Simone Bernhart et al.: E-transfer at the Bonn Correlator, IVS 2012 General Meeting Proceedings, p.115–119
http://ivscc.gsfc.nasa.gov/publications/gm2012/bernhart.pdf

E-transfer at the Bonn Correlator

Simone Bernhart, Arno Mueskens

Institute of Geodesy and Geoinformation, University of Bonn
Contact author: Simone Bernhart, e-mail: simone@mpifr-bonn.mpg.de

Abstract

During recent years, the number of stations that transfer their observational data via high-speed network connections to the correlators has increased significantly. In order to help coordinating e-transfers among correlators and stations, the Geodesy VLBI Group has set up a website that shows ongoing transfers. We present the usage of this website as well as the overall status of e-transfers at the Bonn correlator.

1. Technical Specifications of the Bonn Correlator

The Distributed FX correlator[1] consists of 60 nodes with 8 compute cores on each node (480 cores in total). The correlator cluster is connected via 2 × 1 Gb Ethernet to 14 Mark 5 units used for playing back the data. All Mark 5 can play back all types of Mark 5 data (A/B/C). If more than 14 playback units are required, and in the case of e-VLBI, data are copied to the raid systems prior to correlation. Further technical specifications can be found in the Bonn Astro/Geo Correlator report [1].

A 1-Gbps switch connects the Bonn correlator to the high speed network (German Research Network DFN and GÉANT, the pan-European data network dedicated to the research and education community). Furthermore, a 10-Gbps dedicated fiber connection exists for Effelsberg and the LOFAR station located there, which, however, is not directly connected to MPIfR. Since late 2011, a firewall computer has been set up for the e-transfer servers.

2. E-transfer Status

For the transfers we use Tsunami, which is a fast file transfer protocol that uses UDP (User Datagram Protocol) data and TCP (Transmission Control Protocol) control for transfers over high speed networks (≥ 1 Gbps) for a long distance. The current version is 1.1 cvsbuild42 and can be downloaded at `http://tsunami-udp.sourceforge.net/`. The project is based on original Indiana University 2002 Tsunami source code but has been significantly improved and extended by Jan Wagner. As such, large portions of the program today are courtesy of the Metshovi Radio Observatory.

TCP is the most commonly used protocol on the Internet. The biggest advantage of TCP is the so-called “flow control”, which guarantees the delivery of the data that are transferred. Flow control determines when data needs to be re-sent and interrupts the flow of data until previous packets are successfully transferred, i.e., the client re-requests the packet from the server until the whole packet is complete and identical to its original.

[1] DiFX: A Software Correlator for Very Long Baseline Interferometry using Multiprocessor Computing Environments, 2007, PASP, 119, 318

UDP is another commonly used protocol on the Internet. It offers speed and is much faster than TCP because there is no form of flow control or error correction. This main advantage is, however, at the same time its biggest disadvantage.

Tsunami combines both TCP and UDP; it offers data transmission with default priority for data integrity, but disabling retransmissions may as well enable rate priority. Communication between the client and server applications flows over a low bandwidth TCP connection. The bulk data is transferred over UDP.

Alternatives to Tsunami are, e.g., UDT[2], another UDP based transfer protocol used by colleagues from New Zealand, or VDIF-SUDP which is used by the Japanese colleagues [2], and others.

2.1. Data Storage

At the Bonn correlator, we currently have three machines available with a shared 1 Gbps connectivity. Two of the servers (io03 and io10) are connected to the DiFX correlator cluster via InfiniBand. The total data storage capacity is of the order of ∼70 TB distributed as follows:

- sneezy2: 7.6 TB (`/mnt/rawdata`)
- io03: 20 TB (`/data3`)
- io10: 37 TB (`/data10`) + 8.2 TB (`/data10b`)

The structure of the data storage path is the same for all servers following the scheme

`/parent_folder/experiment_type/station_name/exp_name/`
(e.g., `/data10/ohig/fortaleza/ohig76`).

2.2. *Regular* e-transfers to/from Bonn

Table 1 shows a list of stations and correlators that already have transferred data to or from the Bonn correlator. Names in italics indicate transfers on a regular base, and stations marked with asterisks have their data transferred directly from the Mark 5 module which is mounted via fuseMk5.

Taking into consideration that most of the stations meanwhile transfer the data on their own, the tabulated transfer rates are merely experienced values from the time that the (partially just test) transfers had been started from Bonn, and only one or two were running at the same time. Now that sometimes three or more transfers are running in parallel, the transfer rates per station are usually of the order of ∼ 300 Mbps.

On average $\geq 50\,\%$ of the stations do e-transfer, and the number increases. E.g., in T2081, 21 stations participated in the observations, and 12 of them sent their data via Internet connection. The average amount of e-transferred data per week ranges from 4 to 6 TB considering only the regular INT3 and R1 experiments.

3. Website for Active Transfers

During recent years, the number of stations, that transfer their observational data via high-speed network connections to the correlators, has increased significantly. This necessitates some

[2] http://udt.sourceforge.net/

Table 1. *Regular* e-transfers to/from Bonn.

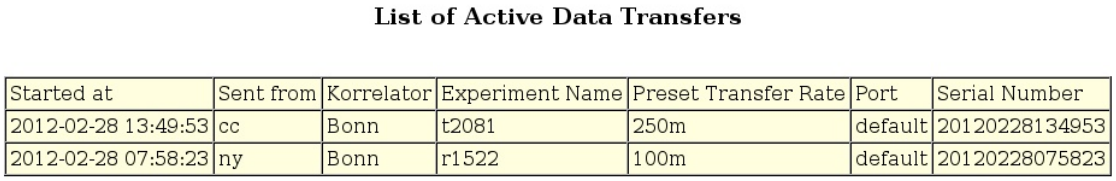
Station	data rate [Mbps]	Experiments
Onsala	500	GEO (R1, EURO, T2, dBBC test data)
Metsähovi	800	GEO (EURO, T2)
Medicina	?	GEO (EURO, T2, R1)
Ny-Ålesund⋆	100	GEO (INT3, R1, EURO, T2)
Wettzell	250	GEO (INT3, dBBC test data)
Yebes(⋆)	800	GEO (R1, EURO, T2)
JIVE	400	ASTRO
Hartebeesthoek⋆	400	GEO (R1, T2, OHIG)
Tsukuba (Aira, Chichijima, VERA-Ishigakijima)	600	GEO (INT3, R1, T2)
Kashima (K1, Kb, Syowa)	600	GEO (R1, T2, OHIG)
Mitaka (VERA-Mizusawa)	400	GEO (T2)
Seshan⋆	250	GEO (INT3)
Hobart⋆ *(Hb, Ho)*	300	GEO (R1)
Warkworth⋆	?	GEO (R1, T2 OHIG)
CSIRO (ATCA, Ceduna, Mopra, Parkes)	500	Astrometry (GEO planned)
Fortaleza⋆	400	GEO (R1, T2, OHIG)
WACO	250	GEO (Hb data)

⋆ via fuseMk5

form of coordination since the transfers are mostly running on the same network connections and thus interfere mainly due to bandwidth limitations. In order to help coordinating e-transfers among correlators and stations, the Geodesy VLBI Group has set up a small set of scripts to show ongoing transfers on a website (`http://www.mpifr-bonn.mpg.de/cgi-bin/showtransfers.cgi`, see also Figure 1). It is important to point out that the website merely shows active transfers and works on a first come first served basis. An overall coordination of e-transfers concerning their importance and priority is still required, and the transfer website should be regarded as a temporary solution.

The aforementioned website is the front end to display information about current transfers and is located on the MPIfR Web server. It is created by a Perl script running as CGI which reads the underlying database. The HTML page is static; there is no mechanism to automatically update the table. Therefore the page needs to be reloaded in order to see the latest status of transfers.

For others, the most important information in the table displayed in Figure 1 is the route on which the data are sent ("Sent from" and "Correlator") as well as the applied transfer rate and the port on which the transfer is running. The following text will shortly describe how an ongoing transfer can be shown on the website and be removed again as soon as it is finished.



List of Active Data Transfers

Started at	Sent from	Korrelator	Experiment Name	Preset Transfer Rate	Port	Serial Number
2012-02-28 13:49:53	cc	Bonn	t2081	250m	default	20120228134953
2012-02-28 07:58:23	ny	Bonn	r1522	100m	default	20120228075823

Figure 1. View of the transfer website at `http://www.mpifr-bonn.mpg.de/cgi-bin/showtransfers.cgi`.

3.1. Start and Stop Files

At the start of a transfer it is necessary to create an (empty) start file which needs to be sent to the MPIfR FTP server (`ftp.mpifr-bonn.mpg.de`) to directory `incoming/geodesy/transfers`. One can, e.g., use the program *ncftpput*:

```
ncftpput ftp.mpifr-bonn.mpg.de /incoming/geodesy/transfers file_start
```

This will send the file via anonymous ftp. As soon as the transfer is finished or aborted(!), it is important to send the corresponding stop file to our FTP server. As soon as the script sees a pair of start and stop files it will delete both of them and remove the corresponding information from the database. In consequence the transfer will disappear from the Web page. The delay time depends on the current configuration of the cron job that calls the script to generate the html page but typically will be of the order of ten seconds to one minute. The name of the start file has to match the following scheme:

```
[sn]_[exp name]_[sent from]_[correlator]_[preset transfer rate]_[tsunami port]_start
```

Table 2. File name information.

`sn`	serial number - time stamp, format: YYYYMMDDhhmmss
`exp name`	Name of the experiment of which the data is transferred
`sent from`	(Two-letter) station code of the recording station
`correlator`	Name of the correlator the data is sent to
`preset transfer rate`	The applied transfer rate
`tsunami port`	Port used for the transfer (default=46224)

The words and the square brackets have to be replaced by the appropriate values which are described in Table 2, e.g.,

```
20120228075823_ny_Bonn_r1522_100m_default_start.
```

The "serial number" serves as a time stamp of the transfer start. It is used both for the time information displayed on the website and, together with the "sent from", as an identifier of the transfer itself. This identifier is also used in the stop file which needs to be named as follows:

```
[serial number]_[sent from]_stop
```

In the aforementioned example this corresponds to

```
20120228075823_ny_stop.
```

In the database there is also an extra primary key independent of the “serial number”. Considering it to be very unlikely that a single station sends two experiments starting the same second, the combination of “serial number” and “sent from” should be sufficient to identify the transfer in practice.

4. Outlook

In the near future, additional (test) transfers will be performed with new stations such as the Japanese stations Koganei and Uchinoura, which took part in the observations of T2081.

Concerning our connectivity, the 1-Gbps network connection is sufficient for the current maximum observing mode of 256 Mbits of experiments that are handled at the Bonn correlator and the number of e-transfer stations per experiment that we are dealing with at the moment. But as soon as the observing mode is upgraded to 512 Mbits and even more stations start doing e-transfer (let alone when the astronomical EVN stations use e-transfer instead of module shipping to the Bonn correlator), no guarantee can be made to meet the 15-day turn-around time that is envisaged for R1 experiments.

In view of VLBI2010, it is still planned to upgrade the network connection to 2 Gbps preferably even 10 Gbps. However, funding problems still tend to be insurmountable.

References

[1] La Porta, L., Alef, W., Bernhart, S., Bertarini, A., Mueskens, A., Rottmann, H., Roy, A. The Bonn Astro/Geo Correlator, In: International VLBI Service for Geodesy and Astrometry 2011 Annual Report, NASA/TP-2012-217505, K. D. Baver and D. Behrend (eds.), June 2012.

[2] Sekido, M., Takefuji, K., Kimura, M., Hobiger, T., Kokado, K., Nozawa, K., Kurihara, S., Shinno, T., Takahashi, F. In: IVS 2010 General Meeting Proceedings “VLBI2010: From Vision to Reality”, NASA/CP-2010-215864, D. Behrend and K. D. Baver (eds.), December 2010.