VLBA Sensitivity Upgrade Memo 52

Multi-bit sampling in VNDA

Walter Brisken, Mark Kettenis

12 Oct 2022

1 Introduction

VLBI systems have traditionally sampled voltage streams with 1 or 2 bits per sample as a means to optimize continuum observation signal-to-noise ratio at a given data rate. The VLBA currently only produces 2-bit quantized data with its ROACH Digital Back End (RDBE) system. The VLBA New Digital Architecture (VNDA) will replace the RDBEs and will offer 2, 4, and 8 bit per sample options. This memo explores choices that must be made and suggests a change in the location that quantization correction is performed.

2 Digitization

VLBI employs digitization of the voltage streams for storage, distribution, and correlation. Digitization consists of two separate processes:

Sampling Sampling is the act of making a series of voltage measurements at a particular time interval. The signal being sampled must have a bandwidth no greater than half the sample rate for real-valued sampling, or no greater than the sample rate for complex-valued sampling.

Quantization Quantization is the act of assigning digital representation to the measured value. Usually this involves thresholding to divide the voltage range into a finite number of intervals, each interval being assigned a particular digital code.

2.1 Digitization within VNDA

The VNDA system has two primary modules: the Producer and the Consumer. The Producer will entirely sample the four VLBA analog IFs and will quantize to 12 bits. These samples are expanded into 16-bit real-valued words and packaged as time-tagged and VDIF formatted. They are then distributed via a multicast-enabled high-speed network switch. The Consumer will take the samples from the Producer, channelize the data in a configuration requested by the user, and finally requantized to 2, 4, or 8 bit values. These are then packaged as complex-valued VDIF data. Finally the data are sent to a recorder or over a wide-area network to an off-site destination.

In most cases, VNDA data will be processed by the DiFX correlator which has support for all of the modes that VNDA will produce.

The VNDA Consumer will quantize using the following process:

1. The sample stream will be multiplied by a gain factor provided by monitor and control
2. Sample values exceeding limits will be clipped
3. An appropriate selection of bits will be retained

A software service loop will be established within the VNDA Monitor Interface Board (MIB) emulator that will determine an appropriate update to the gain value based on the switched power RMS values for each channel. This new gain value will be latched into the Consumer on a 1-second boundary.

3 Some mathematics of quantization

In the absence of Radio Frequency Interference (RFI) the voltage stream coming off an antenna is Gaussian distributed with mean of 0 and standard deviation of σ^2. The quantization process will take a series of samples with voltage values v_i where i is the sample index and will digitally encode them. When the values are to be used, the digital code is mapped to a reconstituted value $\hat{v}_i$. In the case of simple n-bit quantization, there are 2^n separate digital codes (each of the possible values of n bits). The reconstitute value will differ from the original by induced quantization noise, $\epsilon_i = \hat{v}_i - v_i$.

Generally the amount of introduced quantization noise is reduced when the number of quantization states is increased. This is quantified by the quantization efficiency[1], η_Q:

$$\eta_Q = \mathrm{corr}(v, \hat{v})^2 \equiv \frac{\langle v\hat{v}\rangle^2}{\langle v^2\rangle \langle \hat{v}^2\rangle} \tag{1}$$

Quantization efficiency is a quantitative assessment of the negative impacts of quantization on signal-to-noise ration. In order to make up for η_Q, one would need to observe for η_Q^{-2} as long.

3.1 1-bit quantization

1-bit quantization is the simplest possible scheme to use. A positive voltage sample, v, is recorded as a binary '1' and a negative value is recorded as a binary '0'. The decoder would reconstruct a digital voltage stream by assigning some positive value, $\hat{v} = \sqrt{2/\pi}\sigma$, to a '1' sample and $-\hat{v}$ to a '0' sample. It should be clear from its functional form that an absolute scaling of the reconstructed values has no impact on quantization noise. It is more usual to simply reconstruct the values as ± 1 as long as direct calculation of ϵ_i is not being performed. The magnitude of the reconstructed values is chosen to mathematically minimize the RMS quantization noise. The resultant digital stream has constant "power" and thus cannot convey total power information; amplitude scaling is generally determined through system temperature and aperture efficiency measurements. Figure 1 left panel shows the two regions of the voltage axis, the voltage probability function (Gaussian with RMS=σ), and the code ('0' or '1') that would be assigned to each voltage sample value. The true power must be computed based on out-of-band measurements such as T_{sys} which is generally determined on timescales much longer than the sample interval.

The quantization efficiency is exactly $2/\pi$ (approximately 63.6%) for 1-bit sampling, assuming a uncorrelated Gaussian-distributed voltage stream. Note in Figure 1 right panel that the distribution of quantization noise is non-Gaussian. As should be clear from symmetry, exactly half of the samples result in each of the two codes. The two sample states are described in the table below

Code	Voltage range	Reconstructed value	Frac.
0	$-\infty$ to 0	$-\sqrt{2/\pi}\,\sigma$	50%
1	0 to ∞	$\sqrt{2/\pi}\,\sigma$	50%

[1] See Thompson, Emerson, and Schwab, Radio Science 42, RS3022, 2007, available at https://safe.nrao.edu/wiki/pub/KPAF/KFPACorrelator/Quantizationlevel.pdf.

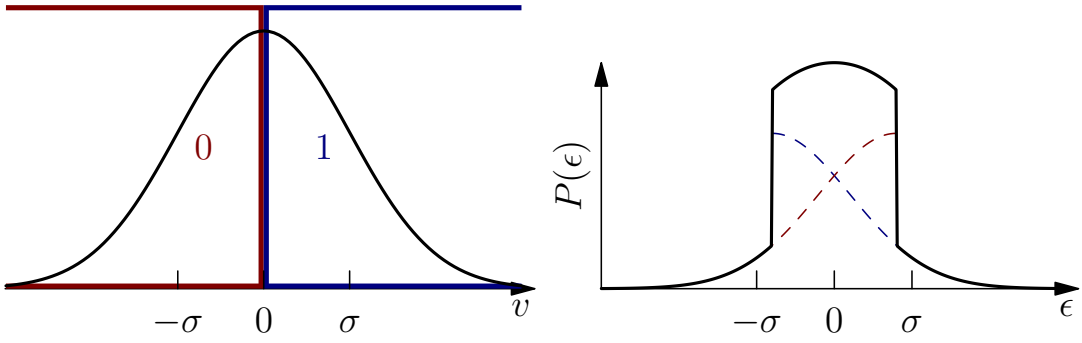

Figure 1: *1-bit quantization visualized.* (right) Voltage distribution, sample regions, and codes for 1-bit quantization. (left) quantization noise distribution. The red and blue dashed curves represent the contribution to quantization noise from the two sample states. See text for more details.

3.2 2-bit quantization

The case of two bits per sample is considerably more complicated than for one bit per sample. Figure 2 gives an overview of the distribution being sampled and the codes being assigned. After

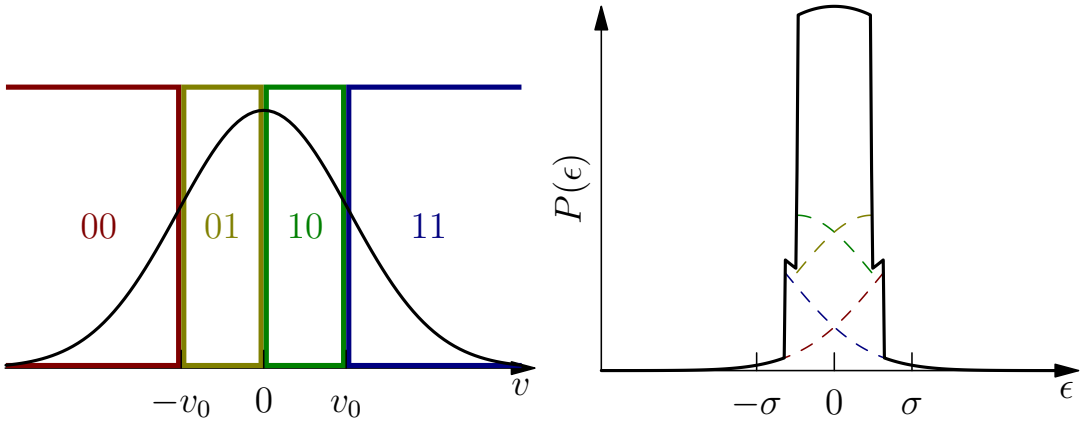

Figure 2: *2-bit quantization visualized.* (left) Voltage distribution, sample regions, and codes for 2-bit quantization. (right) Distribution of quantization noise values. The dashed colored lines show contribution to quantization noise from the four intervals. See text for more details.

applying symmetry about $v = 0$ there are three parameters that can be changed: a threshold, v_0, between "low positive" and "high positive" voltages, the reconstructed voltage values for "low positive", α, and the ratio between the "high positive" and "low positive" reconstructed values, R. These can be determined numerically by maximizing η_Q, resulting in $v_0 = 0.96\sigma$, $\alpha = 0.478\sigma$, and $R = 3.3359$. As for the 1-bit case, the absolute scaling of the reconstructed values (and hence the value of α is not important; values ± 1 and $\pm R$ are more commonly used.

Code	Range	Value	Frac.
00	$-\infty$ to $-v_0$	$-\alpha R$	17%
01	$-v_0$ to 0	$-\alpha$	33%
10	0 to v_0	α	33%
11	v_0 to ∞	αR	17%

The samples falling into the 4 states follows a 17%, 33%, 33%, 17% distribution. The RDBE interatively adjusts the sampler threshold to maintain this distribution. It should be noted that the distribution of quantization noise in the 2-bit case is becoming more rectangular and narrower compared to the 1-bit case. This will continue as number of bits increases.

3.3 4-bit quantization

4-bit quantization leads to 16 possible digital codes. This greatly increases the possible complexity through exponential increase in the number of parameters. In this case, there are eight positive-voltage intervals, each needing a reconstructed value and 7 interval boundaries, each needing a threshold value. A great simplification, at a very modest loss of SNR, would be to linearize the series of reconstructed values and the thresholds. If the first positive threshold is denoted v_0 then the 15 thresholds would be $-7v_0$, $-6v_0$, ... $7v_0$. If the first positive reconstructed value is denoted α, then the 16 reconstructed values would be -15α, -13α, ... 15α. It is natural to set $\alpha = v_0/2$. The first threshold value can be determined numerically based on maximization of η_Q to be $v_0 = 0.3356\sigma$, resulting in $\alpha = 0.1678\sigma$. In this case $\eta_Q = 0.989$.

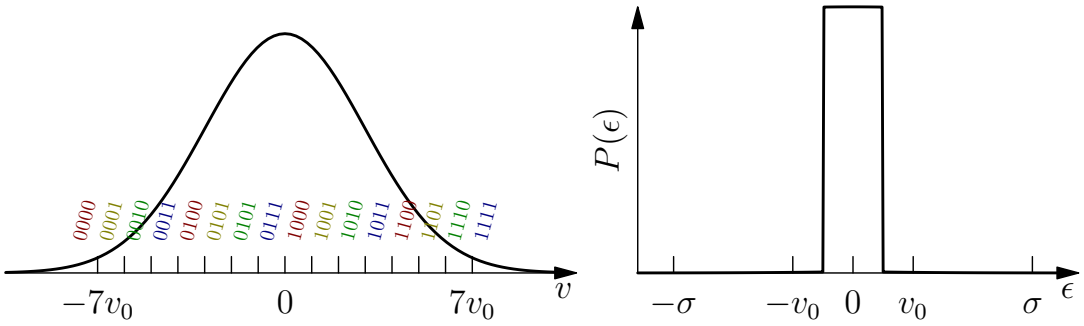

Figure 3: *4-bit quantization visualized.* (left) Voltage distribution, sample regions, and codes for 4-bit quantization. (right) Quantization noise distribution. Note that the horizontal scale of the right figure is 2.5 times smaller than that of the left. See text for more details.

Code	Range	Value	Frac.
0000	$-\infty$ to $-7v_0$	-15α	0.95%
0001	$-7v_0$ to $-6v_0$	-13α	1.26%
0010	$-6v_0$ to $-5v_0$	-11α	2.46%
0011	$-5v_0$ to $-4v_0$	-9α	4.31%
0100	$-4v_0$ to $-3v_0$	-7α	6.73%
0101	$-3v_0$ to $-2v_0$	-5α	9.40%
0110	$-2v_0$ to $-v_0$	-3α	11.8%
0111	$-v_0$ to 0	$-\alpha$	13.1%
1000	0 to v_0	α	13.1%
1001	v_0 to $2v_0$	3α	11.8%
1010	$2v_0$ to $3v_0$	5α	9.40%
1011	$3v_0$ to $4v_0$	7α	6.73%
1100	$4v_0$ to $5v_0$	9α	4.31%
1101	$5v_0$ to $6v_0$	11α	2.46%
1110	$6v_0$ to $7v_0$	13α	1.26%
1111	$7v_0$ to ∞	15α	0.95%

3.4 8-bit quantization

With 4-bit quantization already approaching $\eta_Q = 99\%$, there is not much room to improve signal-to-noise ratio. The primary purpose of more bits is to improve the dynamic range, allowing for a more precise representation of highly non-Gaussian distributed voltages. The most prominent (in VLBI) non-Gaussian signal type is RFI, but narrow-band satellite tones or radar returns also fit in this category. Extending the optimization technique used for the 4-bit case to the 8-bit case results in $v_0 = 0.031\sigma$ with a corresponding $\eta_Q = 0.9999$. It is proposed here to instead use linear thresholds with $v_0 = 0.3356\sigma$ (the same used in 4-bit quantization) with resulting $\eta_Q = 0.991$, but allowing a factor of 16 more voltage headroom. This trades a very small quantization efficiency loss for a power headroom increase of 24 dB, thus improving resilience against RFI.

This memo does not explore non-Gaussian (e.g., RFI-ridden) cases. The subject certainly deserves some attention and may be the subject of a future memo.

4 Digital corrections

Two voltage streams v_a and v_b have cross-correlation coefficient $\rho \equiv \mathrm{corr}(v_a, v_b)$. The cross correlation of their quantized values, $\hat{v}_a$ and $\hat{v}_b$, $\hat{\rho} \equiv \mathrm{corr}(\hat{v}_a, \hat{v}_b)$ will differ. The curve relating $\hat{\rho}$ to ρ depends on the details of the quantization being performed. In the 1-bit case (the most extreme quantization performed) the curve can be calculated analytically and is displayed in Figure 4. This is called the "van Vleck" correction. Similar corrections for other quantizations schemes are often referred to as "van Vleck" corrections as well. The curve shape will depend separately on the quantization scheme used for each of v_a and v_b.

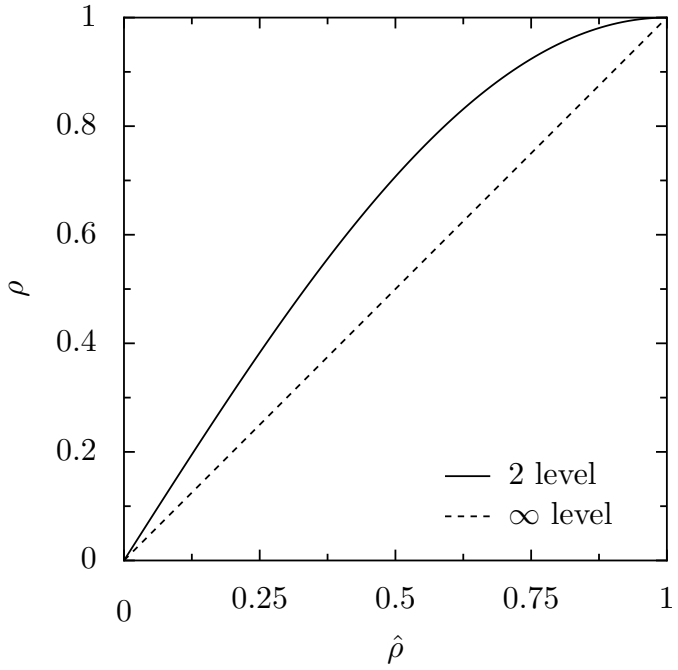

Figure 4: *Van Vleck correction.* A cross-correlation measurement made on quantized data needs to be corrected to produce the equivalent unquantized statistical quantity.

In VLBI the two ends of the van Vleck curve are most important. For cross correlations ρ is very small, rarely exceeding 0.1, except in some narrow-band maser observations. The slope of the

van Vleck curve at this point is the correction factor. The other end of the curve, at $\rho = \hat{\rho} = 1$ is relevant for autocorrelations.[2] Application of the digital correction only to the cross-correlation values provides a potentially *baseline-dependent*[3] amplitude correction that is critical for accurate flux density determination or high fidelity imaging. The table below summarizes the correction that is required on a baseline consisting of antennas using two, possibly different, quantization schemes. The cross-correlations (including cross-hand autocorrelations) should be divided by these corrections. Note that the values in this table assume SNR-optimized quantization at 1- and 2-bits, SNR-optimized linearized quantization at 4-bits, and at 8-bits the same linearized thresholding as used for 4-bits.

	1-bit	2-bit	4-bit	8-bit
1-bit	$2/\pi$	0.752	0.794	0.795
2-bit		0.882518	0.934	0.934
4-bit			0.988	0.989
8-bit				0.991

In the above table, the values for 1-bit × 1-bit ($2/\pi = 0.636620\ldots$) and 2-bit × 2-bit (0.882518) are shown with the same precision as they are used within AIPS[4]. The other values are determined from simulation to approximately 3 digits of precision.

4.1 Handling within DiFX

The program difx2fits converts the raw output of the DiFX correlator into FITS-IDI files[5] This program has been modified to perform the van Vleck corrections if one or more of three conditions are met:

- The new `--vanVleck` command line argument is supplied,
- Antennas with differing quantization are identified,
- Quantization with more than 2 bits per sample is used.

When this occurs, difx2fits will apply the appropriate corrective value to all cross correlations and to cross-polar autocorrelations and a new FITS keyword, `VANVLECK` will be set with value `1`. This keyword will be associated with the Primary Header Unit of the FITS file, which contains other information about the correlator.

After support for the `VANVLECK` keyword has been added to the major data reduction packages (see next sections) and sufficient time has elapsed to allow users to migrate to these updated versions, it is likely that DiFX will be changed to always perform this correction.

Should DiFX ever adopt the option to perform "zero-padded" FFTs, a more sophisticated variant of digital correction could be implemented in difx2fits which allows a more correct correction to be made. Zero-padding approximately doubles the compute complexity of the correlation process.

[2]Strictly speaking, the autocorrelations should have corrections made in the lag domain where a wide range of normalized correlation coefficient could be encountered. However, that is not possible to perform in cases where an FX correlator algorithm is used without zero-padding which is the case for the DiFX correlator.

[3]The correction is actually only baseline-dependent when different quantization methods are used on different antennas.

[4]This value is stored in variable ALFAC within the FITLD.FOR source file.

[5]See https://www.aoc.nrao.edu/~egreisen/AIPSMEM114.PDF for a description of the FITS-IDI convention.

It is possible that additional values for the VANVLECK keyword will be introduced in future versions that could indicate somewhat different means of performing the correction. It is advised that any non-zero value of the VANVLECK keywork should trigger the post-processing software to avoid performing the van Vleck correction.

4.2 Handling within AIPS

The AIPS task FITLD is used to load FITS-IDI data into AIPS. If the `VANVLECK` keyword is absent or set to zero, the current behavior of applying the van Vleck correction will be retained. If keyword `VANVLECK=1` is set, then this correction will be assumed to had occurred at the correlator and will be disabled within FITLD. No changes to pipelines or any data reduction process will be needed. FITLD from the 31DEC22 version of AIPS starting on July 18, 2022 has support for this new keyword and will be required for correct processing of files with `VANVLECK=1` set.

4.3 Handling within CASA

The CASA task importfitsidi is used to convert FITS-IDI into a MeasurementSet that can be used to process data in CASA. If the VANVLECK keyword is absent or set to zero, the current behaviour of applying the van Vleck correction for data with CORRELAT=DIFX will be retained. If keyword VANVLECK is set to a non-zero value, then this correction will be assumed to had occured at the correlator and will not be applied by importfitsidi. When CORRELAT=DIFX, the importfitsidi task will continue to apply the other normalization factors for 1-bit, 2-bit and mixed 1-bit and 2-bit quantization. The target for implementing support for the VANVLECK keyword is CASA 6.6.

Since the relevant code lives in the casacore library, any other software that is based on casacore should pick up the change as well.

5 Recommendations for VNDA

This study of quantization leads to a few recommendations:

- Retain the same, optimized, 2-bit quantization parameters from the RDBEs.
- Use linearized quantization thresholds and reconstructed values for 4- and 8-bit cases.
- Optimize the 4-bit case for maximal continuum signal-to-noise ratio in the absense of RFI.
- Retain the same level spacings and output reconstructed value system for the 8-bit case as for the 4-bit case, allowing the 8-bit case to provide significantly increased (24 dB power) headroom to allow for mitigation of RFI.
- Move the "Van Vleck" digital correction into `difx2fits` rather than in data reduction package fillers (such as "FITLD" in AIPS and "importfitsidi" in CASA).
- Consider a possible future 4-bit mode which sets v_0 to a higher value to allow a different compromise between sensitivity and headroom; this mode would not require any additional features within the FPGA firmware or low-level monitor and control of the VNDA hardware modules.

6 Acknowledgements

Thanks to Matt Luce, Sylas Ashton, and Clay Smith for discussing the mechanics of quantization in VNDA. Thanks to Justin Linford and Steve Tremblay for reviewing. Thanks to Eric Greisen for developing the path forward within FITS-IDI and AIPS. Thanks to George Moellenbrock for interesting discussions on this and related topics.