Russian Data Recording System of New Generation

Ilya Bezrukov[1], Alexandre Salnikov[1], Andrey Vylegzhanin[2]

Abstract The IAA RAS is developing a data recording system for observations with radio telescopes of small diameter (13 m), which are also intended to realize the goals of the VLBI2010 international program. The main system goals are:

- recording of eight data streams (with scalability up to 16) in the VDIF format with a data speed of 2 Gbps from each channel;
- realizing data transfers to the Data Processing Centers at 10 Gbps transfer speed simultaneously with recording and buffering data;
- storing session data up to 20 TB in size in a generic file structure with a set of disk pools.

Keywords BRAS, DRS, VLBI

1 Features of the Data Recording System

In accordance with the experience of our colleagues [1, 2], we decided to create a Data Recording System (DRS) based on Commercial Off-The-Shelf (COTS) hardware.

Our development stand is a rack server (Figure 1) Dell PowerEdge R720 with two Intel CPUs: Xeon E5-2643 3.30 GHz or Xeon E5-2650 2.0 GHz, 96 Gb RAM, and two disk enclosures Dell PowerVault MD1220 (up to 24 2.5″ hot-pluggable small-form-factor drives). With this configuration, the disk subsystem of the server setup consists of three

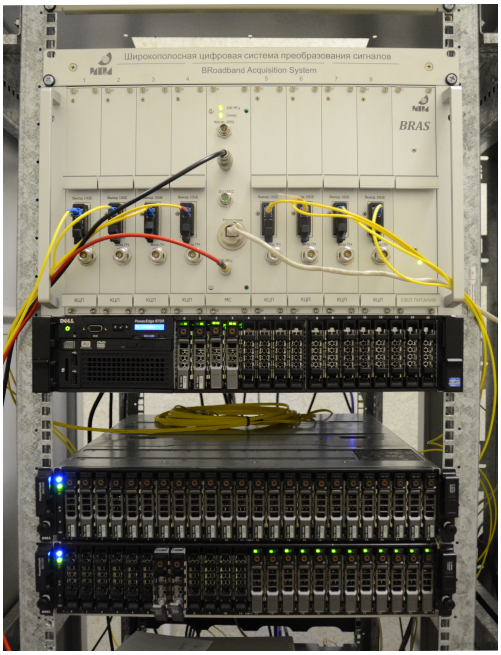

Fig. 1 The development stand. Up: digital backend BRAS [4]. Down: Server Dell PE R720 with storage Dell PV MD1220.

SAS backplanes, up to 64 2.5″ drives maximum. Each SAS-backplane is attached to LSI SAS2008 based SAS HBA by two 24 Gbps SAS 2.0 channels. Up to four dual-port 10 Gig Ethernet Intel network cards are used (Intel X520). The connection of the DRS with the digital backend was carried out in two ways: a direct connection of each channel BRAS to a server 10 GbE network interface (Figure 2a) and a connection through the Cisco Catalyst C4900M switch (Figure 2b). It also revealed first experimental data on registration of test signals and transmission of the data through a fiber

1. Institute of Applied Astronomy, Russian Academy of Sciences
2. Ioffe Physical Technical Institute, Russian Academy of Sciences

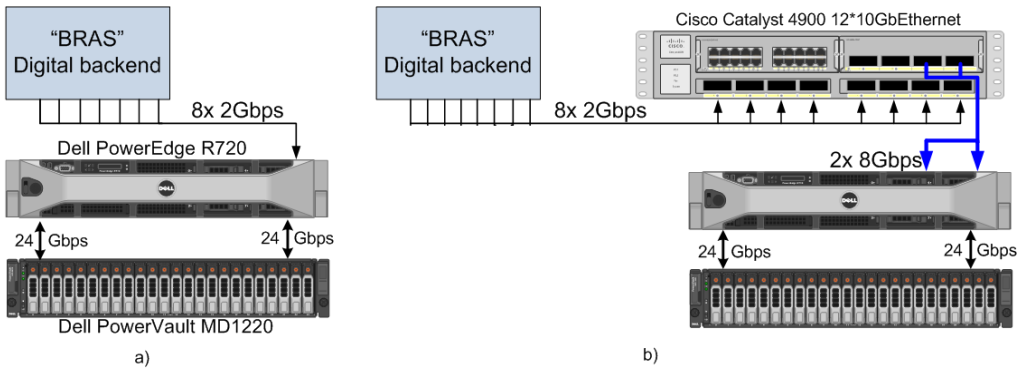

Fig. 2 Experiment diagram. a) direct connection and b) connection through 10 GbE switch.

optic channel with a maximum bandwidth of 10 Gbps between the IAA sites in St. Petersburg, Russia.

2 Packet Capturing and Recoding Algorithm

A block diagram of the data buffering software (application) is shown in Figure 3. The application has a multi-threaded architecture. Packet capture and management of the network interface is performed via a netmap [3] framework controlled by receiver thread. The receiver performs packet pre-processing and data transferring from the netmap circular buffer to an interim buffer. The data recording from the interim buffer on disk volumes is performed by a writer thread. The netmap is a framework for high speed packet I/O which has been a native part of a FreeBSD kernel since late 2011. This software interface, developed and maintained by Luigi Rizzo (member of FreeBSD team), can effectively process I/O packets at the maximum speed of the network interface, bypassing the standard kernel network stack interface. Unlike other frameworks for high packet processing, the netmap operates in user space memory, which excludes the possibility of a kernel crash. To achieve high speeds of packet processing, the application uses some of netmap's performance-boosting techniques, such as memory-mapping the network card's packet buffers and I/O batching. Network interface interrupt handlers and stream instances of thread <<receiver>> were also bound on the same CPU core, and features of Sandy Bridge processor architecture, memory management, and PCI express slot geographical addressing were taken into consideration.

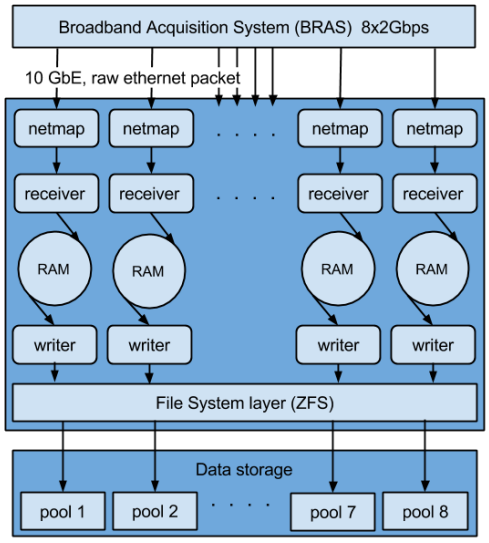

Fig. 3 Block diagram of the data buffering software.

3 Assessment of the Performance of the Disk Subsystem Data Buffering

The DRS is running under the FreeBSD 10.0-RELEASE operating system (in February 2014) with the ZFS (Zettabyte Filesystem). The ZFS is a stable FS actively developed by the world community that combines the functions of file system, logical volume manager, and software RAID. For a range of different disk configurations, we performed testing by

simulating recorded data in Intensive session mode. To emulate this, a 10 GB random data set (white noise) was created in RAM and recorded 60 times by eight threads with 20 second intervals to eight ZFS pools. Data recording to the ZFS disk pools was performed with the standard Unix utility "dd". Scoring of the writing speed to the ZFS pools was performed with utility "zpool iostat". We tested SAS 10k rpm, low cost MLC SSD, NL SAS 7.2k rpm, and 3.5″ SATA 7.2k rpm drives.

Table 1 Disk performance.

Type of disk	Stripe pool, MBps			RAID-Z E5-2643		RAID-Z E5-2650	
	2 disk	3 disk	4 disk	3 disk	5 disk	3 disk	5 disk
SAS	**188**	**262**	**316**	140	**209**	117	143
NL-SAS	133	**200**	**241**	93	**185**	87	130
SSD	**274**	**338**	**403**	**189**	**236**	141	157
SATA	160	**189**	**192**	111	167	117	131

Table 1 shows the results of the disk types for stripe and RAID-Z (ZFS advanced RAID5 analog) pools. Results for the RAID-Z pools are given for two types of Intel Xeon processors: E5-2650 2 GHz and E5-2643 3.3 GHz. The red (thicker) numbers in Table 1 show transfer rates that fit requirements for DRS (180 MB/sec for 40 sec data streaming / 20 sec pause).

4 Experimental Research System Data Buffering from BRAS

During our research, we carried out data stream registration (recording) with 2 Gbps speed from each of the eight channels of BRAS. Figures 2a and 2b show the options of connecting BRAS to the DRS. We simulated the following mode of operation: a one hour session, eight data streams, each with 40 seconds of data recording and 20 seconds of pause. These total $60 \times 8 = 480$ files, 480×10 GB $= 4800$ GB total data. The recording was performed on the following configurations of ZFS disk pools:

- eight stripe pools, three SAS disks each;
- eight stripe pools, four NL-SAS disks each;
- eight RAID-Z pools, five SAS disks each;
- two stripe pools, 12 SAS disks each.

Additional testing was performed for simultaneous data transmission (with a transfer rate of 6 Gbps) to another server (Figure 4) and data recording of eight streams from BRAS (16 Gbps). For future DRS use, we performed successful preliminary testing for recording of a 32 Gbps data stream (4×8 Gbps BRAS channels) to four ZFS stripe pools, 12 SAS disks each (total 48 disks) with insignificant packet loss.

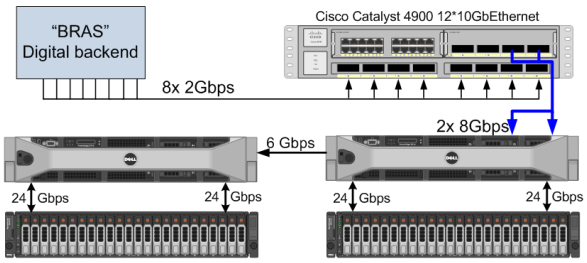

Fig. 4 Simultaneous data transmitting and data recording.

5 Future Plans

In the near future (2015), we plan to install our DRS on the RT-13 radio telescopes at the Badary (Siberia) and Zelenchukskaya (The Caucasus) observatories.

6 Main Results

The average data write speed for 60 scans (one scan/file size is 10 GB) to differently configured ZFS pools with the same disk type is about 2.5 GB/s, which is enough for recording 16 Gbps of a BRAS data stream.

With the multiple simulated Intensive session mode, we have demonstrated the ability to record eight data streams from a digital backend BRAS (total 2 Gbps $\times$ 8 $=$ 16 Gbps) without packet loss in two configurations:

- Direct connection of each BRAS channel to a DRS server;
- 10 Gbit Ethernet switched connection with 4$\times$ BRAS channels (2 Gbps) to one 10 Gbit Ethernet port of the DRS server.

The average measured transfer rate for the Tsunami-UDP protocol is:

- Between IAI RAS sites in city area, 2.7 Gbps (with MTU 1500) with simultaneous BRAS data stream recording;
- In 10GBit LAN segment, 7 Gbps (with MTU 9000).

References

1. Roger Cappallo, Chester Ruszczyk, Alan Whitney. Mark6: Design and Status. http://www.haystack.mit.edu/tech/vlbi/mark6/mark6_memo/05-Mark6_Design_and_Status.pdf.
2. Tomi Salminen, Ari Mujunen. Nexpres WP8 — FlexBuff. http://www.jive.nl/nexpres/doku.php?id=nexpres:nexpres_wp8.
3. Luigi Rizzo. The netmap project. http://info.iet.unipi.it/~luigi/netmap/.
4. Evgeny Nosov, Dmitriy Marshalov. Current Development State of Russian VLBI Broadband Acquisition System. In this volume.