ngVLA Electronics Memo #16
Trade Study: CSP Internal Data Interchange Format

Nolan Denman

NRAO CDL

1. Overview

Data transport within the ngVLA Central Signal Processor (CSP) occurs at an extremely high rate, driving the CSP Switched Fabric (CSF) technical requirements; this increases the importance of efficient data transportation. Figure 1 and Table 1 provide an overview of the main data flows within the ngVLA CSP - between the antennas' Digital Back-Ends (DBE), the Subband Processors (SBP), the Pulsar Engine (PSE), and the CSP Back-End interface of the Online Computing and Software subsystem (ONL).

The internal networking has been explicitly assumed to use Ethernet links over commercial hardware, with IPv6 and UDP as the Network and Transport layer protocols respectively. The question is what the internal data representation should look like; particularly, if there are reasons to adopt any specific pre-existing format.

For fundamental efficiency reasons, only formats which permit the direct encoding of binary data are considered; this disqualifies base64-encoded formats such as JSON and YAML. Additionally, the Hierarchical Data Format version 5 (HDF5) was excluded due to high overhead; it offers many functions for archival use but is not particularly suited for data transport. Sections 2 through 6 describe the primary options under consideration, with summary discussion in Section 7.

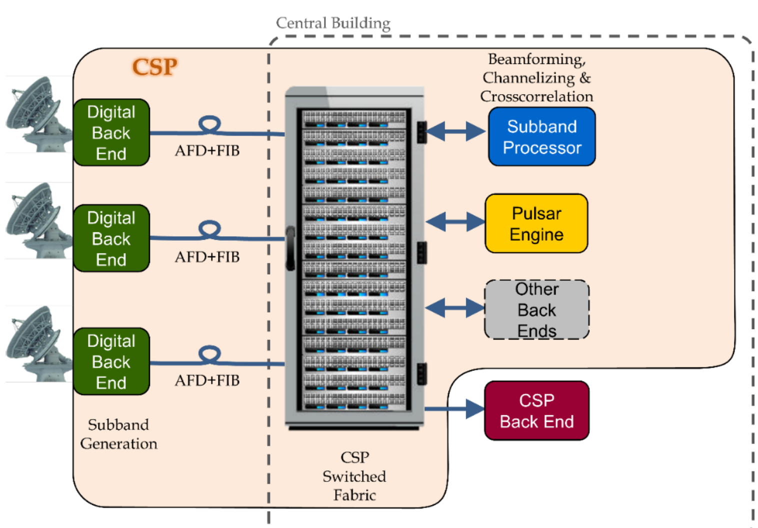

Fig. 1.— Data flows within the ngVLA CSP.

Source	Destination	Data Description	Notes
DBE	SBP	Digitized Voltages	
SBP	PSE	Channelized and Beamformed Data	For further transient processing
SBP	ONL	Channelized and Beamformed Data	For external VLBI
SBP	ONL	Time-Integrated Visibilities	
PSE	ONL	Pulsar Timing Data	Phase-aligned, possibly de-dispersed
PSE	ONL	Pulsar Search Data	Possibly de-dispersed

Table 1: A summary of data varieties

2. FITS and FITS-Derived

The Flexible Image Transport System (FITS, IAU FITS Working Group (2016)) standard is commonly used in astronomy applications to support images, multi-dimensional data arrays, tables, and key/value pairs. A subset of valid FITS files meet the additional PSRFITS standard (Hotan et al. 2004) specifically intended for use with time-domain pulsar observation data. PSRFITS has been specified as the output format for the ngVLA CSP Pulsar Engine (PSE).

FITS files consist of a primary header, optional primary data array, and zero or more extensions each with their own header blocks. All headers are expressed as sequences of 80-character ASCII-text strings (byte values `0x20` through `0x7E`), and both header and data blocks are constructed in units of 2880 bytes.

The broad adoption of FITS and the large number of existing keyword and extension conventions make it an essential archival format for astronomical data, but the inefficiency of the header encoding and the requirement of 2880-byte-aligned data blocks significantly reduces its potential as a choice for transportation.

3. VDIF

The VLBI Data Interchange Format (VDIF, VDIF Task Force (2014)) is a broadly-supported format for the storage and transportation of channelized voltage data for radio astronomy, particularly VLBI.

The VDIF packet header consists of 32 bytes of data, including 15 bytes of extended user data which may be allocated according to a customized Extended Data Version. VDIF makes certain assumptions about the data it contains; the number of channels in a frame must be a power of two and the data packing must occur in a very specific structure - which is highly inefficient for data sizes which do not evenly divide 32 bits. The way in which VDIF stores timing information is also idiosyncratic and has trouble with uniqueness beginning in 2032.

4. VITA 49

The Versa Module Eurocard Bus (VMEBus) International Trade Association number 49 (VITA 49, VITA (2017)) is a family of proprietary standards for digital transmission of Software-Defined Radio (SDR) data and metadata.

Formally, the ANSI/VITA 49.0-2015 VITA Radio Transport (VRT) standard (“V49.0”) describes packet structure and function conventions for time-domain radio-frequency data, specific ‘context’ metadata, additional ‘extension’ data, and arbitrary ‘extension context’ metadata. ANSI/VITA 49.1-2015 defines the VITA Radio Link Layer (VRL) standard, an encapsulation protocol which is not germane to this discussion.

The ANSI/VITA 49.2-2017 VRT Standard (“V49.2”) defines a protocol to express and transmit spectrum observation, spectrum operations, and RF device capabilities; this extends and broadly replaces the V49.0 standard. A number of varieties of data type are predefined, including time- and spectral-domain signal data and specific categories of metadata. The ‘extension’ categories are retained for any data which may not be conveyed through the pre-defined methods. As in V49.0, signal data is the primary use case, although spectral-domain data is now included in this category. A significant portion of the standard concerns parallel transmission of monitoring and control signals in the same connection as the data stream, which is not applicable in the ngVLA context.

5. SPEAD

The Streaming Protocol for Exchanging Astronomical Data (SPEAD, Manley et al. (2012)) includes an application-layer packetized data format designed for lightweight and flexible transfer of data over UDP. The protocol centers around a user-defined data structure - a ‘heap’ of variables - with packets serving to propagate changes to the value of these variables. Each packet consists of a structured header which describes the subsequent data, as well as a set of required and optional identifiers which allow interpretation of the subsequent ‘payload’ of binary data.

The SPEAD system is extremely flexible and powerful, but many of the features would not be required for the ngVLA. As mentioned in Manley (2014), practical considerations lead many implementations to avoid the self-defining and re-structuring features of the protocol in favor of a fixed structure.

The minimum compliant SPEAD packet, for the ‘standard’ 64-40 bit flavor, adds 32 bytes to the data payload (plus 3-8 bytes per item). Some of the parameters included (packet length, counters and timestamps, etc.) would likely be replicated in a custom format (see Section 6), but some of this overhead is not strictly required for application to ngVLA.

Philosophically, SPEAD is intended to propagate changes in the ‘heap’ of variables from one or more sources to a host which maintains a current state; the ngVLA implementation of SPEAD as a packet-level protocol would not involve adopting this structure. Additionally, the use of SPEAD without an implementation of the self-re-defining feature set or predefinition of certain items would not automatically provide any degree of interoperability with other telescopes using SPEAD - or any simplification of the implementation process.

6. Custom Binary

In addition to the above-specified formats, it is possible to define a new packet format or family of formats which can be applied to the ngVLA CSP internal data. This could ‘mix and match’ desired features in any combination; the primary drawbacks are that it requires development effort and that any externally-developed custom back-ends would be wholly reliant on ngVLA documentation for their interface specifications.

7. Discussion

Categorizing FITS along with HDF5 as more relevant to archiving than internal data transport, the primary options are VDIF, V49.2, SPEAD, and a custom format.

VDIF formatting is mandated for VLBI output from the CSP (requirement CSP0035). Internal discussions suggest that VDIF is disfavored for non-VLBI data due to a poor fit with the CSP requirements; the specificity of its layout, which enhances its usefulness for VLBI processing, makes it less well-suited for conveying a wide variety of data. Attempting to extend VDIF in a way which would include all the relevant data would undermine the ability for external parties to interpret the data according to the standard, negating the interoperability which was its primary benefit.

V49.2 offers broad commercial and industrial compatibility, but is focused on immediate signal data transfer; many of the ngVLA CSP use cases would be entirely custom-defined within the 'extension' space provided. As such, it is an excellent fit for DBE-SBP communication, but a high degree of customization would be required to make V49.2 useful to the other communication paths in the CSP.

The implementation of an ngVLA CSP flavor of SPEAD is in many ways similar to the definition of a custom format. The bit widths of the item pointers and item identifiers are selected, and item identifiers may be assigned to the data of interest. Packets may then be constructed with the minimal SPEAD header and any data payload which is desired; the interpretation and unpacking of these packets will be constructed as required for each system.

A custom packet format would differ from SPEAD primarily in that the data structure could be made implicit rather than self-describing, in which case the header and internal address data could be reduced in size to the bare minimum. Almost all of the requirements would be the same, and the implementation process very similar. There would be a small ($\sim 1\,\%$) gain in transmission efficiency; it is unclear what costs would be involved beyond the SPEAD case. Future expansion of the CSP with externally-developed hardware would require consultation of the specific implementation in either case; SPEAD would provide a mechanism to integrate an alternate packet structure but this could just as easily be incorporated into a custom format.

It appears that, no matter which data representation is selected, the essential development effort of defining and documenting data structures and ensuring their correct implementation is unchanged. Similarly, external back-ends would need information from ngVLA to properly interpret the data they receive in all but the most strictly defined of cases.

The output from the DBE, destined for the SBP and possibly some commensal back-ends, fits very well with the V49.2 standard and as such V49.2 has been tentatively selected as the DBE output data format. Output from the SBP which is intended for VLBI use will be in VDIF. At this time, there does not appear to be a material benefit to adopting a pre-existing format for the other data streams; an as-yet-undefined custom binary packet format remains the provisional selection.

Comments and questions would be appreciated by the author, and can be sent to `ndenman@nrao.edu`.

REFERENCES

Hotan, A. W., van Straten, W., & Manchester, R. N. 2004, Publications of the Astronomical Society of Australia, 21, 302

IAU FITS Working Group. 2016, Definition of the Flexible Image Transport System (FITS), Tech. rep., International Astronomical Union

Manley, J. 2014, PhD thesis

Manley, J., Welz, M., Parsons, A., & Ratcliffe, S. 2012, SPEAD: Streaming Protocol for Exchanging Astronomical Data, Tech. Rep. SSA4700-0000-001, SKA SA

VDIF Task Force. 2014, VLBI Data Interchange Format (VDIF) Specification 1.1.1, Tech. rep., International VLBI Service for Geodesy and Astrometry (IVS)

VITA. 2017, VITA Radio Transport (VRT) Standard for Electromagnetic Spectrum, Tech. Rep. 49.2-2017, ANSI/VITA

This preprint was prepared with the AAS LaTeX macros v5.0.