IVS Memorandum 2013-001v01

23 July 2013

“e-transfer Status at IVS Correlators”

Simone Bernhart, Jason SooHoo

e-transfer Status at IVS Correlators

About e-VLBI

Electronic VLBI (e-VLBI) describes the transfer of observational data from radio telescopes to the correlators via high-speed network connections like fiber channel. Advantages of this technology are

- higher data rates for greater sensitivity with the same antennas
- faster turn-around time of experimental processing for use in scientific investigation
- elimination of large and expensive media pool, that is currently being used.

e-transfers at the Bonn correlator

I remember that when I started working with the geodesists in 2007, the first e-transfer tests had been performed at the Bonn correlator. One of our student assistants - Christian Dulfer, a student of computer sciences - had set up some servers and installed software to do the first e-transfers between Bonn and some IVS stations. Among these stations there were Onsala and Metsaehovi, but also Ny-Alesund and Tsukuba. Since autumn 2007, we do regular e-transfers for the INT3 observations which take place every Monday.

Back then Bonn was still working with the hardware correlator and the transferred data had to be copied onto modules before correlation. Besides the Japanese K5 data needed to be converted to Mark5 format before writing them to disk. Meanwhile the Japanese colleagues are so kind as to convert the data before the transfer and almost all stations do transfers to the Bonn correlator on their own. Besides, with the dawning of the DiFX correlator era there is no need anymore to write the data onto modules because they can be correlated straight away from RAID.

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). Since late 2011, a firewall computer has been set up for the e-transfer servers.

At the Bonn correlator, we currently have five machines available with a shared 1 Gbps connectivity. Three of the servers (io03, io10 and io11) are connected to the DiFX correlator cluster via InfiniBand. The total data storage capacity is of the order of ~130 TB distributed as follows:

- io03: 20 TB (/data3)
- io10: 37 TB (/data10) + 8.2 TB (/data10b)
- io11: 38 TB (/data11)
- sneezy1: 19 TB (/sneezy1)
- sneezy2: 7.6 TB (/sneezy2)

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) on 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 Metsähovi 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.

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 (http://udt.sourceforge.net/), another UDP based transfer protocol used by colleagues from New Zealand, or VDIF-SUDP which is used by the Japanese colleagues (http://www2.nict.go.jp/aeri/sts/stmg/K5/Software/VDIF_SUDP/VDIF-SUDP-j.html), and others.

If a station plans to start data transfer to the correlators via internet, it should have a network connection of at least 100 Mbps available. A 'quick-and-dirty' introduction to the transfer from the stations to Bonn using the Tsunami UDP Protocol and fuseMk5(A) in case data need to be transferred from a Mark5 unit can be found here: http://www3.mpifr-bonn.mpg.de/div/vlbicor/geodesy/Docs/etransfer_how-to.txt. The instructions mainly refer to the use of Tsunami scripts that have been developed a couple of years ago in Bonn (for those who are interested, these scripts can be downloaded here: http://www3.mpifr-bonn.mpg.de/div/vlbicor/geodesy/Docs/tsunami-scripts.tar). But at the end of the etransfer_how-to.txt file there is a description on how to run Tsunami without these scripts. I would like to point out that the use of Tsunami and/or the scripts is not mandatory. Stations, that do the transfer on their own, are free to use whatever high speed transfer protocol they prefer.

For a transfer to Bonn the following information is important: login to our servers is only possible via a *passphrase-generated public key* that will be added to the authorized_keys files on our servers. If you plan to do transfers to Bonn, please contact geodesy(at)mpifr-bonn.mpg.de using subject [e-transfer] and send the following information to us:

- IP address(es) of the machine(s) from which you want to send the data
- public key of the user of this machine
- information on available bandwidth of your internet connection

Besides please grant access to your machine(s) for the IP addresses listed in our etransfer_how-to.txt (see link above) and please attach our public key to your authorized_keys file in case you would like us to do some connectivity tests.

e-transfers at the Haystack correlator

At Haystack Observatory we began e-transfers back in the early 2000's. It was mainly for transferring Kashima K5 data which we converted from K5 to M5, wrote the data to Mark5 disk modules, and shipped out to correlator sites. Our e-transfer operations were limited by

our shared 100Mb/s network and it was more practical to ship the disk modules.

In 2011, The Haystack network was upgraded from our 100Mb/s shared connection to a 10Gb/s dedicated connection. Our network runs through MIT's backbone and onto the NOX (Northern Crossroads). The Northern Crossroads supports high speed networking for research institutions in New England allowing access onto the Internet2 network.

We currently maintain 2 data servers for e-transfers. They have a combined capacity of ~48TB of space and are configured as RAID0 to maximize performance though we take the risk of data redundancy.

- evlbi1: ~16TB (/raid0, /raid1, /raid2, /raid3)
- evlbi2: ~32TB (/raid4, /raid5, /raid6, /raid7)

For additional details of e-transfers at Haystack please visit our website:
http://evlbi.haystack.mit.edu

e-transfers at the Washington correlator

...

Web page on 'List of Active Transfers'

During the last years, the number of stations, that transfer their observational data via high-speed network connections to the correlators, has increased significantly. This necessitates some 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 cronjobs to show ongoing transfers on a web page (http://www3.mpifr-bonn.mpg.de/cgi-bin/showtransfers.cgi), see also Fig. 1. In addition to ongoing transfers, the storage capacity at the three IVS correlators in Washington, Haystack and Bonn is listed as well. It is important to point out that the website merely shows active transfers and works on a first come-first served base. An overall coordination of e-transfers concerning their importance and priority is still required and the transfer web page should be regarded as a temporary solution.

List of Active Transfers for VLBI

List of Active Data Transfers

Started at	Sent from	Sent to	Raid	Experiment Name	Preset Transfer Rate	Port	Serial Number
2013-06-28 06:11:49	ny	Haystack	raid1	rd1304	100m	default	20130628061149
2013-06-27 10:59:53	Hb175	WACO	data5	aust09	300m	52100	20130627105953
2013-06-27 09:30:00	on	Bonn	data11	t2090	600m	default	20130627093000

Bonn Storage Information

Raid	Via Server	Size	Free	Note
/data3	io03	19.1 TB	1.6 TB	
/data10	io10	36.4 TB	7.5 TB	
/data11	io11	81.9 TB	11.6 TB	38 TB for e-transfer!!!
/sneezy1	sneezy1	18.2 TB	7.5 TB	
/sneezy2	sneezy2	7.5 TB	7.3 TB	

Figure 1: Screenshot of 'List of Active Transfers' web page

The affore mentioned 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.

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

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 according information from the database. In consequence the transfer will disappear from the web page. The delay time depends on the current configuration of the cronjob 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]_[sent to]_[preset transfer rate]_[port]_[raid]_start

The words and the square brackets have to be replaced by the appropriate values which are

described in Table 1, e.g.,

20120228075823_r1522_Bonn_ny_100m_default_start

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
sent to	Name of the correlator the data is sent to
preset transfer rate	The preset transfer rate of the transfer
port	The port (default=46224)
raid	The storage raid (e.g., data10 or data3)

Table 1: content of start message filename

The "serial number" serves as a time stamp of the transfer start. It is both used 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 affore mentioned 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 pratice.

Outlook

Bonn

Certainly in the near future, additional (test) transfers will be performed with more stations. Concerning the Bonn 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), it cannot be guaranteed to meet the 15-days turn-around time that is envisaged for R1 experiments.

In view of the above limitations and for VGOS (VLBI2010 Geodetic Observing System), it is definitely necessary to upgrade the network connection to 10 Gbps. However, funding problems still tend to be unsurmountable.

A compact summary of the above given information as well as useful links to transfer protocols and network testing tools can be found on the web pages of the Geodesy VLBI Group: http://www3.mpifr-bonn.mpg.de/div/vlbicor/geodesy/evlbi/index.html.

Haystack

Data e-transferred are still being written onto disk modules for processing as our data servers have not been merged with the correlator. Our goal is to have this merged and correlation processing can be done directly from the e-transferred data. We also hope to increase our server capacity as data rates and size increases.

Washington

...