PROCEEDINGS OF SCIENCE

European 4 Gbps VLBI and e-VLBI

Guifré Molera Calvés*
TKK/Metsähovi Radio Observatory, Kylmälä, Finland
E-mail: gofrito@kurp.hut.fi

Jan Wagner
TKK/Metsähovi Radio Observatory, Kylmälä, Finland
E-mail: jwagner@kurp.hut.fi

Jouko Ritakari
TKK/Metsähovi Radio Observatory, Kylmälä, Finland
E-mail: jr@kurp.hut.fi

Ari Mujunen
TKK/Metsähovi Radio Observatory, Kylmälä, Finland
E-mail: amn@kurp.hut.fi

This paper discusses an implemenation of a high data-rate recording system for VLBI and e-VLBI applications. Using commercially available components, a low-cost system has been built and tested to fulfill the EXPReS project requirements. The results described in the paper successfully demonstrate a recording prototype which can handle sustained data rates beyond the goal of 4 Gigabits per second. Furthermore, data can be losslessly streamed over the Internet to its final destination at the correlator center. The improvement of the system is tightly dependent on advances in hardware components that would make achieving a throughput of 8 Gbps feasible in the next couple of years.

This work has received financial support from the European Commission (DG-INFSO), within the Sixth Framework Programme (Integrated Infrastructure Initiative contract number 026642, EXPReS).

Science and Technology of Long Baseline Real-Time Interferometry:
The 8th International e-VLBI Workshop, EXPReS09
June 22 - 26 2009
Madrid, Spain

*Speaker.

PoS(EXPReS09)083

1. Introduction

In the EC-funded EXPReS project (Express Production Real-time e-VLBI Service) the overall objective is to create a production-level e-VLBI service in which the European VLBI Network (EVN) radio telescopes are reliably connected via high-speed optical-fibre connections to the JIVE correlator node in Netherlands, with an aggregate data flow of up to 16 Gbps. The task of Metsähovi Radio Observatory (MRO) is to provide a prototype of a new network-connected multi-gigabit Data Acquisition System (DAS) with real-time data transfer and high-speed storage capabilities. Some test results leading to such a system are described in this article.

Since the beginning of the year 2009 the Finnish radio-astronomical station has been actively participating in many e-VLBI observations. During this period MRO took part in such events as the *24 hours of continuous e-VLBI observation during the International Year of Astronomy opening ceremony* [1] that ran at a 256 Megabits per second (Mbps) rate; the *100 hours of astronomy* [2], scheduled at 512 Mbps and also various 24 hours real time e-VLBI geodetic sessions at 128 Mbps. During the previous year 2008, many data-rate records were broken in the world of astronomy. A 4-Gbps data transfer from the Swedish radio-astronomical station Onsala to the Finnish station, with remote recording, was demonstrated during observations of water maser emission from the Orion constellation. Finally, an 8-Gbps stream from Metsähovi to Onsala using iBOB FPGA-devices was demonstrated during the International e-VLBI Workshop in Shanghai [1].

2. Primary goals for European 4 Gbps e-VLBI

The primary goals for this project are listed below:

- Test the limits of Commercial Off-The-Shelf (COTS) computers.
- Test new 10 Gbps devices: switches, Ethernet cards or FPGA-based devices such as iBOB, BEE2, ROACH [3], dBBC [4].
- Test the performance of the Internet in high speed transfers and long distance baselines.
- Test the performance of hard disks, RAID disk controllers and Port Multipliers (PMP).
- Develop simple UDP-based data transfers protocols: Tsunami-UDP, VDIF UDP packetizer, VSIB multicast [2].
- Build a DAS capable of streaming and recording at 4 Gbps, compatible and interoperable with Mark5 A/B/C systems.

[1] http://http://www.expres-eu.org/iya2009/

[2] http://www.100hoursofastronomy.org

[3] iBOB, BEE2 and ROACH are a set of FPGA devices with radioastronomy purpose created by the CASPER group from Berkeley University, http://casper.berkeley.edu.

[4] digital BaseBand Converter, in production at HatLab.

3. Methodology

In order to achieve these objectives, several COTS components have been extensively tested over the past 18 months. A compromise between cost, performance and OS compatibility is strongly required. Hence, we conducted a rigorous search for equipment capable of supporting high data-rates. This resulted in the selection of two test systems. Table 1 shows both models that were built as new high capacity DAS. Table 2 lists specific components, either selected or discarded, that were used in the tests.

Components	**System 1**	**System 2**
Motherboard	Asus L1N-SLI WS	Asus Rampage Extreme II
Processors	Two dual-core AMD	Pentium Quad-core
RAM memory	DDR-2 4 GB	DDR-3 4 GB
SATA controller	Native ports	PCI Express RAID controller
10 Gbps ETH card	Chelsio 10 Gbps	Myrinet 10 Gbps
HDD disks	Samsung F1 750Gb	Samsung F1 1TB
Max capacity (with 2TB)	24 TB (12 disks)	40 TB (20 disks)

Table 1: High data-rate systems built to record over 4 Gbps streams.

PCI Express RAID controller	**Motherboards**	**Port Multipliers (PMP)**
HighPoint RocketRaid 2522	Asus L1N-SLI WS	AD5SAPM
Hewlett-Packard SC44Ge	Asus Rampage Extreme II	AD5SAMP-E
Addonics ADSA3GPx8-4	Asus Striker Extreme II	
	Asus P5Q Pro	

Table 2: List of components tested. It includes RAID controllers, Motherboards and Port Multipliers from different manufacturers.

Two sets of tools to benchmark the performance of the system, RAID disk controllers and hard disk modules were developed in-house: *wr* [3] and *Tsunami-UDP* [4]. The *wr* software kit was developed to read data from a VSIB board used in the VLBI observations and record onto the PC-EVN. The *wr-nexgen* is a software tool used to perform sequential RAID I/O benchmarks with versatile settings, such as raw and formatted disk writing modes or multi-thread and parallel writing capabilites. The toolkit was improved in 2009 with real-time priority, CPU affinity and similar optimizing load processes features.

Tsunami-UDP is a fast aggressive FTP, that can also be used as a data transmission protocol. Tsunami was developed at Pervasive Technology Labs research center at the University of Indiana and currently is maintained completely at MRO. Transmission data are sent as UDP/IP packets at a significantly higher transfer rate than that of TCP, especially over long distances. The basic idea behind Tsunami-UDP is that the transmission data are chopped into large packets of equal size and the server tolerates a long delay of the acknowledge packets from the client site.

4. Results

The selection of the hardware components was clearly a key factor to achieve our targets. To ensure a good overall performance each single piece of the chain of components must be carefully investigated. There are important concerns such as the true writing/reading rate of the hard disks, the capability of the RAID disk controllers to handle the data rate, the efficiency of 10-Gbps Ethernet cards and in particular the overload of the processor's core due to the amount of simultaneous CPU-intensive tasks running on the computer. As an example, a careful selection of SATA disks resulted in an improvement of up to 33 % of the write disk speed.

Based on hard disk benchmarks published in specialized journals [5], the Samsung F1 Spinpoint HDD was selected to populate our systems. We confirm the previous findings and conclude that these units were at the time far ahead of their competitors in terms of writing/reading capabilities. Regarding disk benchmarking, XFS was used in our tests as a standard File System, because of its faster write & read operations under Linux. Currently we are investigating other possibilities such as HFS+ under Mac OS X 10.5. On a Mac PC clone system a 10-25 % speed improvement over the Linux XFS results was achieved.

Figure 1 shows a comparison between disk capacity and the progress of the write speed as the disks are filled up with data. As seen in the graph the recording was done on both raw (blue) and XFS-formatted (red) disks. Although we at first believed that the type of file system does not have a strong influence on the behavior of the disks, we soon realized that the processor limitation and the poor Linux *pdflush* implementation lowered the performance by up to 25 %. In any case, it seems possible to record over 4 Gbps by using XFS, but we are still working within quite narrow margins to ensure trouble-free data recording. To conclude, the raw/XFS plot demonstrates that 4 Gbps recording can be achieved with carefully selected equipment. In this case, 12 disks recorded for over 5 hours at rates above 4 Gbps.

The chipset SATA ports were another concern. Initially we contemplated using the on-board chipset controllers on the motherboards. Configured as a RAID they proved to be faster than any external controllers. But the almost non-existent support for PMP modules in any of the chipsets is a tough obstacle when using more than 12 disks in RAID mode within a single PC. On the other hand, PCI Express-based RAID disk controllers offer easy expansion capabilities thanks to their PMP support, good write performance and a possible reduction in CPU processor load.

Finally, the long-haul and backbone network link tests were extremely satisfactory. We showed that 8 Gbps continuous streaming over the Internet via NRENs [5] was possible. We found no errors or packet loss on the optical-fibre link during the Shanghai demonstration [1]. The Finnish and Swedish NRENs (Funet and SUNET) played an active role in testing the capacity and VLBI-usability of their international links. These tests led to recognition in the Finnish press [6, 7] for all involved parties. The normal Internet protocols and northern NRENs were demonstrated to support

[5] National Research and Education Network

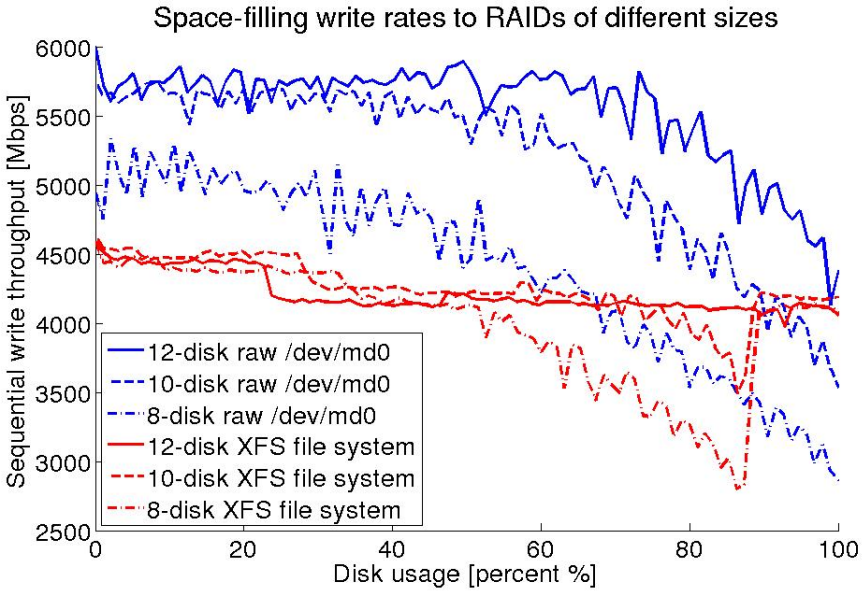

Figure 1: Writing performance on RAID disks, using 12,10 and 8 disks. The blue lines show recording on raw disks and red with XFS as a File System.

all the requirements and constraints of e-VLBI with no disruption to other Internet traffic. Considering the even higher trunk capacity and lower load of Central European backbones and NRENs, extrapolating from our experience with Funet, SUNET, the DFN and GÉANT2, we recommend a simple connection without QoS, BOD or other services for all the EVN radio telescopes in the community for connecting to a central or distributed correlator – for classic single-site, Cloud or Grid correlation, respectively, without traffic routing limitations.

Server microbursting of large data packets combined with the small buffer memory in some low-end switches and routers is a problem that warrants closer inspection for smoother streaming over the Internet. Thus, the next generation of DAS based on FPGA designs that do not exhibit microbursting may contribute to a direct reduction in e-VLBI transfer loss. The FPGA can ensure a fixed inter-packet delay for which the user has full control. When streaming is implemented on standard PCs, even if the user sets a fixed inter-packet delay in software, the generated traffic can still be transmitted as microbursts. The user does not have easy control over certain low levels of the computer architecture such as the PCI Express bus, Ethernet board or the driver. The network layer and driver can re-group packets regardless of userland settings. Disabling interrupt coalescing to reduce microbursts is effective only to a limit. With current hardware it still comes at the cost of reduced throughput or increased system load.

5. Conclusions

The results from the stream recording tests proved that nowadays it is already possible to record network data with a single Linux or Mac OS X computer at a sustained 4 Gbps rate using consumer hardware. Hence, Metsähovi has built and now offers expertise for low-cost, low-power,

and high-speed storage solutions, for such cases as temporary VLBI station data storage, high-bandwidth data acquisition systems with file capture and Network Streaming including file transfer for VLBI, at unprecedented data-rates. An example of a 4G-EXPReS diskpack for use as a new European DAS is shown in the left-hand panel of Figure 2. The right-hand panel shows a Mark5 unit with retrofitted 4G + RAID capability. For hardware suggestions please refer to the Metsähovi web page [8].

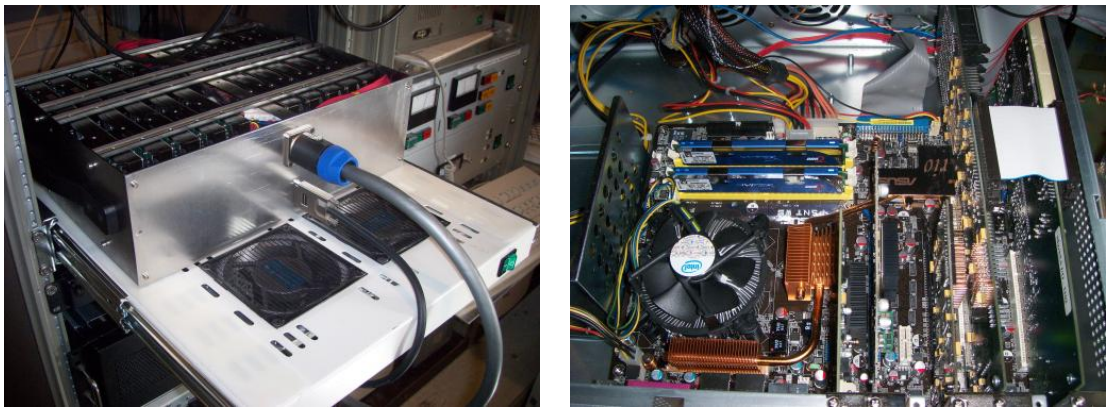

Figure 2: The left-hand panel shows the new disk pack built at MRO. It has a capacity for 20 disks, connected using PMP and a Infiniband cable to the main server PC. Below the disk packs there is a cooling tray to maintain constant temperature over the disks' surface. The right-hand panel shows a possible new configuration for the Mark V units with the existing Conduant boards + a new 10 Gbps Ethernet card and a RAID disk controller. This unit could run both Mark5 A/B and the new Metsähovi-designed system.

We further plan to improve the UDP-based data transfer protocols mentioned above. A UDP multicast mode has already been successfully tested in VLBI experiments with Sweden and Japan for real-time correlation. Future StreamStor FPDP real-time multicasting is in the test phase. For UDP recording of 8 to 32 Gbps streams we plan to evaluate a multi-storage solution. Further applications of the transfer protocols may be in our real-time software correlation, RF spectrometer and VLBI space satellite tracking software currently developed on Intel architecture, and versions for Playstation3 [9] and CUDA graphics [10] architectures, currently under development.

References

[1] *7th International eVLBI Workshop*, http://www.shao.ac.cn/eVLBI2008/, 2008 ;

[2] *Set of tools for VLBI and e-VLBI*, http://www.metsahovi.fi/en/vlbi/vsib-tools/index, 2009;

[3] A. Mujunen, *Installing VSIB/VSIC Test Software*, http://www.metsahovi.fi/en/vlbi/vsib-docs/pre-sw-inst.pdf, 2003 ;

[4] J. Wagner et al. *Tsunanmi-UDP source code*, http://tsunami-udp.sourceforge.net/ ;

[5] *Samsung Spinpoint F1 HDDs: New Winners*, http://www.tomshardware.com/reviews/samsung-overtakes-a-bang,1730.html ;

[6] *Suomalaisille maailmanennätys radiosignaalin siirrossa internetissä* http://m.digitoday.fi/?page=showSingleNews&newsID=200817436 2008;

[7] *Funet network rate more than 8 Gbps, Metsähovi Radio Observatory sets world record* http://www.csc.fi/english/csc/news/news/funet_metsahovi 2008;

[8] J. Wagner, G. Molera et al. *4G-EXPReS Hardware Recommendations*, http://www.metsahovi.fi/en/vlbi/ibob/4gexpres, 2008 ;

[9] *IBM Cell miniature software correlator*, http://cellspe-tasklib.sourceforge.net/ , January 2008 ;

[10] *High Performance Computing with CUDA*, http://www.nvidida.com/object/GPU_computing.html , 2008;