

The initial stage of Big Data transfer testbed over parallel data links with SDN approach

S. Khoruzhnikov,¹ V. Grudinin,¹ O. Sadov,¹ A. Shevel,^{1,2} V. Titov,^{1,3} and A. Kairkanov¹

¹*ITMO University St.Petersburg, Russia*

²*Petersburg Nuclear Physics Institute, Russia*

³*Saint-Petersburg State University, Russia*

The transfer of Big Data over computer network is important and unavoidable operation in the past, now and in any feasible future. A large variety of astronomical projects produces the BigData. There are a number of methods to transfer the data over 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 point in Internet in general over long range distance: many thousands Kilometers. Several free of charge systems to transfer the Big Data are analyzed here. The most important architecture features are emphasized and discussed idea to add SDN Openflow protocol technique for fine grain tuning the data transfer process over several parallel data links.

email: tit@astro.spbu.ru

1. INTRODUCTION

The “Big Data”¹ is known problem for many years. In each period the term “Big Data” does mean different volume and character of the data. Keeping in mind “triple V”: Velocity, Volume, Variety we can pay attention that all those features are relative to current state of the technology. For example in 1980-s the volume of 1 TB was considered as 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 that time Big Data. In 70s80s, the development of FITS-format standardizes storage, transmission and processing of scientific and other images gave the opportunity to exchange the data between different institutes. Nowadays the systems to proceed with Big Data in modern vision are developed and used. The standards for access to big amount of astronomical data (metadata standards, formats, query languages etc), technologies for proceeding Big Data, all this are developed and supported by IVOA (International Virtual Observatory Alliance). The

¹ http://en.wikipedia.org/wiki/Big_data

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 are ESO/VLT, NOAO/CTIO, NASA/Kepler, NASA/HMS etc.

There are a range of aspects of the problem: store, analyze, transfer, etc. In this paper we discuss one of important aspects of the Big Data — the transfer over global computer network.

2. THE SOURCES OF THE BIG DATA

It is known the long list of human activities (scientific and business) which are the generators of large volume of data [1–3], see the projects: SKA², LSST³, FAIR⁴, ITER⁵, and also the sites CERN⁶, and CLDS⁷.

In according [1] total volume of business mails in the World in year 2012 is around 3000 PB ($3 \cdot 10^{18}$). The consensus estimation for the total volume of stored data is growing 1.5–2.0 times each year starting from 2000. In this paper (and for our testings) we will assume that volume of data around 100 TB (10^{14}) and more could be labeled as Big Data.

Another aspect of Big Data — the preservation of the data for long periods of time: several tens or more years. Many aspects of our personal, society, technical, and business life are now held in digital form. Large volume of those data needs to be stored and preserved. For example, results of medicine tests, data generated by important engines of various kinds (airplane engines, power station generators, etc) and other data have to be archived for 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 preserved data have to be stored in several places (continents) to avoid data loss due to technical, nature or social disasters.

Historically one of the first field where Big Data came into reality was experiments in High Energy Physics (HEP). As the result a number of aspects for data transfer were analyzed and a

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/>

range of problems were 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]:

- Hypparcos: 1989–1992, total volume of data 300 GB;
- ESO/VLT: 1999, total volume of data is 65 TB and increases by 15 TB/year;
- NASA⁸/KEPLER⁹: since 2009, 100 GB/month;
- LOFAR: LOw Frequency ARray¹⁰, 2012, till 1 PB/day;
- GAIA: Global Astrometric Interferometer for Astrophysics, 1 PB/year;
- PRAO (Pushchino): all projects, 10-100 GB/day;
- Radioastron: 1.28 TB/day;
- CERN, all projects, several tens of PB/year;
- LSST: Large Synoptic Survey Telescope, 2020, data volume 10 PB/year;
- ITER: International Thermonuclear Experimental Reactor, 2020, 1 PB/day;
- CTA: Cherenkov Telescope Array, 2015–2020, 20 PB/year;
- SKA: Square Kilometer Array, 20192024, 1500 PB/year.

Four of these projects are under design, the longer lead time the bigger data are planned to gather. Last December GAIA mission start the observations. Total volume of data is planned to be equal to 1 PB/year.

Often the observers have to store only selected fraction of the received data [10]. For deep analysis quite often there is need to distributed the obtained data among the collaborators around the World. That means good fraction of experimental data need to be transferred over Internet.

⁸ <http://www.nas.nasa.gov/>

⁹ http://www.nasa.gov/mission_pages/kepler/

¹⁰ <http://www.lofar.org/>

3. FREELY AVAILABLE UTILITIES/TOOLS FOR DATA TRANSFER OVER THE NETWORK

The time to transfer over global computer network (Internet) depends on the real data link bandwidth and volume of the data. Taking into account that we talk about volume 100TB and more we can estimate minimum required time for data copy over the network link with 1 Gbit capacity. It will give us about 100MB/sec, hence $100\text{TB}/100\text{MB}/\text{sec} = 1\,000\,000\text{ secs} = 277.8\text{ hours} = 11.6\text{ days}$. During this time the parameters of the network link might be changed. For example percent of dropped network packages can be varied significantly. The data link might suffered of operation interruptions for different period: 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 1/2 thousand parameters describing the network link in the kernel. Not all of them are equally sensitive or influencing on the data transfer process. Most important of them are TCP Window size, MTU, congestion control algorithm, etc. Of course quite important the number of independent network links which could be used in parallel. Also there are important network parameters like round trip delay time (RTT) and % of lost network packages. Quite obvious that in each data transfer of large volume (running long time) we need to be able to tune (to set) different number of threads, different size of TCP Window, etc during the data transfer process to achieve maximum data transfer speed.

Now it is time to observe freely available data transfer tools/utilities which might be used to transfer Big Data over the network and what they permit to tune.

A. Ideas to Compare the data transfer utilities

First of all quick consideration of parameters to compare the data transfer utilities which might help to transfer Big Data.

- Multi-stream data transfer mode is ability to use several TCP streams in parallel.
- Multi-link data transfer mode is ability to use more than one data link in parallel; important feature especially if it is possible to take into account that available network links are not equal in bandwidth and in conditions (reliability, price, real status, etc).
- Possibility to set parameters low level parameters e.g. TCP Window size, etc.

- The method to bypass the 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?

In reality the data transfer consists of many steps: read the data from the storage, transfer the data over network, write the received data to the storage on remote computer system. In this paper our attention is concentrated more on network transfer process.

B. Low level data transfer utilities/tools

We could mention several utilities for the data transfer over the network (at least part of them are known for many years):

- one of low level protocols to transfer the data over the network is UDT¹¹. UDT is library which implements data transfer protocol which permit to use *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 protocol RDMA over Converged Ethernet (RoCE) [4] has been studied and it was found that in many cases RoCE shows better results than UDP, UDT, conventional TCP.
- MP TCP¹² is interesting protocol which permits to use several data links in parallel for one data transfer. The protocol is implemented as Linux kernel driver.
- (open) ssh family¹³ — well known data transfer utilities deliver strong authentication and a number of data encryption algorithms. Data compression before encryption to reduce the data volume to be transferred is possible as well. There are two well known (open) SSH flavors: patched SSH version¹⁴ which can use increased size of buffers and SSH with Globus GSI authentication. No real restart after failure. No parallel data transfer streams.
- bbcp¹⁵ — utility for balk data transfer. It is assumed that bbcp is running on both sides, i.e. transmitter, as client, and receiver as server.

¹¹ <http://udt.sourceforge.net/>

¹² <http://mptcp.info.ucl.ac.be/>,
<http://multipath-tcp.org/>

¹³ <http://www.openssh.org/>

¹⁴ <http://sourceforge.net/projects/hpnssh/>

¹⁵ <http://www.slac.stanford.edu/~abh/bbcp/>

- `bbftp`¹⁶ utility for bulk data transfer. It implements its own transfer protocol, which is optimized for large files (larger than 2 GB) and secure as it does not read the password in a file and encrypts the connection information.
- `Xdd` [11] — utility developed to optimize data transfer and I/O processes for storage systems.
- `fdp`¹⁷ — Java utility for multi-stream data transfer.
- `gridFTP`¹⁸ is advanced data transfer utility for globus security infrastructure (GSI).

Many of them are quite effective for data transfer from point of view of link capacity usage. However Big Data transfer assumes significant transmission time (may be many hours, days or more). For long time it is not easy rely on so simple transfer procedures. As we mention above the network link might change the capacity and percent of lost network packages, and so on.

C. Middle level File Transfer Service

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

Another interesting development is `SHIFT`²⁰ which is dedicated to do reliable data transfer in LAN and WAN. There were paid much attention to the reliability, advanced tracking, performance of the data transfer and the usage of parallel data transfer between so called equivalent hosts (between computer clusters).

D. High level data management service: *PhEDEx*

`PhEDEx`²¹ — Physics Experiment Data Export is used (and developed) in collaboration around Compact Muon Solenoid (CMS) [12, 13] experiment at CERN. The experiment does produce a lot

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;
<https://svnweb.cern.ch/trac/fts3>

²⁰ <http://fasterdata.es.net/data-transfer-tools/>

²¹ <https://cmsweb.cern.ch/phedex>,
<https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhedexAdminDocsInstallation> and <http://hep-t3.physics.umd.edu/HowToForAdmins/phedex.html>

of experimental data (in 2013 it was written around 130 PB). Data analysis requires to copy of the data in a range of large computing clusters (about 10 locations in different countries and continents) for analysis and data archiving. Later on the fractions of the data might be copied to smaller computing facilities (more than 60 locations). Total data transfer per day is achieved 350 TB/day[13]. It is possible that in nearest future the volume per day will be increased. Because in between several sites there are more than one link in PhEDEx there were developed routing technique which permit to try alternative route when default route is not available.

Finally the system PhEDEx is quite complicated and the management service depends on the physics experiment collaboration environment. It is unlikely that PhEDEx is possible to use without redesign in different environment.

4. CONSIDERATION

Mentioned utilities have several common features. Among them:

- all utilities have 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, use a range 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;
- usage of a number of techniques to make data transfer more reliable;
- the utilities are not equal in number of parameters and scope of suggested tasks. Part 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 collaboration ATLAS²², are dedicated to be used as part of more complicated and specific computing environment.

²² <http://rucio.cern.ch/>

In other words there is stack of toolkit which might help in many cases to transfer the Big Data over networks. At the same time it is seen that quite a few utilities can use more than one network link.

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

Network link parameters are planned to be set with protocol Openflow²³ [14] in the source network switch (hardware or software). The special tool PerfSonar [15] will be used to watch the data links status.

It is obvious the specially customized test installation is required to test data transfer process with mentioned utilities and described improvements. The customized testbed has to be able to simulate at least main network problems, e.g. changing RTT, delays, drop package percent, and so on. Such the testbed development has been started at the Laboratory of the Network Technologies and Distributed Computing Systems²⁴. The direction of the study is attractive for many researchers [16].

The testbed is intended to be platform to compare different utilities in the same environment. As the first step it is planned to perform comparative measurements with the range of data transfer utilities with writing all the measurement conditions details. That permits to compare in future other data transfer methods in exactly same environment in the testbed.

5. THE TESTBED PROGRESS

The testbed consists of two servers HP DL380p Gen8 E5-2609, Intel(R) Xeon(R) CPU E5-2640 @2.50GHz, 64 GB under Scientific Linux 6.5. Because it is planned to test everything in virtual environment for each mentioned data transfer systems two virtual machines will be used. One VM as transmitter and another VM as receiver. In other words we have around ten VMs. To organize those VMs is was deployed and started to use Openstack platform from OpenStack.org. PerfSonar has been deployed as well.

²³ <https://www.opennetworking.org/images/stories/downloads/white-papers/wp-sdn-newnorm.pdf>

²⁴ <http://sdn.ifmo.ru/>

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

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

Developed procedures (scripts) and short descriptions are written in the site <https://github.com/itmo-infocom/BigData>.

6. CONCLUSION

Planning the project dealing with large amount of experimental data is important to take into account the efforts for data moving over network. It is possible to show several points in observation cycle where Big Data transfer over network is real demand:

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

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

ACKNOWLEDGEMENT

The work is supported by the St. Petersburg National Research University of Information Technologies, Mechanics and Optics (www.ifmo.ru).

REFERENCES

-
1. J. Pearlstein, Information Revolution: Big Data Has Arrived at an Almost Unimaginable Scale // <http://www.wired.com/magazine/2013/04/bigdata/>.
 2. L. Borovick, R.L. Villars. White paper. The critical Role of the Network in Big Data Applications // http://unleashingit.com/docs/B13/Cisco%20UCS/critical_big_data_applications.pdf
 3. W.E. Johnston, E. Dart, M. Ernst, B. Tierney // Enabling high throughput in widely distributed data and analysis systems: Lessons from the LHC // <https://tnc2013.terena.org/getfile/402>, <https://tnc2013.terena.org/getfile/716>
 4. B. Tierney, E. Kissel, M. Swany, E. Pouyoul // Efficient Data Transfer Protocol for Big Data, http://www.es.net/assets/pubs_presos/eScience-networks.pdf // Lawrence Berkley National Laboratory // School of Informatics and Computing, Indiana University.
 5. M. Juric, J. Kantor, T.S. Axelrod, et al. 2013, American Astronomical Society Meeting Abstracts No 211, 221, N247.01.
 6. P. Dewdney, SKA1 System Baseline Design, 12.03.2013 07:09, SKA-TEL-SKO-DD-001. Revision: 1.
 7. B.S. Acharya, et al. 2013, Astroparticle Physics, v. 43, pp. 3–18, <http://dx.doi.org/10.1016/j.astropartphys.2013.01.007>.
 8. P. De Teodoro et al. Data Management at GAIA Data Processing Centers // Astrostatistics and Data Mining, ed. by L.M.Sarro et al., Springer Series in Astrostatistic 2, DOI 10.1007/978-1-4614-3323-1_10, 2012
 9. E.A. Isaev, V.V. Kornilov, P.A. Tarasov, V.A. Samodurov, M.V. Shatskaya. Preprint PIAN, No. 8, 2014 (in Russian).
 10. S. Karpov, et al. Acta Polytechnica, **53**, 1, 3843 (2013).
 11. S.W. Hodson, S.W. Poole, T.M. Ruwart, B.W. Settlemyer // Moving Large Data Sets Over High-Performance Long Distance Networks // Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008 Oak Ridge, 37831-6164 // <http://info.ornl.gov/sites/publications/files/Pub28508.pdf>.
 12. The CMS Collaboration 2008. The CMS experimental at the CERN LHC JINST 3 S08004.
 13. R. Kaselis, S. Piperov, N. Magini, J. Flix, O. Gutsche, P. Kreuzer, M. Yang, S. Liu, N. Ratnikova, A. Sartirana, D. Bonacorsi, J. Letts // CMS Data Transfer operations after the first years of LHC colli-

- sions // International Conference on Computing in High Energy and Nuclear Physics 2012 (CHEP2012)
IOP Publishing Journal of Physics: Conference Series 396 (2012) 042033
14. B.A.A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obracht, and T. Turletti // A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks, http://hal.inria.fr/index.php?halsid=ig58511e1q1ekqq75uud43dn66&view_this_doc=hal-00825087&version=5.
 15. J. Zurawski, S. Balasubramanian, A. Brown, E. Kissel, A. Lake, M. Swamy, B. Tierney, and M. Zekauskas // perfSONAR: On-board Diagnostics for Big Data // http://www.es.net/assets/pubs_presos/20130910-IEEE-BigData-perfSONAR2.pdf
 16. D. Gunter, et al // Exploiting Network Parallelism for Improving Data Transfer Performance // <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6496123> // High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion, ISBN 978-1-4673-6218-4, DOI: 10.1109/SC.Companion.2012.337.