SKA Project Series
SKA-LOW transient buffer proposal

G. Comoretto[1]

[1]INAF - Osservatorio Astrofisico di Arcetri

Arcetri Technical Report N° 5/2017
20-dec-2017

Abstract

The SKA-LOW telescope must include a transient buffer. The buffer must be able to capture up to 600 seconds of station data, from all the 512 LFAA stations, for a 150 MHz total bandwidth. Different proposals for the buffer implementation are compared.

1 Introduction

2 Transient buffer specifications

The transient buffer is specified in a series of Lev 1 specifications:

SKA1-SYS_REQ-3081 SKA1_Low, when commanded, shall generate and respond to real-time internal triggers by storing digitized voltage data, with 2-bit or better sampling, for at least 300 MHz of contiguous, tunable observed bandwidth in both polarizations, from every station within the triggering subarray, covering at least 10 seconds before (TBC) and at least 500 seconds after (TBC) the triggering event.

SKA1-SYS_REQ-3082 The SKA1-LOW shall have a system latency of at most 900 seconds from the time that the highest frequency component of a transient signal arrives at the telescope to the time when the transient buffer is frozen

SKA1-SYS_REQ-3083 The SKA1-LOW shall have the capacity of archiving at least 150 terabytes of transient buffer data per day

SKA1-SYS_REQ-3524 When commanded, SKA1_Low shall archive all or part of the transient buffer based on the results of single-pulse searches, independently for each subarray.

A schematic description of the buffer use case is as follows:

- channelized data being sent to the CSP for pulsar search operations is also re-quantized to 2 bits/sample, and formatted in independent SPEAD packets by the last TPM in the beamforming chain;
- these packets are sent to a circular buffer, with a total buffer space sufficient to hold the whole required time interval;
- When a transient is detected, the associated start time (including pre-trigger time) and stop time are sent to the buffers;
- When the stop time is reached, the stored interval is kept in the buffer (not overwritten) and dumped to the SDP for archiving;
- When enough buffer time is available for a second transient to be captured, the circular buffer operation is resumed;

As the time necessary to unload the buffer is considerably large, the possibility of detecting 2 consecutive events would require a double sized buffer.

2.1 Buffer size and data rate

The following assumptions have been made:

- At most 150 MHz of data per station and polarization needs to be stored in the buffer
- Total segment stored in the buffer is 900 seconds long
- No double buffering is required. After a transient has been captured, there will be some reasonable *blind time* in which the buffer is unloaded, and during which further events cannot be stored.
- Data format is 2 bits per sample, complex samples, double polarization. Thus each byte stored contains a complete complex sample for 2 polarizations, one station
- Samples refer to channelized oversampled data, with oversampling factor equal to $32/27 \simeq 1.185$, i.e. 1.185 samples per second per Hz of recorded bandwidth.

All these assumptions can be waived, but at an increased cost by an obvious numeric factor. E.g. increasing the band to 300 MHz would double the buffer size and data rates.

With these assumptions, each station generates 178 Mbyte/second (1.422 Gbps) of data for the transient buffer. The total buffer space required for 900 seconds of data is 160 GByte.

3 Possible solutions

In this section a brief description of the proposed solutions is given. A detailed analysis of the implementation is beyond the scope of this document. A comparison of the relative advantages and disadvantages is made in section 4

3.1 Distributed buffer storage

Each TPM has an internal DDR memory, that could be used as a distributed storage. A station processing chain is composed of 16 TPM, therefore each TPM must provide storage for 1/16 of the buffer space. The samples are available in the last TPM in the station beamformer chain, that disseminates back the re-quantized samples to the other 15 TPMs. When a transient event is detected, the DDR content is dumped to the 40G network. A schematic of this architecture is shown in figure 1

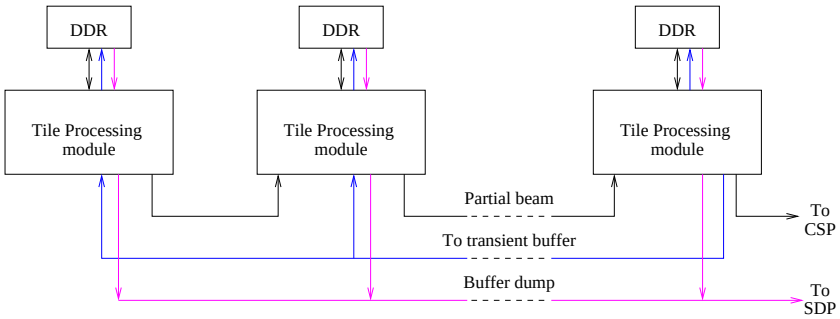

Figure 1: Transient buffer implementation using Tile Processing Module distributed memory

The internal switch network in the TPM rack can easily manage the extra 1.5 Gbps traffic, as the 40 GBE link in each TPM is used only for about 28 Gbps bandwidth. About 10 GByte of memory per TPM is needed. In the current design each TPM hosts 2 GBbyte of DDR, with plans to reduce this to 1 GByte or less. This solution therefore would require to expand the memory in each TPM to at least 10, likely 16 GByte. This is possible, but at a considerable cost.

This solution will require a non negligible design work to implement the required firmware, and to integrate it in the current design.

3.2 Dedicated buffer storage servers

In this solution the re-quantized samples are sent to a buffer server, i.e. a COTS computer with sufficient memory (at least 160 GB per served station) and memory bandwidth to host the transient buffer for all the serviced stations. The computer needs not to be dedicated to this service, as long as there is sufficient memory bandwidth between the network card and the memory. A DMA-capable network card could ease this requirement.

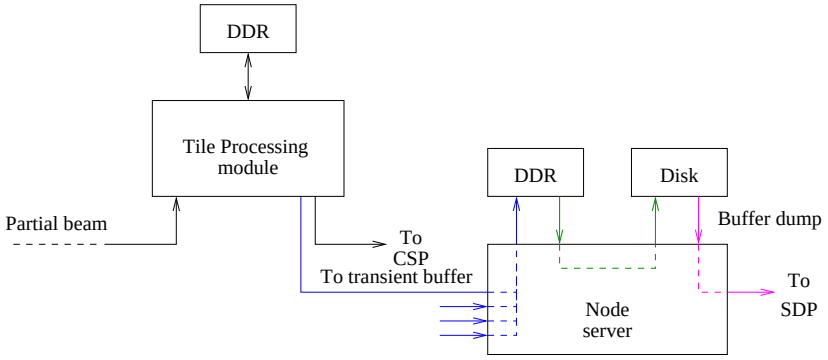

Figure 2: Transient buffer implementation using DDR memory in external servers

The bottleneck is the amount of required RAM. A server with 1024 GB of RAM could host 6 stations, with a corresponding memory bandwidth of 1066 Mbytes/s, i.e. less than 8% of a moderately fast DDR module. The aggregate net bandwidth is about 8.6 Gbps. A server with these capabilities costs a few hundred euros, plus 8-10 kEuro for the memory. A total of 86 such computing nodes are required. The MCCS computer cluster (128 nodes) could be used for the purpose, with 4 stations (640 GB of extra memory) per node. In this case only the memory modules should be purchased.

After detection of a transient event, the memory content can be dumped on disk. Giving current disk writing speed, this requires roughly 2-3 times the capture time, i.e. about an hour. Time can be drastically reduced using large RAID arrays, using faster SSD disks, or both. Once data have been transferred to disk, transient capture will resume, while the collected data on disk can be transferred to SDP using a relatively slow link, in a time constrained only on the expected average number of transients per day (e.g. 10 hours for an average of 2 events per day).

3.3 Disk based storage

A variant of this approach is to use RAM memory only as a temporary buffer, and store data directly on disk. This is limited by the available disk transfer speed, currently around 120 MB/s. A large RAID disk array (about 2 disks/station) would be needed to meet the required bandwidth. The disk capacity is not an issue. At least two whole buffers per station are required, but the minimum disk capacity is currently around 1 TB, i.e. 12 times the buffer size for a RAID cluster with 2 disks.

The structure is the same shown in figure 2, only the RAM size and data transfer speeds change.

RAM speed should be sufficient to allow both buffer write and buffer read at the same time. It should be noted that the same buffer can be written to memory multiple times (e.g. from Ethernet temporary buffer to a temporary buffer, to a disk buffer in the operating system and finally to disk), so an optimal transfer strategy must be found. Data transfer should not be impended by normal server activity. Also the disk bandwidth and disk transfer size must be sufficient to allow for simultaneous high speed disk write and slower disk read to transfer the data to the SDP.

An advantage of this approach is that the recover time after an event could be drastically reduced, or eliminated, providing that the dump to SDP operation be interleaved in such a way not to significantly reduce the disk write rate.

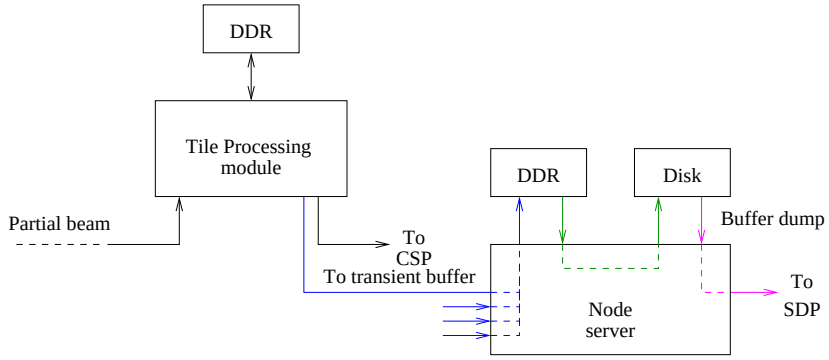

Figure 3: Transient buffer implementation using data storage on disk

An alternative to this approach is to use the existing hardware adopted for VLBI data storage. This hardware interfaces to the receiver (in our case the TPM generating the beamformed data) using high speed links, and uses large packs of COTS disks to store the sampled data. The Mark6 equipment [2], based on commercial COTS hardware, stores up to 16 Gbps of streamed data on a pack of 32 disks. The system is being currently tested to extend the recording speed to 32 Gbps. With 16 Gbps a single unit could record data for 8-10 stations. The unit is however quite bulky, using a standard 6U cabinet plus roughly other 6U of rack space for 4 disk packs of 8 disks each. The amount of rack space required would then be significant.

Some adaptation of the firmware and software would be required to comply with the different storing on trigger policy and playback feature. This has the obvious advantage of relying on existing, well tested, equipment, but may prove to be too expensive in terms of extra hardware and rack space. In any case the experience gained by the Mark6 programming team would be precious in developing the transient buffer software.

3.4 Mixed disk and RAM solution

A SSD device has a much higher bandwidth, for contiguous large files, currently exceeding 400 MB/s. A solution with a single SSD disk per station would therefore provide enough bandwidth to store the incoming samples without Disk capacity is much smaller, with fastest disks currently having a capacity of 256 GB, that could be sufficient for storing about 1400 seconds of beam samples.

SSD however have a very low endurance (number of writes per memory cell before the cell fails), that are in the order of 1000 writes per cell. A circular buffer can then be written just 1000 times, i.e. about ten days, before wear out occurs. High reliability SSDs exist, but the improvement is at most a factor of 10 at a large increase in cost. Increasing the SSD capacity will limit the number of write accesses to each cell, but in general reduce the endurance per cell. So, in the best case, we should replace 512 SSD modules every few months of continuous usage. SSD could be a viable option only if the total time spent in single pulse search is a small fraction of the total.

SSD can however be used as a transient buffer. With good quality SSD modules, and 2 events per day, several years of operation are expected before SSD wear out. A possible solution would be then to use DDR memory to store a time interval of the order of a few ten seconds (20-60 s), sufficient to accommodate the triggering delay ad a reasonable pre-trigger interval and, after a trigger has been detected, to dump the data on DDR directly on the SSD drive. The DDR then acts as a FIFO buffer, and the SSD is used only when an actual event has been detected.

4 Comparison of the proposed solutions

4.1 Storage costs

In all these solutions the dominant cost is linked to the huge amount of required storage. This amounts to a bare minimum of 160 Gbyte per station, 80 terabyte total. Te current retail cost for DDR chips and modules range from 8 (chip) to 9 (computer DIMM module) euro per Gbyte. Assuming a 30% saving for the high volume, this ranges to approx. 500 kEuro, or something less if the TPM distributed solution is adopted.

A large saving could be possible if disk storage is used. In this case the capacity is irrelevant (it would be larger than required), and the critical factor is the disk write bandwidth. At least 2 disks per station, 1024 total, would be required. A good quality, high speed standard disk costs around 80-100 euro, so the cost cap would be around 100 kEuro. SSD devices current performances would allow a single SSD disk per station to be used, with a total cost around 70 kEuro (140 euro per 256K SSD). Due to limited endurance, SSD can be used however only for temporary storage.

Costs can be reduced if the buffer is required only for the stations within the PSS beamforming core. This would be the case in a situation where PSS operations are performed using a sub-array with these stations, while the other stations will be used for other purposes. It is however conceivable that PSS operations will be conducted in commensality with other observations, and that the whole array will point in the PSS search area.

4.2 Network cost

An increase in bandwidth from the site to the SDP would be required. To support the TPM solution, at least 100 Gbps would be required. To download 160 GB per station, at 100 Gbps, about 6600 seconds would be required. This limits the time between two detectable pulses to something less than 2 hours.

In the RAM solution, the time required to transfer the RAM buffer to disk would be comparable to the capture time. Using SSD this time can be even shorter by a factor of 2–3, allowing for almost continuous transient capture. Network bandwidth is dependent only on the assumed average number of events per day. If about 14 events per day are expected, no bandwidth reduction is possible. If the number of expected events is smaller, of the order of a few, and/or a shorter event duration is acceptable, then 15 Gbps would be sufficient. Requirement 3083 also limits the required band to about 15 Gbps.

4.3 Software/firmware cost

In the TPM solution, significant modification of the current firmware is required. An estimate of the required work is about 0.5-1.0 man-year. Some work (0.1-0.2 man-year) is required for the 2 bit quantization and packet formatting for the samples to be sent to the transient buffer.

In the disk buffer solution, significant software must be written to optimize memory and disk transfers. It would be advantageous to leverage on the experience of the VLBI Mark6 data acquisition system.

The control software to integrate the buffer in the TM and in the elements LMC is similar for all the proposed solutions.

Type of buffer	TPM	RAM	Disk	Hybrid
Memory	500	450	40	50
Disk	–	70	100	140
FW/SW development	140	30	100	100
HW/rack space	–	–	70	–
Total	640	550	310	290
Risk	high	low	medium	low

Table 1: Comparison of costs (kEuro) for the different solutions

4.4 Risk costs

The TPM solution appears the more prone to risks. The added functionality interferes with the normal beamforming data paths, and arbitrage must be added to memory and Ethernet interface to prevent this. A nontrivial state machine must be designed to manage the buffer storage and download operations. These functionalities are added on top of an already saturated FPGA (the smallest one that fits the design), and timing closure of the design could be compromised, requiring a larger, and more expensive, component.

Using the existing MCCS computers could also cause interferences with the calibration task. In this case an independent cluster would be required, with a likely cost around 150 kEuro and one-two racks of floor space.

Disk storage has to be proven. This is a relatively easy task, as it has to be proven on a single node, with a simplified environment.

SSD endurance values for different SSD devices must be explicitly measured if SSD storage is assumed to be used.

Additional rack space is not requested in solutions that use existing hardware. Using standard disk storage in MCCS can increase the rack occupancy for each node, thus increasing the total number of racks. A very rough estimate (wild guess) of 0.5 euro per node has been made to account for this risk.

Both memory cost, memory size (thus number of required nodes) and, to a lesser extent, disk transfer speed is expected to improve in the future. So a cost mitigation option would be to limit the performances of the buffer (e.g buffer only the inner stations, decrease the buffered bandwidth, decrease the post transient time) and upgrade/replace the hardware at a later time.

4.5 Comparison result

An estimate of the relative costs for the proposed solutions is shown in table 1. Only costs that are different between these solutions has been considered, i.e. costs for the extra bandwidth between LFAA/CSP and SDT, and for modifications in the TM/LMC software have not been considered.

Considering the very large uncertainties in these figures, costs appear comparable. Risk factors are likely the dominant cost factors, and should be evaluated more accurately.

References

[1] Zsolt Kerekes: "SSD endurance myths and legends"
http://www.storagesearch.com/ssdmyths-endurance.html

[2] Roger Cappallo, Chet Ruszczyk, Alan Whitney: "Mark6: Design and Status"
http://www.haystack.mit.edu/ftech/vlbi/mark6/mark6_memo/05-Mark6_Design_and_Status.pdf

List of acronyms

ADC: Analog to Digital Converter

COTS: Commercial Off The Shelf

CSP: Central Signal Processor

DDR: Double Data Rate: Implementations of DRAM using both clock edges for data transfer. DDR3 and DDR4 versions of the standard are used in the design

DSP: Digital Signal Processing

FPGA: Field Programmable Gate Array

FW: Firmware

GBE: Giga Bit Ethernet

HDL: High Level Design Language

I/O: Input/Output

LFAA: Low Frequency Aperture Array Element or Consortium

LMC: Local Monitor and Control

MATLAB: MATLAB simulation language and application

MCCS: Monitor, Control and Calibration Servers

M&C: Monitor and Control

PSS: Pulsar Search

RAM: Random Access Memory

RFI: Radio Frequency Interference

RS: Requirement Specification

SDP: Science Data Processing

SDRAM: Syncronous Dynamic Random Access Memory: Standard for bursting, fast memory. DDR3 and DDR4 implementations of SDRAM are used in the design

SKA: Square Kilometre Array

SKAO: SKA Organization (or office)

SPEAD: Streaming Protocol for Exchanging Astronomical Data

SSD: Solid State Disk

SW: Software

TBC: To be confirmed

TBD: To be decided

TM: Telescope Manager

TPM: Tile Processing Module

VLBI: Very Large Baseline Interferometry

WBS: Work Breakdown Structure

Contents

List of Tables

List of Figures