Mark 5C Specification

(V1.0 19 Feb 2008)

MIT Haystack Observatory
Westford, MA 01886

National Radio Astronomy Observatory
Socorro, NM 87801

1. Introduction

The Mark 5C is being designed as the next-generation Mark 5 system, with a capability of recording sustained data rates to 4096 Mbps. It will use the same disk modules as the Mark 5A and Mark 5B, thus preserving existing investments in disk modules.

The data interface for both recording and playback will be 10 Gigabit Ethernet, which is rapidly becoming a widely supported standard. The use of 10GigE interfaces comes with some significant implications, however. Firstly, data sources must be designed to provide data streams in a format compatible with the Mark 5C requirements. And secondly, data playback through a 10GigE interface is a good match for a rising generation of software correlators. In the interests of backwards compatibility, the Mark 5C will also support a mode which writes disk modules in Mark 5B data format which can be correlated on existing Mark 4 correlators that support Mark 5B.

This document is intended as a specification for the Mark 5C, which will be implemented using the existing Amazon StreamStor disk-interface card (used in the Mark 5B+) from Conduant Corp along with a new 10GigE-specific interface daughterboard being designed by Conduant. Unlike the Mark 5A and Mark 5B, no separate specialized I/O card will be necessary in the Mark 5C.

2. Data Sources

One major implication of the Mark 5C model is that the Digital Data Source (DDS) is responsible for all data time-tagging, formatting and creation of Ethernet packets. This is a departure from the VSI-H model used by the Mark 5B, which has basically only 32 parallel sample bit-streams, a clock, and 1pps tick flowing between the data source and the Mark 5B, with the Mark 5B being responsible for creating data frames with higher level time-tagging and formatting.

Fortunately, VLBI DDSs capable of creating such formatted Ethernet packets are now being developed in both the U.S. and Europe as part of the development of digital downconverters and backends. Suitable 10GigE DDSs to drive the Mark 5C are expected to be available sometime in the first half of 2008. The details of the data formats to be provided to the Mark 5C are specified in a separate document “Mark 5C Data-Frame Specification”. Normally, each Ethernet packet from the DDS will contain sample data from only a single frequency channel, although a Mark 5B-compatible data mode is specified which will write disk modules in a format that can be re-played on a Mark 5B playback unit; this will provide the ability to process the recorded data on existing Mark 4 hardware correlators.

3. Correlators

A major shift is currently developing to move from hardware-based correlators to software-based correlators, some of which already exist. Unlike the Mark 5A and Mark 5B, the Mark 5C will have no streaming hardware playback interface. Instead, the data files will appear to the user as standard Linux files and will be read as such. We expect that the standard interface for playback

to a correlator will be through a standard 10GigE interface implemented on a commercial NIC. Unlike existing hardware correlators, software correlators do not demand constant-rate streaming inputs. As such, the Mark 5C playback is well suited for interfacing to software correlators, but not well suited or intended to interface to hardware correlators.

4. General Mark 5C Characteristics

The Mark 5C will have the following characteristics:

- Mark 5C will be fully compatible with all existing Mark 5 disk modules, however some older modules may limit record/playback data rates.
- At data rates above about 2 Gbps, it will be necessary to record to two 8-disk modules simultaneously, in so-called 'non-bank' mode, which is not normally used by Mark 5A or Mark 5B/5B+.
- A 10GigE interface for *receive only* will be implemented on the Amazon StreamStor disk-interface card (currently used in the Mark 5B+) by replacing the FPDP I/O daughterboard on the Amazon card with a newly designed 10GigE daughterboard. This 10GigE interface will be receive-only, optimized for sustained real-time recording of *at least* 4096 Mbps from a DDS. Received Ethernet packets can be OSI Layer 2 or higher, but will only be processed by the Mark 5C at the Layer 2 level. Jumbo Ethernet packets up to 9000 bytes will be supported. The DDS is required only to transmit Ethernet packets, and is not required to process any received packets.
- The entire data payload from each arriving Ethernet packet, sans a specified length of payload header (which may contain higher OSI Layer parameters or other information), will be recorded to disk. In this sense, the Mark 5C is entirely 'formatless'; i.e. all data formatting must be done by the data source. This allows each user to format the recorded data according to his/her needs.
- The Ethernet data payload may contain a user-generated 32-bit "Packet Sequence Number" (PSN), whose position within the data payload can be specified to the Mark 5C. The Mark 5C can be commanded to a "PSN monitor mode" that will parse this serial number from every packet to identify missing or out-of-order packets. Out-of-order packets, within some reasonable limits, will be restored to proper order, while the user data from each missing packet will be replaced by user-specified "fill-pattern" data. The MSB of the PSN may also be used as an 'invalid' marker to prevent recording data from a packet. If "PSN monitor mode" is disabled, data are recorded to disk in the order that packets are received; no checks are made for out-or-order or missing packets.
- Similar to the Mark 5A/B, the Mark 5C will record data as "scans", where a scan is defined as the period between starting and ending the recording of a particular observation. The duration of a scan may be from several seconds to many minutes. The host application software will maintain a directory of scans for easy identification and access. No duplicate-named scans are allowed.
- All control of Ethernet link setup and recording parameters of the Mark 5C will be setup using Streamstor XLR calls.
- Scans will appear as normal Linux files to the host PC. Data playback on the Mark 5C will be through a 10GigE NIC interface on the host PC . A planned upgrade by Conduant

of the Amazon card, which interfaces to the PCI-X bus, will support the PCI-e bus to allow substantially higher playback rates from Mark 5 disk modules. A sustained average playback rate of at least 3Gbps is needed to ensure that data processing keeps up with data taking on dedicated VLBI arrays such as the VLBA.

5. Physical connections

5.1 Connection Architectures

The normal connection for receipt of data by the Mark 5C, illustrated in Figure 1a, is a one-to-one connection between a DDS and the Mark 5C. This type of connection guarantees proper packet ordering.

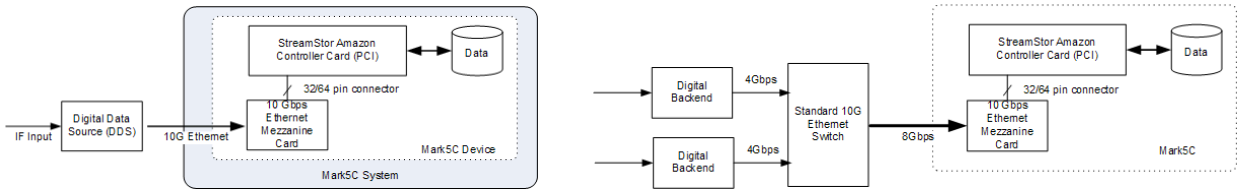

Figure 1a): One-to-one DDS-to-Mark5C connection, 1b) Example of several-to-one connection

Multiple DDSs, shown in Figure 1b as Digital Backends, may be connected to a single Mark 5C through a standard 10GigE Ethernet Switch. In this configuration, there is no guarantee of packet order arriving at the Mark 5C.

Conversely, a single DDS may spread data to multiple Mark 5Cs through an Ethernet switch, as shown in Figure 2. Obviously, in this case, packet order is guaranteed.

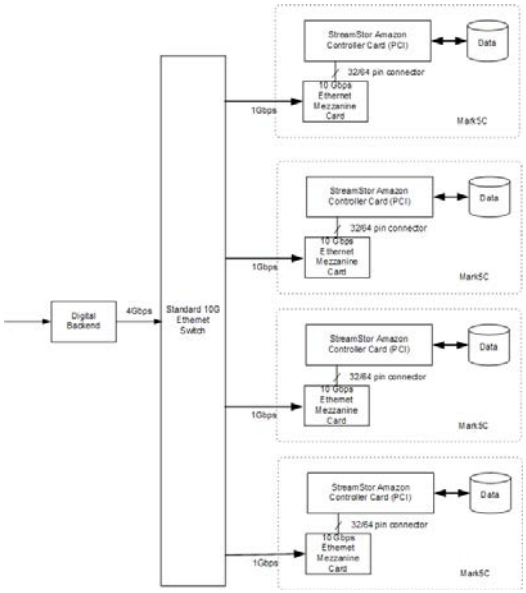

Figure 2: Example of one DDS connected to several Mark 5C systems

5.2 Physical Ethernet Connector

The physical Ethernet connector on the Mark 5C shall be compatible with a standard 'CX4' connector (10GigE copper).

6. Ethernet packet structure

Packet structures will adhere to standard IEEE 802.3 Ethernet packet specification or the Ethernet jumbo-packet packet structure[1]. The general structure of an Ethernet packet is shown in Figure 3.

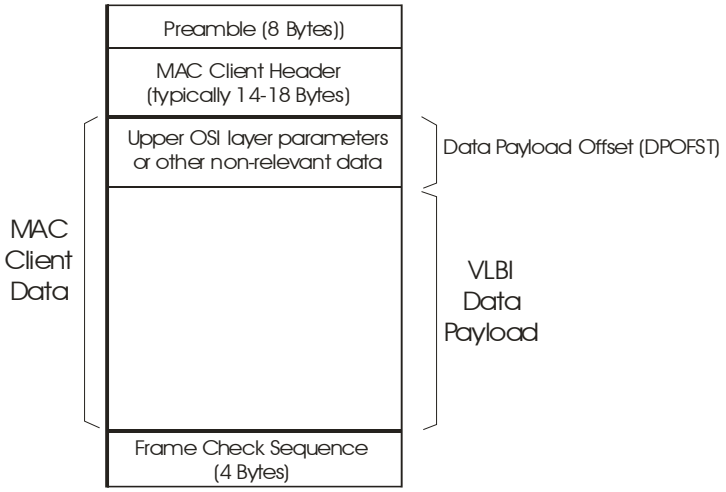

Figure 3: Ethernet packet format

The components of an Ethernet packet include:

Preamble – a synchronization pattern that allows the Ethernet hardware to synchronize properly to the Ethernet packet

MAC Client Header – contains such information as Source and Destination MAC addresses, type/length, and tag control information

MAC Client Data – this is the data that is delivered by the Ethernet receiver to the user

Upper OSI layer parameters – such parameters as OSI Layer 3 IP and higher-layer parameters may be contained here; they are assumed to be of fixed length for a given scan.

VLBI Data Payload –the VLBI Data Payload will be recorded to disk; the user must specify the byte offset from the beginning of the MAC Client Data at which recording is to start (dubbed 'Data Payload Offset' or 'DPOFST')

Frame Check Sequence (FCS) – a cyclic redundancy error-checking code (not delivered to the user)

Within each scan recorded to a Mark 5C, all Ethernet packets will have exactly the same length and structure.

[1] The Ethernet 'jumbo' packet is bigger than the standard Ethernet packet size, which is 1518 bytes (including the Layer 2 header and FCS) The definition of packet size is vendor-dependent, as these are not part of the IEEE standard. IEEE 802.1Q-2005 QTag fields are used to support jumbo packets.

7. VLBI Data Payload

The VLBI Data Payload may optionally contain a 4-byte DDS-generated Packet Sequence Number (PSN) that may be used by the Mark 5C to monitor packet ordering and replace the data from missing or bad packets with 'fill-pattern' data, or to prevent recording of packets marked as having 'invalid' data. Normally the PSN, if it exists, will be located at the beginning of the VLBI Data Payload, as shown in Figure 4, though it may be placed on any 4-byte boundary within the VLBI Data Payload. If the PSN is present, the user must specify its position as a byte offset (dubbed 'PSNOFST') from the beginning of the VLBI Data Payload (PSNOFST=0 as shown in Figure 4). The user may also specify an offset from the beginning of the VLBI Data Payload (dubbed 'DFOFST') to the first byte of the 'Data Frame' to be recorded to disk (DFOFST=4 as shown in Figure 4); all bytes from the beginning to the end of the Data Frame will be recorded[2]. For compatibility with the StreamStor bookkeeping, the length of a Data Frame must be a multiple of 64 bits. The VLBI Data Payload is always in *little endian* format. The endian format of OSI-layer-specific fields will conform to established Internet protocols (i.e. big endian).

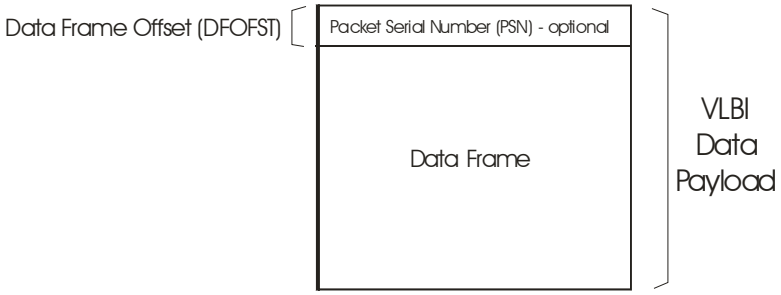

Figure 4: Structure of VLBI Data Payload (shown with PSNOFST=0, DFOFST=4)

Figure 5 shows an example of a VLBI Data Payload in which the PSN is embedded within the Data Frame. In this case, the PSN will be recorded as part of the Data Frame.

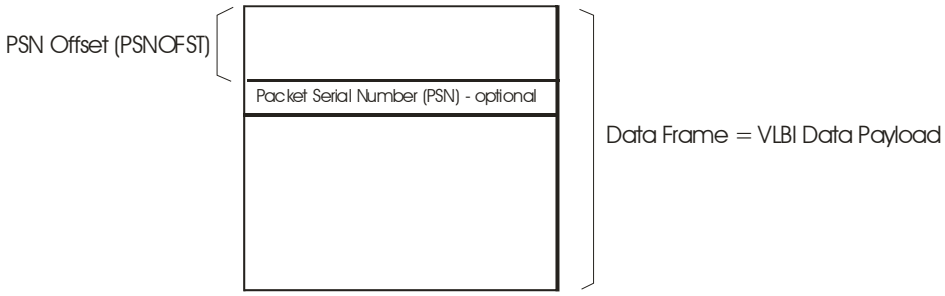

Figure 5: VLBI Data Payload coincident with VLBI Data Payload;
PSN embedded within Data Frame (PSNOFST>0, DFOFST=0).

[2] It is assumed that there is no padding at the end of the Data Frame since padding is normally necessary only for packets with MAC Client Data fields of length ~46 bytes, which we do not expect to use.

8. Packet Checking and Monitoring

8.1 Use of Packet Sequence Number

The optional Packet Sequence Number (PSN) may be used for one of two purposes:

1. Monitor for missing or bad packets and replace the Data Frame of such packets with a user-defined 'fill-pattern' ('PSN-monitor' mode 1).
2. Mark a packet as invalid to prevent it from being recorded ('PSN-monitor' mode 2).

These two modes are mutually exclusive and may not be used together.

8.1.1 PSN-monitor mode 1

In order for the Mark 5C to use the PSN to monitor for missing or bad packets, the DDS must guarantee that the PSNs arriving at the Mark 5C are properly generated, i.e. increment by one for every packet sent[3]. The 32-bit PSN can start at any arbitrary number on initiation of recording, and will rollover to zero on overflow.

If PSN-monitor mode 1 is enabled, the Mark 5C will do the following:

1. Decode and read the PSN on each arriving error-free packet.
2. A ring buffer of modest size (>=16 packets) will be maintained to manage the writing of packets to disk.
3. If a packet is identified as bad (as indicated by a bad FCS word) or missing, a replacement fill-pattern Data Frame of the correct length will be written as necessary to maintain the proper sequencing and completeness of Data Frames.

8.1.2 PSN-monitor mode 2

If PSN-monitor mode 2 is enabled, any packet with the most-significant bit set to one will be interpreted as 'invalid' and its Data Frame *will not be recorded* to disk.

Note: Though Figure 4 shows the PSN at the beginning of the VLBI Data Payload, the PSN can be placed anywhere within the VLBI Data Payload. If the specified byte offset to the PSN is greater than the byte offset to the Data Frame, the PSN will be recorded as part of the Data Frame.

8.2 Packet length checking

As a guard against recording spurious packets which might be received (particularly over switched networks) during a scan, the Mark 5C will include an option to check each received packet against a user-specified MAC-Client-Data length and record only those packets matching the expected length. If this option is enabled, packets of incorrect length will not be recorded. If packet-length checking is disabled, the Mark 5C will record all packets with a MAC-Client-Data length up to and including a user-specified value.

[3] This condition can usually be guaranteed only in a one-to-one DDS-to-Mark 5C connection.

9. Setup

The following parameters will be sent to the Mark 5C (presumably through standard StreamStor function calls) before accepting data from a DDS:

1. List of MAC source addresses (16 max) to accept; or set to 'promiscuous' mode to accept all.
2. Length of MAC Client Data (to be checked against received packet if packet-length check is enabled); if packet-length checking is disabled, this parameter corresponds to the maximum length of MAC Client Data (aka MTU) that will be accepted.
3. Byte offset to beginning VLBI Data Payload (DPOFST).
4. PSN monitoring mode: mode 0 (no PSN monitoring), 1, or 2. If mode is 1 or 2, value of PSNOFST.
5. Byte offset to beginning of Data Frame (DFOFST).
6. Byte length of Data Frame to be recorded (must be multiple of 64 bits).
7. 32-bit fill-pattern

All setup parameters will remain fixed for the duration of a scan[4]. The Mark 5C must be able to report to the user the MAC address of its 10GigE data port.

10. Monitoring

The following monitor parameters will be available to the user while recording is in progress:

1. Total number of packets received since start of recording.
2. Number of packets with FCS errors.
3. If packet-length monitoring is active: number of packet-length errors.
4. If PSN-monitor mode 1 is active: number of missing or bad packets (including packets with errors); should be same as number of fill-pattern Data Frames written to disk.
5. If PSN-monitor mode 2 is active: number of invalidated packets.

These monitor parameters are in addition to the normal StreamStor monitor functions.

The data stream from DDS may exist before the Mark 5C is setup and recording. In this case, the Mark 5C will simply start accepting data as soon as setup is complete and a recording command is issued. Similarly, the data stream from the DDS may continue after recording is stopped.

Conversely, the DDS may be quiet for a period after recording is started, or may become quiet before the recording is stopped. Additionally, packets may appear in bursts, with gaps between bursts (pulsar recording). These conditions should not be cause for alarm. The user may request monitor parameters to ascertain the status of the recording.

[4] The astute reader will note that only two byte-offset values really need to be provided to the Mark 5C hardware, namely DPOFST+PSNOFST and DPOFST+DFOFST.

11. Playback

Playback will normally be managed through a software wrapper which makes the StreamStor data files appear to the host PC to be standard Linux read-only files. Playback data rate will depend on the combination of host-hardware capabilities, as well as the limitations of the PCI bus. Conduant is planning to replace the Amazon StreamStor board with a version which uses the much faster PCI-e serial bus, but the timescale is not certain.

12. Compatibility

Recordings made with the Mark 5C must be readable with the XF2 StreamStor board as well as the Amazon board.

13. Bad disk management

8.1 Recording

In order that data not be lost when one or more 'slow' disks are part of the ensemble of data disks, the StreamStor system must, when necessary, divert data to other disks to prevent data loss or interruption.

As part of the bad-disk management strategy, the module directory must be redundantly recorded on all disks so that the loss of any one or more disks will not affect the ability to recover the directory. [Note: This redundant directory structure may, we believe, already be implemented.]

8.2 Playback

In the case of missing disks on playback (usually when a bad disk has been removed), the Mark 5C must be able to recover data on the remaining good disks. An option should be provided to write a user-specified 32-bit fill pattern in place of missing data, as this will be required by some correlators. For playback to software-based correlators, it will not normally be necessary to supply fill pattern Data Frames in place of missing data, but an indication must be provided to indicate to the user that requested Data Frames are unavailable.

14. VSN label management

At data rates above ~2 Gbps, it will be necessary to record to two disk modules (16 disks) simultaneously in so-called 'non-bank' mode. While operating in non-bank mode, it is essential that both modules maintain their individually assigned VSN labels and that the VSN labels of both modules be accessible to the user.[5]

15. Data-capacity goal

A long-term goal of the Mark 5C system is to support 4Gbps recording continuously for 24 hours with no operator intervention, which will require ~44 TB of accessible disk storage. This might be accomplished with sixteen disks of ~2.75TB capacity each, or by allowing hardware extensions to support additional disk modules.

[5] Normally, VSNs are assigned in 'bank' mode. It would be satisfactory to require that the modules be placed in 'bank' mode when reading the individual VSNs. However, placing the modules into 'non-bank' mode or erasing the data in 'non-bank' mode must not modify the VSN assigned to either module.

Acronyms

10GigE- 10 Gigabit Ethernet
DDS - Digital Data Source
FPDP - Front Panel Data Port (data interface to StreamStor card used by Mk5A/B/B+)
MAC - Media Access Control (lower sublayer of the OSI data-link layer protocol)
MSB - Most Significant Bit
MTU - Maximum Transmission Unit (maximum length of MAC Client Data)
NIC - Network Interface Card
OSI - Open Systems Interconnect (reference model for network communications)
PPS - Pulse per second
PSN - Packet Sequence Number