Full steam ahead - 1024Mb/s data rate for the e-EVN First Draft

Paul Boven

January 8, 2008

1 Introduction

The e-EVN has recently announced that observations at 512Mb/s are now available for production. The highest speed that the current EVN correlator can support is 1024Mb/s, but this exceeds the capacity of 1Gb/s Ethernet connections. This document investigates what changes to the network would be needed to support this highest data rate for some or all e-EVN telescopes.

2 Bandwidth requirements

The 1024Mb/s is only the data generated by the formatter at the radio-telescope: in order to transport it across the network, address headers and flags need to be added. Each UDP packet contains 20 bytes of IP header, 8 bytes of UDP header and 8 bytes for an internal sequence number. The number of bytes in the payload also needs to be multiple of 8. De number of payload bytes per packet is therefore (MTU - 36), rounded down to a multiple of 8. If an MTU of 9000 can be used on the network, there will be 8960 bytes of e-VLBI data in each packet. The MTU however does not take into account the 8 bytes of Ethernet header and 2 bytes of Ethernet checksum, so the actual size of the packets will be 8960 + 36 + 10 = 9006 bytes. The required network throughput for 1024Mb/s will be $1024 \cdot \frac{9006}{8960} = 1029.3$Mb/s. With an MTU of 1500, this increases to 1055.5Mb/s.

3 Lightpaths

The majority of the telescopes that are active in the e-EVN are connected to JIVE by a dedicated lightpath. A lightpath is an end-to-end tunnel provisioned within one or more SDH/SONET networks. It is created by dedicating a number of time-slots on consecutive SDH links to the lightpath. The capacity corresponds to the number of time-slots available per link, and is usually configurable in steps of 150Mb/s (a VC-4 channel). A 1Gb/s lightpath is configured as 7 or 8 of these channels, giving a theoretical capacity of 1050 or 1200 Mb/s. This means that in theory, a 1024Mb/s stream, including its overhead, would fit into an existing 1Gb/s lightpath. Supporting the highest e-VLBI data rate would therefore be possible without using additional SDH networking resources.

The lightpath is terminated at both ends on Path Terminating Equipment (PTE) which provides connectivity through regular Ethernet interfaces. Ethernet comes in several distinct speed-grades (10, 100, 1000 and 10000 Mb/s) and the speed of the interface that is provisioned to connect to the lightpath does not necessarily match the speed of the lightpath, nor do both ends of a lightpath need to have the same kind of Ethernet termination. At the moment, all our lightpath are terminated with 1Gb/s Ethernet interfaces. And although the lightpath itself might have enough spare capacity for full rate e-VLBI, the gigabit Ethernet connections at both ends of each lightpath are the bottlenecks.

3.1 L2SS functionality in PTE

In most SDH Path Terminating Equipment, there is a direct one-on-one mapping from an SDH lightpath (concatenated channels) to its associated Ethernet interface. Recently, SDH equipment vendors have started including layer 2 (Ethernet) switching capabilities within their Ethernet interface cards. This adds a lot of flexibility in how Ethernet traffic is delivered into the lightpath, and turned back into Ethernet frames again. Some examples of new capabilities enabled by this concept are:

VLAN support Gives a sender access to multiple lightpaths based on VLAN tags, without needing additional Ethernet ports on the PTE.

Trunking By assigning multiple Ethernet ports to a single lightpath, it becomes possible to better use the capacity without resorting to higher speed interfaces.

Aggregation The ability to terminate multiple 1024Mb/s lightpaths on a single 10Gb/s Ethernet port, properly separated by VLAN tags. A single 10Gb/s Ethernet link could carry up to 9 1024Mb/s e-VLBI streams, which would otherwise require 18 gigabit ports.

The SURFnet OME6500 in Dwingeloo does not currently include any Ethernet cards with L2SS functionality. But from discussions with SURFnet it seems possible to terminate some lightpaths on equipment in Amsterdam that does have this capability. There is also a 10Gb/s routed connection from SURFnet to JIVE, which is rate-limited to 5Gb/s for routed traffic. The surplus capacity could be used to deliver the lightpath traffic to JIVE. Such an experiment would of course only make sense if the transmitting telescope is able to deliver the traffic at the 1024Mb/s rate at their end of the lightpath, either by using L2SS, dual lightpaths or a 10Gb/s interface.

4 Path overview

In order to increase the overall throughput from telescope to correlator, we need to take all parts of the network connection into account:

- Telescope Mark5
- Network between Mark5 and PTE
- PTE
- Lightpath capacity
- PTE (In Dwingeloo or Amsterdam)
- Network between PTE and JIVE switch/router
- JIVE switch/router
- JIVE Mark5

At this time almost all network interfaces and links in the path are 1Gb/s, with the notable exception of the lightpath itself which has a slightly higher capacity.

5 Trunking

The capacity of a connection can be increased by adding one or more connections in parallel to the existing one, and distributing the traffic evenly over all members of that trunk. This is often much easier to achieve than upgrading an existing connection to a higher speed. The complication however is in distributing the traffic evenly over all trunk members. The standard way of creating Ethernet trunks is by use of the LACP protocol. But this standard requires that all traffic that

belongs to a single TCP or UDP stream be sent over the same link member, to prevent packets from arriving out-of-order at their destination. LACP works fine for adding capacity on links that are used by many users, but will not work for e-VLBI because it uses a single TCP or UDP stream. Most routers and switches only support LACP-like trunking.

Recent versions of the Linux kernel support link aggregation by distributing packets in a round-robin fashion over two or more interfaces. Although most switches and routers are not able to send traffic in this way, they have no problem in receiving it. By keeping the channels of the trunk separate (e.g. by ways of separate fibers, VLANs or lightpaths) only the sending Mark5 determines which packet goes into which link member, and trunking should work. Creating trunks in this way might create some trouble for the return traffic, but fortunately, e-VLBI doesn't use any. We have even successfully used one-way lightpaths for observations in the past.

In general, if we want to use trunking, we should have the sending Mark5 decide on the packet distribution. Once the two links are joined somewhere in the network, the remaining part of the link must have enough capacity for the full 1024Mb/s because only LACP trunking would be available from that point onward.

A case in point is the JIVE switch/router: as a router, it separates the Ethernet segment with the remote Mark5 from the Ethernet segment with the Mark5 at JIVE. As this switch only understands LACP trunking, the link between the switch and the receiving Mark5 cannot be a trunk, and therefore must be a 10Gb/s Ethernet connection.

6 IP-level trunking

Instead of having the kernel of the sending Mark5 distribute the packets across the link members, it would also be possible to have the application software take care of this at the IP level. Alternating packets would be addressed to the two IP addresses of the receiving Mark5, and proper setup of the routing tables at the sending Mark5 must ensure that the difference in destination IP address will actually cause the packets to exit from different networking interfaces.

One of the disadvantages of this method is that it doubles the required number of IP addresses at both the sending and receiving Mark5, and the software interfaces would need to be changed to accommodate that. Another drawback is that the higher network capacity would only be available to applications that are specifically written to make use of it. Network diagnostics software such as e.g. iperf would not be able to use or test such a link at full speed. But despite these disadvantages, this might still be an avenue of research worth exploring.

7 Overview per station

7.1 Onsala Space Station

Onsala is connected to JIVE via regular routed IP networking through NORDUnet and SURFnet. The network connection at OSO has recently been upgraded to 10Gb/s. Upgrading their Mark5 with a second Gigabit Ethernet interface and using trunking between the Mark5 and their switch/router would probably be sufficient to support 1024Mb/s operations from Onsala.

7.2 WSRT

The Westerbork Synthesis Radio Telescope is run by Astron, who provide the office space for JIVE as part of their contribution to the EVN. There is a 34km dark fiber connection between the WSRT and the Astron/JIVE building. We are planning to upgrade the capacity of this connection by using CWDM equipment in January 2008, and adding a second gigabit Ethernet interface to the WSRT Mark5.

7.3 Effelsberg

There will be a 10Gb/s connection between the Effelsberg radio telescope and the LOFAR processing center in Groningen, which is run by Astron. This connection is currently in the final stages provisioning. This connection will be used for both connecting to German e-LOFAR stations, and for e-VLBI with the Effelsberg radio telescope. There is also enough fiber capacity between Groningen and Dwingeloo to forward the e-VLBI data to the JIVE switch/router. MPIfR, ASTRON and JIVE are currently discussing how to provision this connection.

7.4 Jodrell Bank

We have two 1Gb/s lightpaths from Jodrell Bank to JIVE, to be able to use two telescopes from the Merlin network for e-VLBI. We generally use either the Lovell Mk1 or Lovell Mk2, and one of the other Merlin telescopes at the same time. The Merlin telescopes that are not located at Jodrell Bank are connected to the observatory by microwave links with a 128Mb/s capacity. It would be fairly simple to add a second 1Gb/s interface to the 'local' Mark5 at Jodrell Bank, and connect it to both lightpaths. This way we would be able to use one of the Lovell telescopes at 1024Mb/s, but using a second Merlin telescope at the same time would of course be impossible.

7.5 ATNF

We still have three 1Gb/s lightpaths from Australia to JIVE, which were used to observe SN1987A, amongst others. These lightpaths are connected to the ATCA, Mopra and Parkes telescopes. All the lightpaths terminate in Sydney, with the ATNF network providing connectivity between each lightpath endpoint and a telescope. Depending on the networking resources available within the ATNF it might be possible to connect one telescope to two of the lightpaths using trunking. As these systems are not based on the Mark5 hardware, they can run a much more recent Linux kernel, which has better support for the Ethernet bonding options we would like to experiment with.

It might even be possible to use two telescopes at 1024Mb/s each by carefully engineering how the packets are distributed over all 3 lightpaths, for an average rate of 682Mb/s per lightpath. As Europe and Australia are at almost opposite ends of the earth, it is hard to find sources that are in view at the same time from both locations, so the ability to use 2 Australian telescopes at the same time would make it much easier to detect fringes at this data rate.

7.6 Poznan, Medicina, Torun

The networking situation at these 3 telescopes is fairly similar. In each case there is a single Mark5, some local networking resources to connect it to the PTE, and a 1 Gb/s lightpath through GÉANT2 to JIVE. Achieving 1024Mb/s for these stations depends on their local networking resource, as well as the capacity of the lightpath. Upgrading these stations is only possible if GÉANT is willing to upgrade these lightpaths. This upgrade could consist of splitting it into two smaller lightpaths (e.g. 2x 600Mb/s), or upgrading the lightpath to support 1024Mb/s, with 2x 1Gb/s ports with L2SS, or a 10Gb/s port at the originating PTE. This would also depend on the willingness of SURFnet to upgrade these lightpaths at their end and provide the right kind of termination, either in Dwingeloo or Amsterdam (see 3.1).

7.7 JIVE

Because the JIVE switch/router is the boundary between Ethernet segments from the telescopes on one end, and the 16 receiving Mark5's at JIVE at the other end, it is not possible to use trunking between the switch and the JIVE Mark5's. This mandates the use of 10Gb/s technology between the switch and the JIVE Mark5's. Originally we had planned to use 10Gbase-T (RJ-45 based) links for that, but unfortunately the availability of such interfaces for our switch has been delayed by at least half a year by HP.

The cheapest short-term alternative would be to use CX-4 (Infiniband) based copper connections between switch and Mark5, as these are readily available. The JIVE 5412zl switch has one J8707A 4-port 10Gb/s switch module, of which 3 ports are still unused. For each Mark5 that we want to upgrade to 10Gb/s, we would need a J8440B X2-CX4 transceiver to plug into the module, an CX-4 cable, and a CX-4 capable 10Gb/s PCI-X or PCI-Express card. If more than 3 Mark5 were to be upgraded in this way, we would also need additional 4-port J8707A modules.

In 2008Q3, HP is planning to introduce a 6 port 10Gbase-T module for the 5400-series of switches. This would be the most cost-effective way to upgrade all the Mark5 to 10Gb/s connectivity by then.

8 Conclusion

Upgrading some, or even all of the e-VLBI 1Gb/s connections to support 1024Mb/s seems feasible, with only modest investments in networking equipment and capacity. The different networking situations at each of the telescopes will require detailed planning together with each station, GÉANT and SURFnet.