transfer-tools/`

3.4. High-Level Data Management Service: PhEDEx

PhEDEx[22,23,24] (Physics Experiment Data Export) is used (and developed) in collaboration around the Compact Muon Solenoid (CMS) [12, 13] experiment at CERN. The experiment does produce a lot of experimental data (in 2013 around 130 PB were written). Data analysis requires copying the data to a set of large computing clusters (about 10 locations in different countries and continents) for analysis and data archiving. Later on, fractions of the data might be copied to smaller computing facilities (more than 60 locations). The total data transfer per day attains 350 TB [13]. It is possible that in the near future the daily volume will increase. Because there may be more than one link between several sites, a routing technique was developed in PhEDEx which permits using alternative routes when the default route is not available.

Finally, the PhEDEx system is quite complicated, and the management service depends on the physics experiment collaboration environment. It is unlikely that PhEDEx can be used in a different environment without redesign.

4. DISCUSSION

The mentioned utilities have several common features:

- all the utilities have a client–server architecture;
- are able to set the buffer size, TCP window size, etc.;
- have the ability to perform various operations before real data transfer and after data transfer, e.g., compression/decompression, using a set of drivers/methods to read/write files to/from secondary storage;
- use a number of authentication techniques;
- use more than one stream, more than one network link for data transfer;
- use several authentication algorithms;
- use a number of techniques to make data transfer more reliable;
- the utilities are not equal in the number of parameters and scope of suggested tasks; some of them are well suited to be used as independent data transfer utilities in almost any environment, others, like PhEDEx (in CMS) and comparable systems in the ATLAS collaboration,[25] are dedicated to be used as part of a more complicated and specific computing environment.

In other words, there is a set of tools which might help in many cases to transfer the Big Data over the networks. Quite a few utilities can use more than one network link.

At the same time, no tool offers fine-grain tuning of parallel data links. Fine tuning is considered as a possibility to apply a different policy to each data link. In general, parallel data links might be completely different in nature, features, and conditions of use. In particular, the usage of QoS is assumed for each network link in the data transfer as well as the ability to change the policy on the fly. All that suggests the idea that a special application is required to monitor the status of the data links and to change the parameters of data transfer according to the real situation in the data links.

The network link parameters are planned to be set with the OpenFlow[26] [14] protocol in the source network switch (hardware or software). The special tool PerfSonar [15] will be used to monitor the data link status.

Evidently, a specially customized test installation is required to test the data transfer process with the mentioned utilities and instrumentation. The customized testbed has to be able to simulate at least the main network problems, e.g., changing RTT, delays, percentage of lost packages, and so on. The development of such a testbed has been started at the Laboratory of Network Technologies in Distributed Computing Systems[27] of the ITMO University. This field of work attracts many researchers [16].

The testbed is intended to be a platform for comparing different utilities in the same environment. As a first step, it is planned to perform comparative measurements of a set of data transfer utilities, recording all the measurement details. In the future, this will permit one to compare in the testbed other data transfer methods in exactly the same environment.

5. THE TESTBED PROGRESS

The testbed consists of two servers HP DL380p Gen8 E5-2609, Intel(R) Xeon(R) CPU E5-2640 @2.50GHz, 64GB under Scientific Linux 6.5. Since everything is planned to be tested in the virtual environment, for each mentioned data transfer system

[22] https://cmsweb.cern.ch/phedex
[23] https://twiki.cern.ch/twiki/bin/view/CMSPublic/ /PhedexAdminDocsInstallation
[24] http://hep-t3.physics.umd.edu/HowToForAdmins/ /phedex.html
[25] http://rucio.cern.ch/
[26] https://www.opennetworking.org/images/stories/ /downloads/sdn-resources/white-papers/ /wp-sdn-newnorm.pdf
[27] http://sdn.ifmo.ru/

two virtual machines will be used: first VM as a transmitter, and another one as a receiver. In other words, we have around ten VMs. To organize those VMs, the OpenStack platform (`www.openstack.org`) was deployed and put in use. PerfSonar has been deployed as well.

To study different types of data, a special procedure has been developed to generate a test directory with files of random length; the total volume of the test directory is defined by a parameter of the procedure. During the generation of the test data, it is possible to set the mean value for the file size and the dispersion of the file size. The data inside each file in the test directory is intentionally prepared so as to eliminate the possible effect of data compression (if any) during data transfer.

At the initial stage, it is planned to compare all the above data transfer systems in a local area network to be sure that everything (all scripts) is functioning properly. A separate problem is to write all the logs, parameters, etc. during the measurement. In particular, there is the requirement to write automatically the whole `/proc` directory into some place, let's say "log directory." Also, it is required to write all the parameters and messages from the data transfer engine/utility. Finally, the data link status is intended to be written as well. All the mentioned information has to be saved in the "log directory." Obviously, everything has to be performed by the scripts dedicated to conducting the measurements.

The developed procedures (scripts) and short descriptions can be found at `https://github.com/itmo-infocom/BigData`.

6. CONCLUSION

When planning a project dealing with a large amount of experimental data, it is important to take into account the efforts used to move data over a network. It is possible to list several points in the observation cycle where Big Data transfer over a network is in real demand:

- data gathering;
- quick data quality checking (and/or filtering);
  - possible data transfer (may be local or remote);
- storing the data in secondary storage;
  - possible data transfer to remote computing facilities (may be in several locations) for further analysis;
- data analysis.

This paper describes just the data transfer technique, which is an unavoidable part of the observation cycle. In the coming experiments, where a huge volume of data is expected, it is clear that the more effective the data transfer, the more productive the scientific analysis.

ACKNOWLEDGMENTS

This work is supported by the St. Petersburg National Research University of Information Technologies, Mechanics, and Optics.

REFERENCES

1. J. Pearlstein, `http://www.wired.com/magazine/2013/04/bigdata/`
2. L. Borovick and R. L. Villars, `http://unleashingit.com/docs/B13/Cisco%20UCS/critical_big_data_applications.pdf`
3. W. E. Johnston, E. Dart, M. Ernst, and B. Tierney `https://tnc2013.terena.org/getfile/402/`; `https://tnc2013.terena.org/getfile/716/`
4. B. Tierney, E. Kissel, M. Swany, and E. Pouyoul `http://www.es.net/assets/pubs_presos/eScience-networks.pdf`
5. M. Juric, J. Kantor, T. S. Axelrod, et al., American Astron. Soc. Meeting Abstracts, No. 221, 247.01 (2013).
6. P. Dewdney, W. Turner, R. Millenaar, et al., SKA1 System Baseline Design, SKA-TEL-SKO-DD-001.
7. B. S. Acharya, M. Actis, T. Aghajani, et al., Astroparticle Phys. **43**, 3 (2013).
8. P. de Teodoro, A. Hutton, B. Frezouls, et al., in *Astrostatistics and Data Mining*, Ed. by L. M. Sarro, L. Eyer, W. O'Mullane, and J. De Ridder, Springer Ser. in Astrostatistic **2**, 107 (2012).
9. E. A. Isaev, V. V. Kornilov, P. A. Tarasov, et al., Preprint No. 8 (Lebedev Physical Institute, Moscow, 2014).
10. S. Karpov, G.Beskin, S.Bondar, et al., Acta Polytechnica **53**, 38 (2013).
11. S. W. Hodson, S. W. Poole, T. M. Ruwart, and B. W. Settlemyer, `http://info.ornl.gov/sites/publications/files/Pub28508.pdf`
12. The CMS Collaboration, J. Instrumentation **3**, S08004 (2008).
13. R. Kaselis, S. Piperov, N. Magini, et al., J. Phys. Conf. Ser. **396**, 042033 (2012).
14. B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, et al., Communications Surveys and Tutorials, IEEE **16**, 1617
15. J. Zurawski, S. Balasubramanian, A. Brown, et al., `http://www.es.net/assets/pubs_presos/20130910-IEEE-BigData-perfSONAR2.pdf`
16. D. Gunter, R. Kettimuthu, E. Kissel, M. Swany, et al., in *Proc. Meeting on High Performance Computing, Networking Storage and Analysis (2012 SC Companion), Salt Lake City, USA, 2012*, p. 1600.