On the Precision Required in SDP Pipelines

Document Number ..... SKA-TEL-SDP-00000??
Document Type ..... MEMO
Revision 1B
Author ..... S. Salvini
Release Date 2017-08-09
Document Classification ..... Unrestricted
Status ..... Draft

Lead Author	Designation	Affiliation
Name of the lead author - to be filled in	Designation - to be filled in	Affiliation - to be filled in
Signature & Date:		

Owned by	Designation	Affiliation
Name of the owner - to be filled in	Designation - to be filled in	Affiliation - to be filled in
Signature & Date:		

Approved by	Designation	Affiliation
Name of the approver - to be filled in	Designation - to be filled in	Affiliation - to be filled in
Signature & Date:		

Released by	Designation	Affiliation
Paul Alexander	SDP Project Lead	University of Cambridge
Signature & Date:		

Revision	**Date of issue**	**Prepared by**	**Comments**
1B			

ORGANISATION DETAILS

Name	Science Data Processor Consortium
Address	Astrophysics Cavendish Laboratory JJ Thomson Avenue Cambridge CB3 0HE
Website	http://ska-sdp.org
Email	ska-sdp-pa@mrao.cam.ac.uk

Table of Contents

List of Figures

List of Tables

List of abbreviations

List of symbols

D	Required dynamic range, absolute factor (e.g. 10^3)
N_R	Number of receivers (Antennas/stations)
B_R	Number of bits from receiver
Ξ	SNR ? Signal to Noise Ratio, absolute factor (e.g.10^5)
S	Sampling rate (e.g. 600 MHz)
N_C	Number of channels (frequencies)
T_D	Cross-correlation dump-time
T_C	Dump-time for calibration
T_O	Total Observation time in seconds
H_O	Total Observation time in hours
X_2	Number of significant binary digits (bits) of X
X_10	Number of significant decimal digits of $X_{10} = X_2 \log_2 10$
L_G	General (unquantifiable) loss of digits (FFT, etc.) – $3 \leq L_G \leq 5$?

Summary

This document describes a simple, semi-qualitative approach to the determination of the numerical precision required by the SKA data processing, and use the results to suggest a strategy for precision handling.

Although this paper concentrates on the imaging pipeline, its arguments could be used for other SKA pipelines. Also, the discussion of this paper applies both to unpolarised as well as fully polarised situations. This point will not be laboured further,

Applicable and reference documents

Applicable Documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, *the applicable documents* shall take precedence.

Reference Number	Reference

Reference Documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, *this document* shall take precedence.

Reference Number	Reference

1 Symbols and General Considerations

Table 1 below lists the symbols used in this document

Symbol	Description
D	Required dynamic range, absolute factor (e.g. 10^3)
N_R	Number of receivers (Antennas/stations)
B_R	Number of bits from receiver
Ξ	SNR ? Signal to Noise Ratio, absolute factor (e.g.10^5)
S	Sampling rate (e.g. 600 MHz)
N_C	Number of channels (frequencies)
T_D	Cross-correlation dump-time
T_C	Dump-time for calibration
T_O	Total Observation time in seconds
H_O	Total Observation time in hours
X_2	Number of significant binary digits (bits) of X
X_10	Number of significant decimal digits of X; $X_{10} = X_2 \log_2 10$
L_G	General (unquantifiable) loss of digits (FFT, etc.); $3 \leq L_G \leq 5$?

Table 1: Symbols used in the document

Table 2 reports the number of digits (binary as well as decimal) for IEEE single and double precision floating point numbers. IEEE is specific in definition and behaviour of operations on such numbers and is virtually universally accepted. In my opinion non-IEEE definitions and operations could cause potential difficulties.

Precision	N. bits (binary digits)	N.Decimal Digits	Accuracy
Single (4 Bytes)	$B_{SP} = 23$	$6 \leq D_S P \leq 7$	$\epsilon = 2^{-23} = 1.210^{-7}$
Double (8 Bytes)	$B_{DP} = 52$	$15 \leq D_D P \leq 16$	$\epsilon = 2^{-52} = 2.10^{-16}$

Table 2: Floating point precision IEEE definitions

Significant digits are the digits in a quantity that carry meaning: any other digits beyond them, represent just numerical “dirt” or “noise”, so to speak and nothing can be read into them. The number of zeros preceding or following the digits does not alter the number of significant digits.

As an example, the value of π could be given as below.

Value of π	3	3.1	3.14	3.142
N. significant digits	1	2	3	4

For example, when using 2 significant digits, $2.1 + 0.012 = 2.1$ with *two* significant digits, not 2.012, which would have increased arbitrarily the number of significant digits to *four*: we have **no** information about the third and fourth digits of 2.1 and we can make no assumptions at all about them. Fixed point numbers may cause problems by producing computational errors with non-Gaussian distribution.

Generally speaking, *normalizable* floating point numbers are used (mantissa and exponent; fixed point numbers may cause problems by producing computational errors with non-Gaussian distributions.

2 Digits Gained and Lost

To all effects, the imaging pipeline from gathering signal at the receivers to producing an image is an averaging/summation process. For signals, noise, errors with Gaussian distribution, the number of significant bits in the mean and standard deviation increases with the logarithms (in base 2) of the square root of the number of samples S, namely $\log_2 \sqrt{S}$.

Receivers? samples are, in general, quantized to few bits: 8 bits have been suggested for SKA. Most likely, noise would reduce the number of significant digits in a specific sample (possibly losing all digits). However, as the noise is uncorrelated, averaging would reduce it (it has zero mean), while increasing the non-zero mean signal.

The dynamic range required can be seen as the number of significant digits in the imaging process: for example, a 10^5 dynamic range would require at least 5 significant decimal digits, equivalent to ~ 17 binary digits (bits). The removal of the stronger sources does not alter this in any particular way and can be ignored, as it cannot increase the number of significant digits: only extending the observation time can do so.

During the imaging process, the visibilities of different baselines are also added together, thus providing further samples for the overall averaging and increasing the number of significant bits in the final image: an interferometer consisting of N_R receivers has $\sim \frac{{N_R}^2}{2}$ baselines, attaining an approximate $\log_2(N_R/sqrt(2))$ increase in the number of significant bits (binary digits). Averaging steps, include cross-correlation, any summation/averaging of time dumps before calibration etc., summation/averaging of visibililties, stacking independent images, summation/averaging of channels, etc. Significant digits are lost by any computation carried out on. Some operations, for example FFT, summations, averaging, various matrix and vector products cause minimal numerical error thus incurring in minimal loss of significant digits. Others, such as solvers, for example the solution of linear systems of equations), can leads to the loss of significant digits.

The relative (forward) error δ caused by an operation can be represented by:

$$\delta \leq f(N)\ \kappa(\ldots)\ \|\ldots\|\ \epsilon \tag{1}$$

where the symbols are defined in Tables 3 below:

Symbol	Description
$f(N)$	Some (undefined) function of the problem size. In general, it varies very slowly with N. Pathological cases can be specifically built.
$\kappa(\ldots)$	The condition number of the operation being carried out. This indicates the sensitivity of a solution obtained to small changes of the problem
$\|\ldots\|$	Some normalization factor of the problem – see also below
$epsilon$	The machine accuracy. It is the smallest number such that $1 \neq 1 + \epsilon$.

Table 3: Symbols used in the document

The condition number is a very important concept in analysing numerical performance. It is intrinsic to the problem, and little that can be done to improve on it (i.e. to reduce it) without changing the computational workflow. Unfortunately, some algorithms can induce higher effective condition numbers, thus leading to greater digit losses. The oft-told butterfly effect

describes, of course, an ill-conditioned problem. The worse (larger) the condition number, the more initial data (i.e. significant digits) are required to counteract loss of digits (i.e. information loss).

The normal equation method, very often used, has an effective condition number $\kappa(A)^2$ where $\kappa(A)$ is the matrix of the linear minimization system; other methods, such as SVD, QR, etc. have condition number $\kappa(A)$. So, for example, if $\kappa(A) \sim 10^4$, the normal equation method will lose 8; the other methods may incur losses of 4 digits: single precision would be inadequate and higher precision be needed; the other methods could use single precision and return solutions accurate to 4 digits.

Table 4 reports the quantities of Table 3 for some operations which could be used in the imaging pipeline:

Operation	**Normalization factor** $\|\| \ldots \|\|$	$\boldsymbol{\kappa(\ldots)}$
Matrix vector product Ax	$\|\|A\|\|\|\|x\|\|$	~ 1
Sum of vectors $x + y$	$\max(\|\|x\|\|, \|\|y\|\|)$	~ 1
FFT of vector x	$\|\|x\|\|$	~ 1
Solver (e.g. calibration), e.g. $Ax = b$	$\|\|x\|\|$	$\kappa(A) > 1$ ($\gg 1$?)

Table 4: Values of the various symbols for some imaging pipeline operations

Matrix and vector operations could incur the loss of very few binary digits. Experiments with the FFT has shown that binary digits loss is limited to 2 or 3 in cases likely to occur in SKA, including rather extreme cases.

Noise can also be interpreted as causing the loss of significant digits at the receivers. Basically, it lengthens the time series duration required for some specific image quality, i.e. dynamic range. For SNR smaller than one, noise dominates each receivers? sample, but, of course, signal can be recovered by using extensive time dumps or observations.

These considerations lead only to some qualitative, at best semi-quantitative, analysis. None the less, we can now define the “economics” of the digits? acquisition and loss for an observation, from the receivers to the final image, and that can provide guidance to the choice of precision required for the various steps. Some pipelines? component could include monitoring to possibly change dynamically (and automatically) computational requirements, similarly to what is done in other fields. For example, solvers? condition numbers could be tracked, varying the convergence tolerance etc. as required. For example, StefCal allows to estimate the condition number very cheaply as the iteration progresses.

Table 5 reports the contribution for each components of the ?digits ledger book? in terms of the quantities defined

By reducing the convergence tolerance by κ^{-1}, the inverse of the condition number, it is possible to avoid the loss of $\log_2 \kappa$ binary digits at the cost of increased number of solvers iterations. This could have a very important effect in limited the observation length.

We can finally put it all together to it all together we finally get I_2, the number of digits in the image as a function of the quantities defined above in the case of no extra digits in the

Step	Value
Initial number of bits for each receiver's sample	$\min(B_R, \log_2 \Xi)$
Gain: summation over each calibration interval	$\frac{1}{2}\ \log_2 \frac{S T_C}{N_C}$
Loss: solver (from condition number)	$\log_2 \kappa$
Gain: overall observation summation	$\frac{1}{2}\ \log_2 \frac{T_O}{T_C}$
Gain: summation/averaging over the baselines	$\log_2 \frac{N_R}{\sqrt{(2)}}$
Loss: generic losses (FFT, gridding, etc.)	L_G

Table 5: Binary digits gain/losses for the computation stages

solver:

$$I_2 = \log_2 D = \min(B_R, \log_2 \Xi) + \frac{1}{2} \log_2 \frac{S\, T_O}{N_C} + \log_2 \frac{N_R}{\sqrt{2}} - L_G - \log_2\, \kappa \tag{2}$$

where, as already explained, the κ term within square bracket could be removed.

This can be inverted to obtain the observation time in terms of the dynamic range required:

$$H_0 = \frac{2^{2[\log_2 D - \min(B_R, \log_2 \Xi) + \log_2 \kappa + L_G] - \log_2 \frac{S\, N_R}{\sqrt{2} N_C}}}{3600} \tag{3}$$

Table 6 below show some examples:

N_R	200	200	512	512
B_R	8	8	8	8
Ξ	0.1	0.1	0.1	0.1
N_C	20000	60000	20000	60000
N_R	200	200	512	512
κ	100	10	100	10
L_G	4	4	4	4
solver converged digits (δ the tolerance				
$-\log_{10} \delta^{(1)}$	2.4	2.5	2.2	2.0
$-\log_{10} \delta^{(2)}$	4.4	4.5	4.2	4.0
Observation Time to dynamic range				
H_O	100	100	100	100
$D^{(1)}$	$1.2 10^3$	$1.5 10^3$	$2.1 10^3$	$1,2 10^3$
$D^{(1)}$	$1.2 10^5$	$1.5 10^5$	$2.1 10^5$	$1,2 10^5$
Dynamic range to observation time				
D	10^5	10^5	10^5	10^5
$H_O^{(1)}$	$7.1\, 10^5$	$1.5\, 10^3$	$2.1\, 10^5$	$6.5\, 10^3$
$H_O^{(1)}$	71	43	22	65

Table 6: Some examples

In the table, the superscripts $^{(1)}$ and $^{(2)}$ refers to results obtained without and with the

κ correction. As it can be seen, it is important to avoid a very significant increase of the observation time, the increase being proportional to a factor κ^2.

It should, however, be remarked that in the first place, receivers-based calibration problems have in general low condition numbers. In second place, convergence tolerances used are well in excess of the worst case scenarios, possibly above the minimum requirements to avoid loss of digits for ill-conditioning.

3 Conclusions

From the previous discussion, it can be seen that only the averaging computational stages contribute to any increase in the number of significant digits. This suggests that CLEAN major as well as minor cycles would not contribute to any increase in accuracy. Depending on how the operations are carried out, loss of accuracy is, in all likelyhood, small.

Sources removal cannot add to the absolute accuracy, thus the relative accuracy would decrease, which is hardly surprising. This leads to an interesting and open question: should the computation reverts to single precision after the brightest sources have been removed?

This paper would like to suggest that a viable "prescription" could be, under likely SKA conditions:

Dynamic Range	**Prescription**
$D < 10^5$	Use single precision throughout
$D \geq 10^5$	Use single precision until the last accumulation step, then switch to double precision there and for all subsequent steps. Should the computation revert to single precision after subtraction of the strongest sources, thus having reduced the number of significant digits?

Although the effects of ill-conditioning when using solvers can be rather dramatic, simple steps can and should be taken

Prescription for ill-conditioning κ
Always carry out the solution to preserve the number of significant digits by correcting for the condition number. This simply requires setting the convergence tolearance appropriately. hline Monitor the condition number (it can be done trivially with some algorithms, e.g. StefCal) and correct the convergence requirements accordingly on-the-fly. hline Avoid normal equations and any other methods that induce higher condition numbers unless the condition number is demonstrably very low.