

Mark 5/e-VLBI Newsletter

MIT Haystack Observatory

July 2004

Issue #6 (updated 21 July 2004)

The Mark 5/e-VLBI Newsletter is issued from time to time to keep users informed regarding Mark 5/e-VLBI progress, plans, problems, solutions and workarounds. All back issues of the Newsletter are available at the Mark 5 web site at www.haystack.edu/mark5. We invite input from anyone on subjects we should discuss or questions that need answers; please send them to mark5@haystack.mit.edu. General information about the Mark 5 and e-VLBI is available at the Haystack web site at <http://www.haystack.edu/>.

Contents of this Issue

1. Disk conditioning for 1 Gbps recording (updated 21 July 2004)
2. Support for serial-ATA disks (updated 16 July 2004)
3. Disk-module labeling procedures (again!)
4. Making disk-module labels with a Dymo Labelwriter
5. e-VLBI Network 'Weather Map' Introduced
6. UT1 Result in 4½ hours!
7. Conduant return policy update

1. Disk conditioning for 1 Gbps recording (updated 21 July 2004)

In the original release of this Newsletter, we recommended a condition pass for *every* module at a station before it is recorded, particularly for experiments recording or playing back at 1 Gbps where good disk performance is most critical. It was pointed out to us that this is a near impossibility in many cases and, furthermore, should not really be necessary.

OK, we concede! After further input from Conduant and further internal discussions, we recognize that our recommendation to do a full conditioning pass on all modules before an experiment is unrealistic. Furthermore, observed disk failures during shipping, though rare, usually tend to render disks inoperable rather than slowing their performance, though we believe we may have rarely seen the latter case. We were being more conservative, I think, than is probably justified. So, herewith, we would like to like to amend our recommendations for pre-experiment module verification as follows:

1. Mount each module into a Mark 5 and verify that it comes Ready (not flashing).
2. Read the VSN's and verify that all disks respond properly:'

vsn?;

3. Record and playback a couple of minutes of test pattern data at the maximum rate for which the module is rated (128Mbps times # of disks). For a module with 8 disks, the commands are:

mode=tvgr;

play_rate=32; (set tvgr rate to 32 MHz for 1024Mbps data rate for modules
with 8 disks)

start_stats=; (clear disk statistics)

record=on; (start recording)

record=off; (stop recording)

get_stats?;	(get disk statistics)
start_stats=;	(clear disk statistics)
scan_play=on;	(playback tvg data)
play=off;	
get_stats?;	(get disk statistics)
reset=erase;	(erase module)

This procedure should take only a few minutes per module. If the statistics for any individual disk are clearly out of line, you may wish to perform further testing on that module, including perhaps a full conditioning pass if time is available. If an under-performing disk persists, it should be replaced. Though the StreamStor is designed to shift the load away from slow disks, overall module data capacity will suffer as more data is shifted to normally performing disks.

Writing to a disk is an open-loop process; there is no read-after-write. If there are undetected bad sectors on a disk, the writing process will appear to proceed normally, but unreadable sectors will be written. On playback, the disk will go to extraordinary measures to try to recover the data, re-reading each bad sector many times and, after giving up after as much as 1 or 2 seconds later, will spare the sector, return an error, and proceed. Additionally, a slow-performing disk can tie up the Master/Slave bus for long periods of time, essentially causing the output from *both* disks on the bus to be lost during that time. On playback, the StreamStor card must retrieve the disk data in a timely fashion if it is to keep up with the playback rate. At a playback rate of 1 Gbps, there is little time to lose; if a disk does not return sector data in a timely manner, a special fill pattern is substituted for the data from that sector, which is marked invalid and is not correlated. As a result, even a relatively small fraction of bad sectors on a disk can have a dramatic impact on its ability to keep up with a high playback data rates.

The conditioning process reads all disk sectors and spares any bad sectors, ensuring that playback performance will be as good as possible. Obviously, this is most critical at high playback rates.

On recording, the StreamStor insists that all disks in a newly inserted pack be present and able to at least read their individual serial numbers; *a flashing Ready light indicates that these conditions have not been met and data cannot be recorded on the module*. Once recording is started, however, the StreamStor card is designed to shift the recording load from slow or failed disks to normally operating disks to minimize data loss. Though a module with only 6 well-performing disks can keep up with 1 Gbps, operating in this mode is a bit risky since the inherent data-rate capability of the disks decreases substantially as the heads move towards the inner diameter of the disk platters.

On playback, as mentioned above, unrecoverable data from bad or slow disks are replaced by a special fill pattern. If a slow/failed disk ties up a Master/Slave bus, all data from that disk pair may be lost. At the Haystack correlator, we have observed a number of instances of slow or failed disks in a module, which have been read and correlated uneventfully except for the data loss from the affected disk. We have tested the loss of two disks in an 8-pack module by simply disconnecting power to two of the disks and correlating; in this case, ~25% of the data are lost, as one would expect. The data loss is spread evenly over all channels.

2. Support for serial-ATA disks (updated 16 July 2004)

Work is continuing to develop a satisfactory solution to support serial-ATA (SATA) disks in the Mark 5. SATA disks are rapidly becoming quite popular, and the desire is to support both the

current parallel-ATA (PATA) disk modules and SATA disk modules interchangeably in the Mark 5 system.

The biggest stumbling block has been finding a connector that is both compatible with SATA requirements (which must support multi-Gbps serial data through the connector -- a very tough requirement!) and has sufficient durability for a large number of insertion/removal cycles. There is currently no connector on the market that meets the necessary specifications. One connector under consideration is the Fujitsu FCN-260D, which is based on the Infiniband connector, and is suitable in all respects except for an insertion/removal cycle specification of only 250 cycles (this compares with a 5000-cycle specification for the current 200-pin connector on current disk modules). Though few disk modules will likely exceed the 250 cycle count over their lifetime, backplane connectors at heavily utilized stations and at correlators may exceed this count in a year or less.

The fundamental durability problem is that the gold plating on the connectors wears on each insertion/removal cycle. In tests conducted at Fujitsu, the gold plating was not worn through after 250 cycles, but was worn through after 500 cycles, hence the 250-cycle rating. When the plating is worn through, the impedance of the connection is changed and the underlying metal is exposed to potential corrosion. Conduant is currently working with several connector manufacturers to try to solve this problem, but the prospects for convincing a connector manufacturer to alter their product for such a low-volume customer are slim. Nevertheless, Conduant is conducting in-house testing of candidate connectors to better understand the consequences of plating wear-through.

An improved mechanical design for the 8-pack module for the SATA disks has already been prototyped by Conduant, of which we have a sample at Haystack and with which we are quite happy. If the connector problem can be overcome, Conduant will proceed with the design of a new chassis backplane which will accommodate both the legacy PATA modules (with the existing 200-pin connector) and the new SATA modules. If the connector problem cannot be overcome, some other approach will be required. Given the nature of the difficult connector problems, we do not expect the SATA upgrade to be available before 2005. We will keep the community informed regarding progress.

If anyone in the VLBI community is aware of connectors which may be suitable for the Mark 5 SATA application, Conduant is very open to receiving suggestions. The requirements for the SATA module connector include:

- 1.5GHz+ signal integrity
- Differential design (grounds shielding each pair and same-length leads) - 100 ohm impedance
- Enough pins to support 16 differential pairs plus grounds plus power (min 16A @ 12V)
- Minimum durability of 1000 cycles (estimated 3-year life at correlator)
- Maximum backplane height of 10mm
- Right angle (or straddle) version available for drive module PCB, straight version available for backplane PCB

Please send your suggestions to Ken Owens at Conduant (ken@conduant.com).

3. Disk-module labeling procedures (again!)

We have learned that some disk modules arriving at correlators have had confusing labels. In particular, labels have sometimes appeared on the metal disk-module covers and/or on the cardboard box that conflict with the actual module VSN. *Please do not apply labels to anywhere*

other than to the module itself. There should be one extended-VSN barcode label on the front of every module and one ‘track’ VSN barcode label on the back, as well as a Module Conditioning/Problem label on the side.

4. Making disk-module labels with a Dymo Labelwriter

We use and recommend the Dymo Labelwriter Turbo 330 printer as an inexpensive and convenient way to create barcode labels for Mark 5 disk modules. Two sets of software are available to drive this printer:

From Windows: The current application software shipping with the Dymo printer has a flaw and should not be used. Instead, retrieve an older correct version at <http://web.haystack.mit.edu/mark5/downloads.html> and use VSN label template files that are also available at the same site.

From Linux: Dave Graham has created a Linux application to create VSN labels on the Dymo printer. It is available at <http://www.mpifr-bonn.mpg.de/EVN/dymolab.html>.

5. e-VLBI Network ‘Weather Map’ Introduced

Researchers at Haystack have developed and deployed an automated system for monitoring the quality of e-VLBI transport networks. This new system allows users to view current network usage statistics on a link-by-link and segment-by-segment basis using a combination of passive and active monitoring. Statistics include current link load over various timescales, as well as current TCP throughput between nodes in the network and between various performance monitoring servers throughout the US Abilene network and the TransPAC/APAN networks. Passive monitoring statistics are provided courtesy of Abilene/TransPAC APAN network monitoring sites, while active monitoring statistics are provided through the use of Internet2's Bandwidth Control tool (<http://e2epi.internet2.edu/bwctl/>) and Haystack's Network State Database (NSDB) tool.

Currently, Tokyo XP, Abilene, Washington ISI-E and Haystack are participating in this network monitoring, which can be viewed at <http://web.haystack.mit.edu/staff/dlapsley/tsev7.html> and is shown in Figure 1.

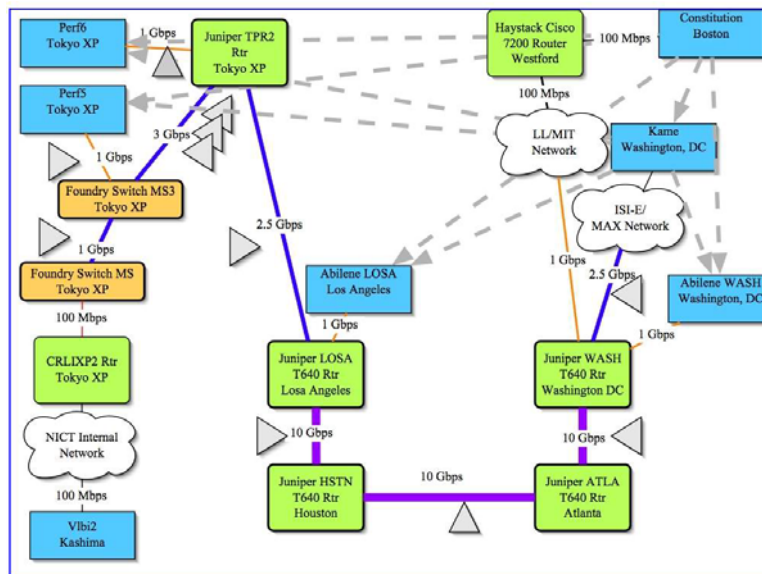


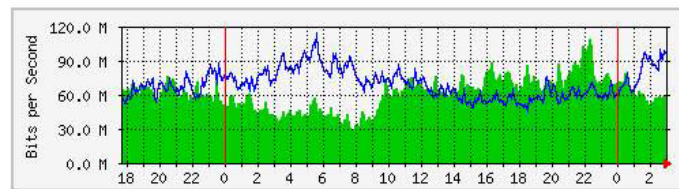
Figure 1: Clickable e-VLBI 'Weather Map' for Haystack to Kashima

Any of the triangles may be ‘clicked’ to show the performance details of that segment. For example, Figure 2 shows a portion of the statistics for the TransPAC link between Los Angeles and Tokyo XP.

Traffic Analysis of TransPAC LA link

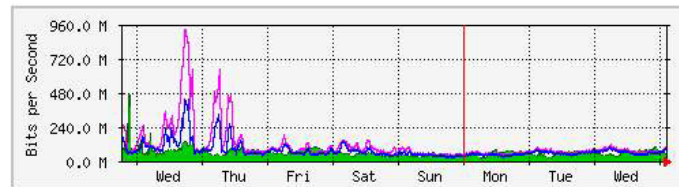
The statistics were last updated **Thursday, 8 July 2004 at 3:01**,
at which time 'tpr2' had been up for **78 days, 12:30:19**.

'Daily' Graph (5 Minute Average)



Max In:109.4 Mb/s (4.6%) Average In:61.3 Mb/s (2.6%) Current In:61.1 Mb/s (2.5%)
Max Out:114.0 Mb/s (4.8%) Average Out:70.4 Mb/s (2.9%) Current Out:96.9 Mb/s (4.0%)

'Weekly' Graph (30 Minute Average)



Max In:475.8 Mb/s (19.8%) Average In:62.0 Mb/s (2.6%) Current In:56.1 Mb/s (2.3%)
Max Out:927.5 Mb/s (38.6%) Average Out:78.8 Mb/s (3.3%) Current Out:85.0 Mb/s (3.5%)

Figure 2: (Partial) Traffic history on Los Angeles - Tokyo XP segment

This website marks the first time passive and active network performance statistics have been integrated for use with e-VLBI. This is also the first time that it has been possible for network servers at the edge of the network to test into servers located within the core of the network to allow partial path analysis (the ability to determine throughput on a link-by-link basis). These statistics help to provide an overall view of the current network state, track changes in network throughput over time, and detect and isolate network/end system faults. Indeed, it was instrumental in helping to isolate and correct network faults in recent network transfers.

In the near future, we plan to extend the e-VLBI ‘Weather Map’ to include Onsala, Wettzell, Kokee, Westford and GGAO. If you would like to add your site to the Weather Map, we would be happy to cooperate with you in doing so; please contact David Lapsley at Haystack.

The software used is freely available (although some is still in alpha stage). If you would like to add your sites/servers to this network, please feel free to contact David Lapsley at Haystack (dlapsley@haystack.mit.edu).

6. UT1 Result in 4½ hours!

Researchers at Kashima Space Research Center and MIT Haystack recently conducted an experiment in which UT1-TAI estimates were achieved 4 hours, 27 minutes and 58 seconds after the recording of the last scan for an observing session had finished. The ‘tsev8’ experiment was conducted on 29/30 June 2004 and involved the Kashima and Westford telescopes. Data were recorded at Kashima in K5 format and at Westford in Mark5 format. The data were then transferred from Westford to Kashima using high-speed research and education networks that

included Internet2's Abilene network, the APAN/TransPAC and JGN2 networks. The data were then translated into K5 format and correlated using software correlation at Kashima. The new elapsed time of 4 hours, 27 minutes and 58 seconds is a significant improvement over the previous record of 21 hours that was set a year ago with the 'tsev6' experiment.

The e-VLBI 'Weather Map' (see above) was instrumental in helping us track down an elusive problem in the Haystack/Kashima connection, which turned out to be a bad NIC card in one of the servers along the path!

Many thanks to Koyama-san, Hirabaru-san and their colleagues at the National Institute for Information and Communications Technology, Brian Corey, Mike Poirier and Jason Soohoo at MIT Haystack Observatory, the staff of the APAN Tokyo XP NOC and Tom Lehman from the University of Southern California, Information Sciences Institute East for their hard work in support of this experiment.

7. Conduant return policy update

Due to the large number of Mark 5 systems and disk modules in service, Conduant has had to update its current fairly casual return policy to be somewhat more formal. Conduant hopes these changes will result in better service to its customers. For details, see <http://www.supportcenteronline.com/dmfiles/639/669/Support%20Downloads/Request%20to%20Return%20Items%20to%20Conduant/ConduantReturnMaterialAuthorization.pdf>.

For any questions you may have for Conduant, please contact Kipp Hayes at kipp@conduant.com. If you are contemplating a purchase, please contact Kipp for the latest Conduant price sheet.