Reduced-Precision Acceleration of Radio-Astronomical Imaging on Reconfigurable Hardware

Citation for published version (APA):
Corda, S., Veenboer, B., Awan, A. J., Romein, J. W., Jordans, R., Kumar, A., Boonstra, A. J., & Corporaal, H. (2022). Reduced-Precision Acceleration of Radio-Astronomical Imaging on Reconfigurable Hardware. *IEEE Access*, *10*, 22819-22843. https://doi.org/10.1109/ACCESS.2022.3150861

Document license:
CC BY

DOI:
10.1109/ACCESS.2022.3150861

Document status and date:
Published: 01/01/2022

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 31, 2022, accepted February 7, 2022, date of publication February 10, 2022, date of current version March 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150861

Reduced-Precision Acceleration of Radio-Astronomical Imaging on Reconfigurable Hardware

STEFANO CORDA[1,2], BRAM VEENBOER[3], AHSAN JAVED AWAN[4], JOHN W. ROMEIN[3], ROEL JORDANS[1], AKASH KUMAR[2], (Senior Member, IEEE), ALBERT-JAN BOONSTRA[3], AND HENK CORPORAAL[1]

[1]Electronic Systems Group, Eindhoven University of Technology, 5600 Eindhoven, The Netherlands
[2]Chair of Processor Design, CFAED, Technische Universität (TU) Dresden, 01062 Dresden, Germany
[3]Netherlands Institute for Radio Astronomy (ASTRON), 7991 Dwingeloo, The Netherlands
[4]Department of Cloud System and Platforms, Ericsson Research, 16480 Stockholm, Sweden

Corresponding author: Stefano Corda (s.corda@tue.nl)

This work was supported in part by the European Commission by the Horizon 2020 Program for the Project ''Near-Memory Computing (NeMeCo)'' through the Marie Sklodowska-Curie Innovative Training Networks European Industrial Doctorate under Project 676240, in part by the Open transPREcision COMPuting (OPRECOMP) Summer of Code 2020, in part by the European Network on High Performance and Embedded Architecture and Compilation (HiPEAC) Collaboration Grant, and in part by the Xilinx Donation Program.

ABSTRACT Radio telescopes produce large volumes of data that need to be processed to obtain high-resolution sky images. This is a complex task that requires computing systems that provide both high performance and high energy efficiency. Hardware accelerators such as GPUs (Graphics Processing Units) and FPGAs (Field Programmable Gate Arrays) can provide these two features and are thus an appealing option for this application. Most HPC (High-Performance Computing) systems operate in double precision (64-bit) or in single precision (32-bit), and radio-astronomical imaging is no exception. With reduced precision computing, smaller data types (e.g., 16-bit) are used to improve energy efficiency and throughput performance in noise-tolerant applications. We demonstrate that reduced precision can also be used to produce high-quality sky images. To this end, we analyze the gridding component (Image-Domain Gridding) of the widely-used WSClean imaging application. Gridding is typically one of the most time-consuming steps in the imaging process and, therefore, an excellent candidate for acceleration. We identify the minimum required exponent and mantissa bits for a custom floating-point data type. Then, we propose the first custom floating-point accelerator on a Xilinx Alveo U50 FPGA using High-Level Synthesis. Our reduced-precision implementation improves the throughput and energy efficiency of respectively 1.84x and 2.03x compared to the single-precision floating-point baseline on the same FPGA. Our solution is also 2.12x faster and 3.46x more energy-efficient than an Intel i9 9900k CPU (Central Processing Unit) and manages to keep up in throughput with an AMD RX 550 GPU.

INDEX TERMS Accelerator architectures, approximation methods, astronomy, central processing unit, field programmable gate arrays, graphics processing units, high level synthesis, high performance computing, reconfigurable architectures, scientific computing.

I. INTRODUCTION

The future generation of radio telescopes, such as the Square Kilometre Array (SKA) [1], will have to process a massive quantity of data (in the order of TeraBytes per second) using high-performance computing systems (in the order of Exaflops per second) [2] with high energy efficiency [3]. The demanding data and computation requirements are mainly caused by the high-resolution images that must be processed to discover new objects in the sky such as stars, supernovas, galaxies, etc. [4].

The most dominant compute kernels of the radio-astronomical imaging pipeline are the gridding, and degridding algorithms [5]. These kernels can be executed highly

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

efficiently in single-precision floating-point accuracy using GPUs. A GPU accelerated imaging solution is shown to meet the computing power and energy efficiency requirements of SKA [6]. FPGAs (Field Programmable Gate Arrays) are well known to be energy-efficient platforms for fixed-point computation [7], [8]. However, recent FPGA platforms can also be efficiently used for floating-point computations, e.g., Intel Arria 10 and Stratix 10 [9]. Furthermore, High-Level Synthesis (HLS) toolchains enable FPGAs to be programmed much more easily compared to using a Hardware Description Language (HDL). This makes FPGAs an attractive accelerator platform [10]. However, the prior art shows that FPGAs are less energy-efficient than GPUs for the radio astronomy application domain [11].

Reduced precision computing is a technique where smaller data types are used to reduce area usage, execution time, and power consumption within noise-tolerant applications without losing information [12]. It has been widely applied in different application domains, especially, in deep learning applications [13], [14]. Existing studies propose the use of reduced precision also for the deconvolution kernel [15], apply mixed precision to other steps of the radio-astronomical imaging acquisition pipeline, e.g., correlator [16], or other radio-astronomy domains, e.g., computation of tomographic reconstructors [17]. However, the works mentioned above employ standard data types supported by CPUs and GPUs, such as double-, single- and half-precision floating-point, and do not evaluate custom data types. While Intel FPGAs such as Arria 10 and Stratix have native support for single-precision floating-point operations (2 flops per DSP, Digital Signal Processing, per cycle), Xilinx FPGAs do not. Xilinx Ultrascale FPGAs, and the Alveo U50 in particular, have many DSPs though [18], and it is tempting to employ these to implement reduced-precision arithmetic.

This paper evaluates the noise tolerance in a state-of-the-art radio-astronomical imaging algorithm to reduce precision data types, highlighting possible optimization opportunities. We evaluate the performance of the radio-astronomical imaging algorithm on a Xilinx Alveo U50 FPGA employing High-Level Synthesis and traditional floating-point data types. Then, we demonstrate how custom floating-point can be efficiently employed to improve performance while maintaining high output quality. To the best of our knowledge, this is the first work that focuses on assessing the applicability of reduced precision for radio-astronomical imaging.

The main contributions of this work are:

- An in-depth analysis (*Section V*) to determine the precision requirements of radio-astronomical Image-Domain Gridding [19], included in the state-of-the-art imager WSClean [20]. It highlights that standard data types such as half-precision and brain floating-point do not meet them. Furthermore, we show how the analysis can be carried out faster (4x in our case).
- A custom floating-point gridding accelerator for radio-astronomical imaging on reconfigurable hardware. To the best of our knowledge, this is the first FPGA implementation applying reduced precision for radio astronomical imaging with negligible quality loss (*Section VI*). Furthermore, we determine several guidelines for accelerator design on Xilinx FPGAs by using Xilinx Vitis with regard to the state of the art.
- An in-depth performance evaluation (*Section VII*) of our accelerator prototypes and of state-of-the-art architectures with similar features such as peak performance, thermal design power and lithography technology. The Xilinx Alveo U50 outperforms an Intel i9 9900k CPU in terms of energy efficiency. Moreover, our best accelerator prototype outperforms its single-precision baseline and keeps up in throughput with an AMD RX 550 GPU.

The paper is structured as follows: *Section II* explains the background information regarding radio-astronomical interferometry and imaging. Then, in *Section III* we discuss background information about reduced precision and data types. *Section IV* shows our methodology. In *Section V* we discuss the analysis results of the WSClean imager. *Section VI* reports details on how we design the accelerator architecture. It is evaluated in *Section VII*, where we also present the lessons learned in this work. Finally, we describe the related work in *Section VIII* and we conclude the paper in *Section IX*.

II. RADIO-ASTRONOMICAL IMAGING

A radio telescope detects electromagnetic waves that originate from radio sources in the universe. The signals are used, among other things, to construct a map of the sky containing the positions, intensity, and polarization of the sources. Radio telescopes such as LOFAR [21] and SKA1-Low [22] are comprised of many (small) dipole antennas that measure two orthogonal polarizations of the radio sources, while other radio telescopes (such as the VLA [23], MeerKAT [24] and SKA1-Mid [25]) are based on an array of dishes. As shown in *Figure 1*, a station consists of multiple antennas, for which the signals for every distinct frequency channel are combined. The signals of a pair of stations (a baseline) are multiplied and integrated (correlation ②) for a short period of time (in the order of seconds), thus producing a single visibility (a 2 × 2 matrix). The data that the telescope produces (the `visibilities`) is, therefore, a three-dimensional matrix (with indices number of baselines, frequency channels and correlations). The relation between visibilities and sky brightness is given by a measurement equation; see [26] for complete details.

The visibilities are first calibrated (③) and next used to reconstruct the sky brightness in the observation direction using an imaging step (④) [27].

This work focuses on ④. The imaging step (see *Figure 2*) starts with an empty sky model and it consists of an iterative process: 1) the `inversion` step is used to produce a `dirty image`; 2) one or more bright sources are detected in this image by a deconvolution algorithm such as `CLEAN` (see also Section II-B); 3) a `model image` is created, which contains all of the sources in the sky model; 4) visibilities

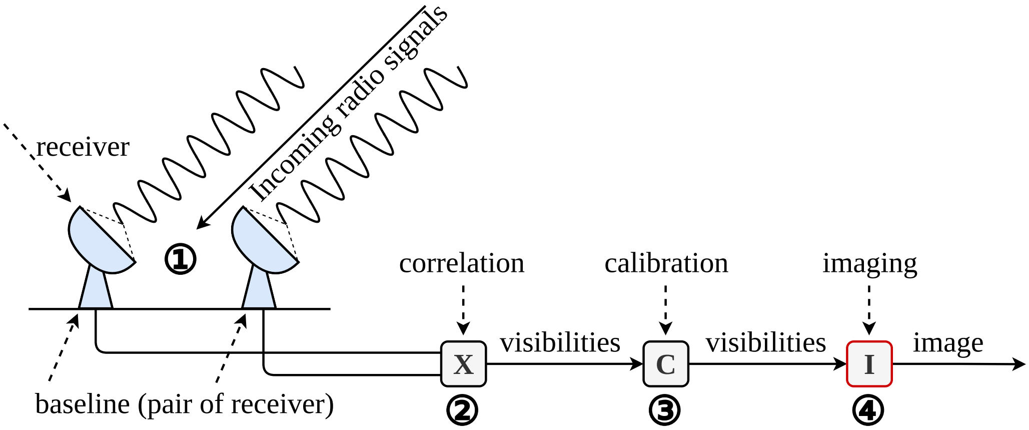


FIGURE 1. Radio astronomy image acquisition: the incoming radio signals are digitized and then correlated and calibrated before the imaging step is executed. We focus on the imaging step, which is highlighted in red.

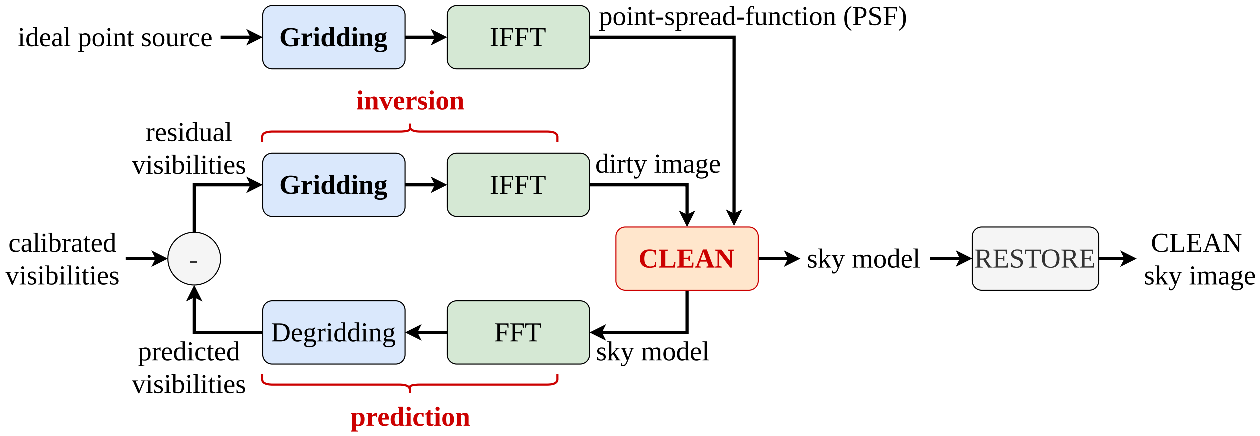


FIGURE 2. High-level schematic of the radio-astronomical imaging step. The three main phases, inversion, prediction, and deconvolution (or CLEAN) are highlighted in red. The critical kernel, the gridding, is included in the inversion step (it is highlighted in bold). We included the point-spread-function (PSF) computation, an inversion step with an all-ones matrix of visibilities used during the CLEAN.

corresponding to this model image are `predicted`; 5) subtracting the predicted visibilities from the measured (and calibrated) visibilities yields `residual visibilities`. This process subtracts strong sources from the measurements, which mask the more interesting weak sources. This step is repeated until the sky model converges. Finally, the sky model is used to create the sky image.

The inversion and prediction steps comprise of 2D FFT and a `gridding` or `degridding` step. The gridding and degridding steps are typically the most compute-intensive image processing steps. To attain high-quality sky images, they need to correct for Direction-Independent Effects (the curvature of the earth, `W-Term` correction) and Direction-Dependent Effects (such as ionospheric effects, `A-Term` correction). The W-Term can be corrected by applying a convolution kernel to every visibility. The required convolution kernel could be huge depending on parameters such as the field-of-view and distance between receivers. A-Term correction requires these convolution kernels to be different for every receiver and change over time according to changes in the Direction-Independent effects. These properties make imaging with correction for W-Terms and A-Terms particularly challenging.

The processing facilities that the SKA consortium is planning to build for the low and mid frequencies consist of large Science Data Processors. Each of them has a peak performance in the order of 6.50 PFLOP/s and a thermal design power in the order of 125 MW [6].

A. IMAGE-DOMAIN GRIDDING

Image-Domaing Gridding (IDG) is a state-of-the-art algorithm for both gridding and degridding [19]. IDG performs both W-correction and A-correction in the image domain, avoiding large convolutions functions. The algorithm performs gridding and degridding using `subgrids`, which represent low-resolution sky images for a subset of visibilities. This approach exposes a lot of parallelism (subgrids can be processed in parallel), which makes it highly efficient on parallel hardware such as GPUs [6]. In IDG, gridding comprises three steps: 1) visibilities are gridded onto subgrids;

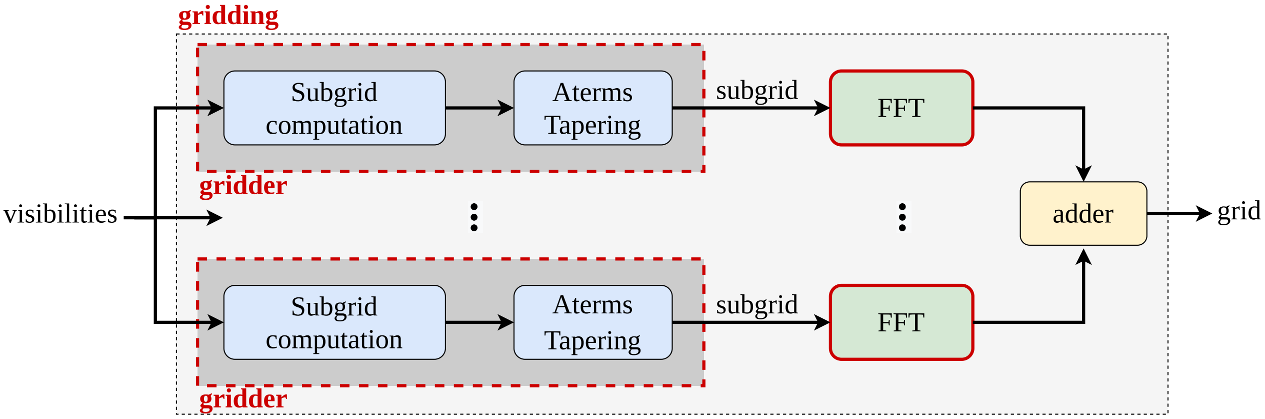


FIGURE 3. Gridding high-level representation. It consists of multiple gridder computations (subgrid computation and tapering) and FFTs that process the input visibilities into subgrids. Then the subgrids are processed by the adder to obtain a grid. The gridder and FFT, red boxes, are the focus of this and the related work [11].

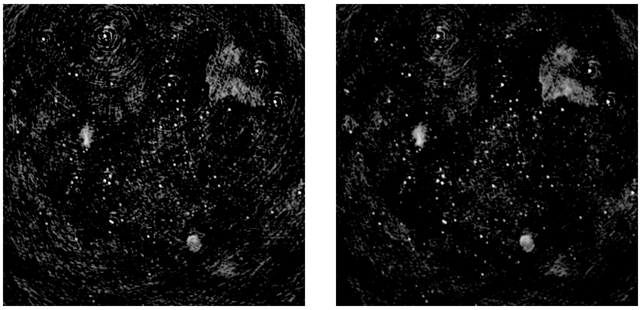


FIGURE 4. Comparison between dirty image (left) and CLEAN image (right) applying Cotton-Schwab algorithm. The CLEAN image has reduced noise, e.g. rotation lines around the image(visible especially at the corners).

2) subgrids are Fourier transformed; 3) subgrids are added to the larger final grid. IDG degridding comprises these steps in reverse order. Refer to [6], [19] for all the details on this algorithm and a formal derivation.

Image-Domain Gridding performs much better [28] than classical gridding/degridding algorithms, such as W-projection [29] or AW-projection [30]. It also employs W-terms to solve artifacts around sources away from the phase center in wide-field imaging. Moreover, IDG image quality is higher than W-projection because IDG, like AW-projection, corrects for DDEs (direction-dependent effects, also called the A-terms), but the computational costs for such DDE corrections are much lower for IDG than for AW-projection [6]. IDG also has higher per-visibility accuracy compared to the other algorithms [19].

B. DECONVOLUTION

The objective of a CLEAN (or deconvolution) algorithm is to detect sky sources by iteratively finding the brightest peaks in a dirty image and fitting a sky model. In *Figure 4* we show an example of a dirty image and the corresponding image after a deconvolution algorithm has been applied. The CLEAN image shows lower noise compared to the dirty image.

Many deconvolution (CLEAN) algorithms have been proposed in the literature. The simplest one is the so-called `Högbom` [31]. After the dirty image is generated from the imaging step (gridding and IFFT), the Högbom CLEAN tries to remove the noise in the image. This is done iteratively, looking for the maximum value in the image. Then, the algorithm subtracts the Point Spread Function (PSF), which is a function dependent on the telescope used. It is computed like the dirty image using as input a Visibility array with values equal to 1, multiplied by a gain factor. After a certain number of iterations or when a certain threshold (e.g., 3σ of the standard deviation) is reached the algorithm stops. This algorithm does not include the prediction step.

The `Clark` CLEAN [32], which is an improvement of the previously described algorithm, adds a feedback loop and tries to remove alias errors. It is possible to distinguish between `major` and `minor iterations` in this case. The minor iterations are represented by the peak search, similar to Högbom (the CLEAN box in *Figure 2*). Then, the model image is Fourier transformed and subtracted from the dirty image. This is the so-called major iteration.

The `Cotton-Schwab` [33] and `Multiscale` [34], [35] are the most employed and modern ones, they have in common the prediction phase, thus including the degridding algorithm. Here the subtraction is done at the visibilities level reducing the pixelation error. More precisely, for these algorithms, a major iteration consists of an entire iteration to transform the data from the visibility domain to the image domain (inversion). The major iterations are executed until a threshold is reached, e.g., until 80% of the flux (which is a power density measure) is removed from the dirty image during the minor iterations.

The Multiscale algorithm operates on a set of residual images obtained by convolving the dirty image with different scale sizes. The peak subtraction step is performed on all the scaled imaged, and only the subtracted components are stored in the CLEAN component table. After being scaled, positioned, and convolved, the final image is obtained by adding the components. It decreases the effects of the pedestal of uncleaned flux and strong sidelobes present in the dirty beam (or Point-Spread Function), which are referred to as ‘‘clean bow,’’ around bright resolved structures and has better convergence properties [36].

III. REDUCED PRECISION

Reduced precision is a software and hardware technique employing smaller data types to improve performance. It can be applied at the software level if the architecture supports reduced data types, e.g. half precision in modern GPUs. However, a custom architecture should be designed on FPGA (or ASIC) hardware when a non-supported data type is needed. The main benefits of this technique are the reduced processing elements (PEs) size, which leads to higher throughput, and reduced memory requirements, which increase the effective memory bandwidth. Key factors for reduced precision are data types that are described in *III-A*. In *III-B* we briefly introduce reduced precision and the advent of appealing tools for exploring custom data types in software and hardware.

A. DATA TYPES

Standard architectures, such as CPUs and GPUs, typically support single-precision and double-precision floating-point applications that perform scientific computations. Commonly, single precision is the most widespread data type since the major part of systems supports it. Indeed, even if double precision is supported on most GPUs, they often do not have dedicated units for double-precision computations except for high-end GPUs like Tesla V100 or Ampere A100 [37], thus reducing performance [38]. Radio-astronomical imaging runs precisely enough using single-precision floating-point. For this reason, we focus on data types up to 32 bits. We present in *Figure 5* the most commonly used data types today. In particular, we recognize two main categories: standard data types where the bit length is fixed and custom data types where the data length is defined at design time, compile time, or runtime.

Except for posit and fixed points, the other data types are floating-point representations that can be expressed by *Equation 1*. More precisely, the `exponent` and the `mantissa` bits are responsible for respectively the dynamic range and the precision of the data type. The dynamic range limits the smallest and largest number representable, while the precision is the represented number's resolution (number of digits).

$$(-1)^{sign} * mantissa * 2^{exponent} \tag{1}$$

In *Table 1* we present the mantissa and exponent sizes of the main standard data types. **Single-Precision** [39] (or `binary32` [40]) and **Double-Precision** (or `binary64`) **Floating Point** are commonly supported on GPUs and have been added to the **IEEE 754** standard. With the advent of deep learning applications and their noise tolerance and need for reduced length data types, half precision usually offers 2x the performance of single precision in applications that can tolerate the noise introduced by the lower dynamic range and less precision. For the same reason, NVIDIA presents the new `Tensor Float-32` with the release of the NVIDIA A100, the new AI and HPC flagship GPU. This format has the same dynamic range of the **binary32** but reduced precision, which is claimed to be sufficient for most of today's AI applications. **Brain Floating Point** [41] offers a further precision reduction while keeping the same dynamic range offered by single precision. This is especially employed by Intel [42] and Google [43].

TABLE 1. Standard floating-point formats.

Name	Exponent	Mantissa
Single-precision	8	23
Half-precision	5	10
NVIDIA-Tensor	8	10
Brain	8	7

Apart from the standard data types mentioned above, other data types are not available in mainstream architecture or do not have a pre-defined number of bits. These data types are described below:

Fixed Point [44]: this format is represented by *Eq. 2*. A real number is represented by two numbers, one for the integer part and one for the fractional part. Compared to floating point, it has a smaller dynamic range since there is no exponent. Still, the hardware implementation is easier since it considers two numbers (integer and fractional), and it is usually employed on custom accelerators, and FPGA [45], [46].

$$(-1)^{sign} * integer.fractional \tag{2}$$

Custom Floating Point [47]: a custom floating-point is super-set of the floating-point above mentioned. It consists of all the possible floating-point format combinations. They are typically used in embedded systems, where the numeric representation is customized for the specific application by selecting specific bit lengths for the mantissa and exponent fields.

Posit [48]: is a numeric representation that has been proposed as a substitute for floating-point data types. Usually, posit has a higher dynamic range than the floating-point with the same bit length (see *Figure 6*). Moreover, posits have a tapered decimal accuracy (see *Equation 3*, where x and y are two numbers with same sign) which means that the decimal accuracy reported in *Figure 6* is roughly symmetrical, and the highest precision is achieved for numbers near 1 (the horizontal axis is reporting the base-2 logarithm of the numbers).

$$decimal\_accuracy = -log_{10}(|log_{10}(x/y)|) \tag{3}$$

This feature differs from floating points that have the constant accuracy across the dynamic range, except for small numbers (left side), and the accuracy suddenly falls off a cliff (right side) to accommodate all the NaN (not a number

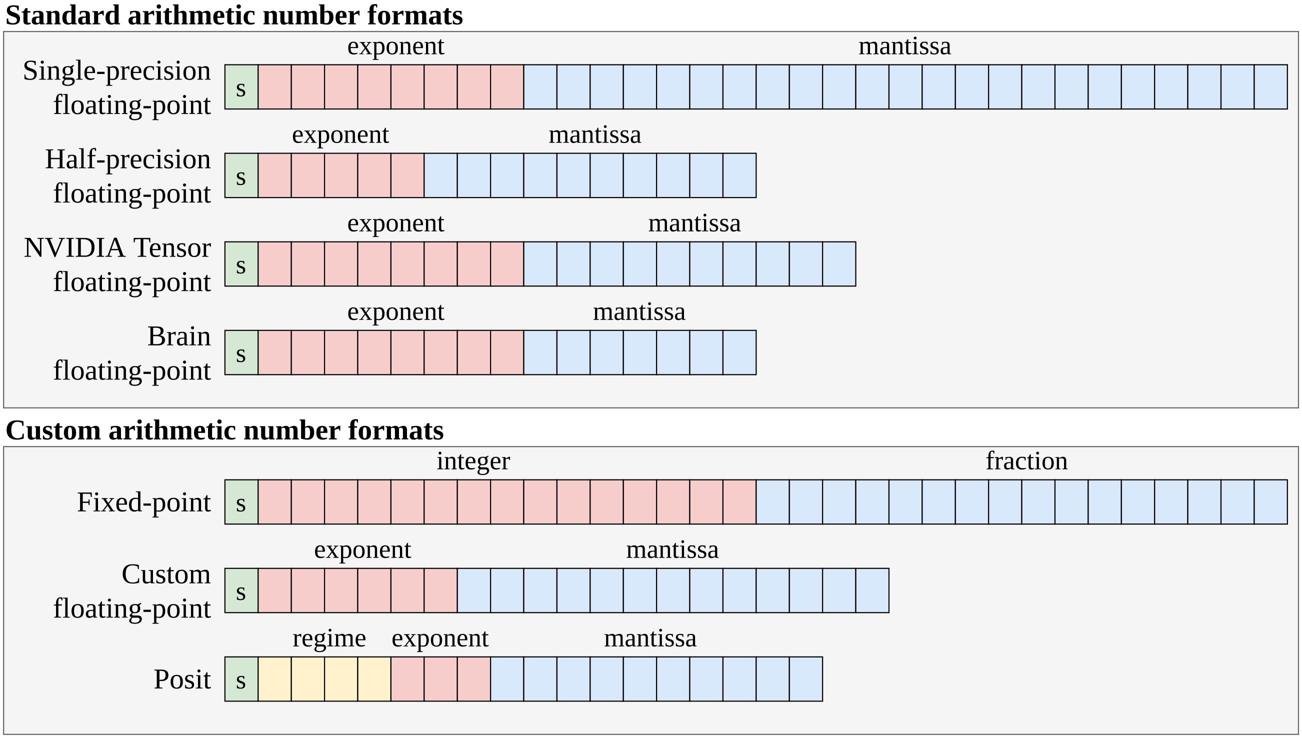

Standard arithmetic number formats
Single-precision floating-point
Half-precision floating-point
NVIDIA Tensor floating-point
Brain floating-point
exponent
mantissa
s
Custom arithmetic number formats
Fixed-point
Custom floating-point
Posit
integer
fraction
regime
exponent
mantissa

FIGURE 5. Data-types overview. Standard arithmetic number formats such as single-precision and half-precision floating-point are data types usually supported by modern CPUs and GPUs. Custom arithmetic number formats comprise some of the main data types employed in research and are often deployed on FPGAs. The custom formats reported are examples, and the number of bits may differ.

values). However, we are not considering posits as a possible datatype candidate since it is usually more expensive compared to floating-point for multiplication and addition operations [49]. For more details refer to [48].

We employ custom floating point for both analysis and hardware design, which comprises single precision, half precision, NVIDIA Tensor, and brain floating point. Based on the analysis in *Section V* we evaluate the precision requirements of the target application and discuss the excluded data types.

B. REDUCED PRECISION TOOLS

Reduced precision is a branch of approximate computing, which usually consists of either reducing the bit size of standard data types or employing more efficient data types [12]. Recently, there has been the rise of automated/assisted precision tuning tools and emulation libraries to help and improve the selection of custom data types inside applications. However, the major part of the above-mentioned works support only standard data types [52]–[54] or fixed point [55].

Flegar *et al.* [47] designed `FloatX`, a C++ template library capable of emulating `custom` floating point, which we employ in our analysis. FloatX also has a reduced execution time overhead compared to the previous library since it employs hardware-supported floating-point types as back-end.

Recently, High-Level Synthesis libraries for supporting custom floating-point precision have been researched. These libraries are easy to use and portable compared to RTL approaches [56]–[58].

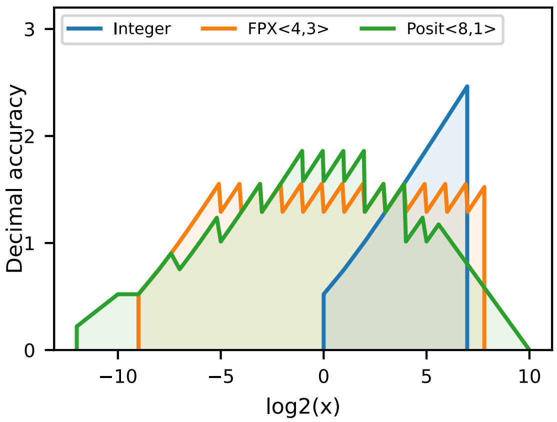


FIGURE 6. The figure is inspired from [48]. Decimal accuracy for 1) 8-bit signed integers; 2) a custom floating points with 4-bit exponent and 3-bit mantissa; 3) 8-bit posits with 1-bit exponent [50], [51]. Posits have a larger dynamic range compared to floating points and integers (very small). Integers have higher decimal accuracy for larger numbers. While posits have a tapered decimal accuracy, floating points have approximately a constant accuracy across the dynamic range.

DiCecco *et al.* [56] propose a custom-precision floating-point library (`CPFP` [59]) for High-Level Synthesis, and they evaluate it on a small convolution neural network. While the custom floating-point IP employs fewer resources than single precision, the FPGA design has a lower throughput than the CPU. Thomas [57], [58] proposed a more efficient (see *Appendix A*) templatized floating-point library for high-level-synthesis (`THLS` [60]). This library, which we employ in this work, is also templatized and eventually supports

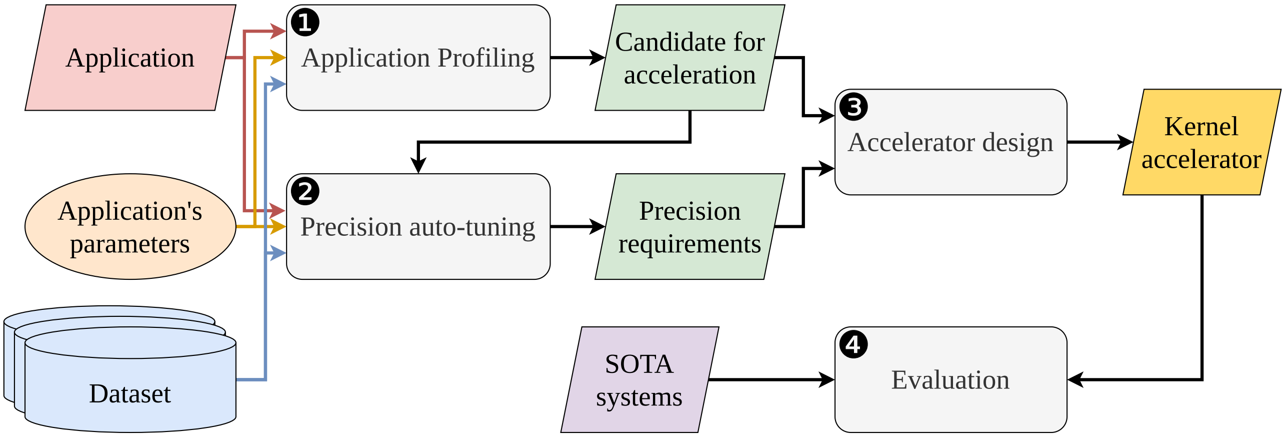


FIGURE 7. High-level overview of the employed methodology. The application's bottleneck is detected by applying application profiling. Then a precision auto-tuning technique defines the precision requirements for the accelerator, which is designed and deployed. Finally, the accelerator performance is assessed with state-of-the-art (SOTA) architectures.

heterogenous custom floating-point operations. The proposed solution has similar resource consumption when comparing standard data types and notable resource reduction when employing reduced-precision data types. The two approaches mentioned above are easier and portable for FPGA development than employing custom floating-point IPs similar to FloPoCo [61], which need to be used as black-boxes.

IV. METHODOLOGY

We present the methodology employed in *Figure 7*. We first ❶ profile the radio-astronomical imager to determine the most time consuming and thus critical kernels. We perform this analysis on different datasets employing the application's parameter described in *IV-A*. Then, ❷ we evaluate the required minimum precision requirements for the selected kernel using an auto-tuning script based on binary search (see *IV-B*). Through emulation, it identifies the minimum bit sizes for both the exponent and mantissa for custom floating-point data types. Afterwards, ❸ we perform an accelerator design phase to individuate the best design optimization for the selected kernel and the precision requirements(see *IV-C*). Finally, ❹ we assess the accelerator performance with state-of-the-art systems (see *IV-D*).

A. APPLICATION PROFILING

We profile the most widely used and state-of-the-art radio-astronomical imager, WSClean [65], which also includes the state-of-the-art gridding and degridding algorithm (Image-Domain Gridding) [64], by evaluating the execution time breakdown. We report the software version used in this work in *Table 2*.

We employ the LOFARSCHOOL dataset, which is usually employed as a test case for practical examples and contains real sky observations [66]. This dataset contains 16 observations and around 30 subbands per observation, available in the LOFAR Long Term Archive (LTA) [67]. We select 14 observations with similar observation parameters such as integration time, observation duration, frequencies, etc. (see *Table 3*), and we select the 10th subbands from every dataset; therefore, all the datasets have the same central frequency.

TABLE 2. Software versions employed. We report the checksum of the commit of the master branch we used for WSClean and IDG. The other packages are WSClean's dependencies.

Software	Version
boost	1.68
OpenBLAS	3.9
python	3.8
wcslib	6.3
cfitsio	3.450
casacore [62]	3.3.0
dysco [63]	1.2
IDG [64]	master 011dfb18
WSClean [65]	master 2680c6a

TABLE 3. Datasets observation parameters.

Name	Description
Central Frequency	120.1172-125.7812 MHz
Channels per subband	4
Channel width	48 828.125 Hz
Declination	50.9410-54.8590
Duration	7199 s
Integration interval	2.002 78 s
Right Ascension	311.2500-318.7500

The WSClean parameters that we use are listed in *Table 4*. The sky images are plotted using Kstars FITS Viewer [68] setting the following parameters: shadows 0.0080, midtones 0.0625, and highlights 0.6009.

B. PRECISION AUTO-TUNING

To determine the application precision requirements, we employ binary search over the number of mantissa bits like [54], and over the number of exponent bits for custom precision floating-point data, as the execution time overhead (about 5-10x with respect to the optimized single-precision code) of emulation of custom data types in software makes evaluating the entire search space unpractical. As shown in *Figure 8* we manually instrument the application code to support the software emulation of `custom` data types. To emulate custom floating-point we employ a template header C++ libraries: FloatX [69] for floating-point [47]. We do not include fixed-point numbers since they would result in very long fixed-point representations based on the analysis results, which shows that a large dynamic range is required. This binary search algorithm first evaluates the mantissa size and then the exponent size. Since the application usually runs in single precision, the starting mantissa size of 23 bits is divided by two and evaluated. The process consists of first determining the size to evaluate and then updating the headers containing the mantissa and exponent sizes. Then, the application is compiled and run. Finally, the algorithm evaluates the output precision employing the Structural Similarity Index Measure (SSIM) [70] metric.

We select SSIM [70] as the assessment metric since, differently from Peak Signal to Noise Ration (PSNR), we can measure the perceived image quality and thus evaluate how two images are similar. SSIM computation is more complex than PSNR and it is shown in *Equation 4*.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2\mu_y^2 + C_1)(\sigma_x^2\sigma_y^2 + C_2)} \tag{4}$$

SSIM evaluates different windows images (x and y) using the window average (μ_x and μ_y), the window variance (σ_x and σ_y), the windows covariance (σ_{xy}). The remaining two variables (C_1 and C_2) are employed to avoid instability when $\mu_x^2 + \mu_y^2$ is close to zero. More precisely, these variables are obtained with the following formula $C_{1/2} = (k_{1_2}L_{1/2})^2$, where L is the pixel-values dynamic range k is a constant (usually $k_1 = 0.01$ and $k_2 = 0.03$).

For completeness, we describe the PSNR formula since we report it alongside SSIM. Like Mean Squared Error (MSE), PSNR is straightforward to compute and has specific physical meaning. PSNR computation is shown in *Equation 5*, where *MAX_I* is the maximum value of the original image and the MSE is calculated with *Equation 6*, where I is the original image and K is its approximate version.

$$PSNR = 20 \cdot log_{10}(\frac{MAX_I}{\sqrt{MSE}}) \tag{5}$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{6}$$

The auto-tuning algorithm gives as output the mantissa and exponent sizes that should be used to avoid noticeable precision loss. As threshold, we use 0.99 in order to have an approximate image as much similar to the original one [71]. We employ the precision tuning method with the overhead of running the application multiple times to detect the target precision. However, we use binary search because it has lower time complexity ($O(log_2N)$) compared to brute force algorithms such as linear search, which has time complexity in the order of $O(N)$, where N is the search space, which in this case is the sum of the number of mantissa and exponent bits.

C. ACCELERATOR DESIGN

Modern Intel FPGAs such as Intel Arria 10 and Stratix 10 have support for single-precision floating-point operations [9]. Their DSPs can perform Fused-Multiply-Accumulate (FMA), which counts as two floating-point operations. However, when employing fixed-point representation, they usually do not reduce the DSP usage [72]. Differently, Xilinx FPGAs have smaller DSPs [73], thus causing a larger use of DSPs for floating-point operations, e.g., 3 DSPs for a single multiplication. This property makes Xilinx FPGAs a worse candidate for single-precision application compared to Intel FPGAs, but, at the same time, an interesting candidate for exploring smaller data types that can be mapped on a smaller number of DSPs. Indeed, we target a Xilinx Alveo U50 [74] for deploying our accelerators. The FPGA mentioned above is a small form factor board with a large number of resources, PCIE3 connection, HBM2 memory, and a TDP (Thermal Design Power, commonly used to report peak power consumption) of 75W. The Alveo U50 is connected to a host system through a PCIE3 X16 connection as shown in *Figure 9*. Modern large FPGAs such as the Alveo U50 are built with multiple Super Logic Regions (SLRs). An SLR is a single FPGA die slice contained in an SSI (Stacked Silicon Interconnect) device [73].

We develop the accelerators into the Xilinx Vitis 2020.2 [75] tool flow, which is shown in *Figure 9*. Xilinx Vitis needs a source code with embedded OpenCL API to run on the host processor to schedule and control the execution of the accelerators. A similar approach is followed for the accelerators code: the source code containing HLS pragmas or optimization directives is compiled and linked by the Xilinx Vitis compiler. Unlike Intel, Xilinx FPGAs expose developers to a deeper level of optimization details, e.g., array partitioning and IP implementation with resource constraints. The kernel compilation step consists mainly of transforming the source code into HDL. At this stage, programmers can detect possible optimization opportunities and or stalls. The linking stage maps the accelerator on the FPGA by employing user configuration directives such as computing units and memory channel connections.

Arbitrary precision libraries generally provide support for fixed-precision computation, short integers, and custom floating-point formats. Unfortunately, such libraries are missing for the Xilinx Vitis OpenCL development environment [75]. The THLS library [60] implements templatized custom floating-point and is built on top of C++ and HLS. Therefore, we also use C++ with HLS pragmas to implement our kernels for Xilinx Vitis.

TABLE 4. WSClean parameters: the top part of the table shows parameters that exclusively depends on the observation and on the radio-telescope structure; the bottom part reports the CLEAN parameters employed for Cotton Schwab run. These parameters are the most commonly used. However, the CLEAN algorithm heavily depends on the user parameters. More complex CLEAN can be used for extracting the sky image, such as the Multiscale CLEAN.

Parameter	Value	Description (unit)
size	6000 6000	output x and y dimensions (pixels)
scale	5 asec	scale of a pixel (degrees)
use-idg	active	-
auto-threshold	3	CLEAN stop condition (sigma)
niter	50000	number of minor CLEAN iterations
mgain	0.85	gain per major CLEAN iteration
weight	briggs 0	weighting mode and robustness
taper	gaussian 2amin	

TABLE 5. Hardware employed for comparison.

Intel i9 9900k	8 cores, 2 threads per core, 4.0 GHz all cores, 16 MB L3 Cache, 64 GB DDR4 3600 MHz
NVIDIA GTX 750	512 CUDA cores, 1.14 GHz, 2 MB L2 Cache, 2 GB GDDR5
AMD RX 550	8 compute units, 1.09 GHz, 512 KB L2 Cache, 4 GB GDDR5
Xilinx Alveo U50	872 K LUTs, 1743 K Registers, 5952 DSPs, 8 GB HBM2

D. EVALUATION WITH STATE-OF-THE-ART ARCHITECTURES

Each evaluated architecture needs to be profiled with specific tools. While applications running on CPU can be profiled with a large number of profiling methods, we choose `perf` [76] since it is available in most Linux distributions and is easy to extract information such as floating-point operations count, DRAM memory traffic, and power consumption (it is usually not needed to have root-access). On the other hand, GPUs typically have proprietary tools. Indeed, we use NVIDIA nvprof [77] and AMD CodeXL [78] for profiling flops and DRAM accesses on the GTX 750 and the RX 550. We count the memory requirements and the number of floating-point operations placed for evaluating the FPGA performance. For measuring the power consumption on FPGA and GPUs, we extend libpowersensor [79].

We select CPU and GPU architectures with similar characteristics such as peak performance and TDP, which are reported in *Table 5* and *Table 6*). The NVIDIA GTX 1050 or GTX 1050 Ti GPU have a 14 nm lithography (see *Table 6*) and would have been the preferred choice for this comparison. However, due to incorrectly functioning power measurement hardware counters on these devices [83], we opted to use an NVIDIA GTX 750 GPU instead.

According to [84], we would expect, for the same chip size, an improvement of ~2x in terms of power consumption efficiency. Furthermore, the TDP values reported for the GPUs are the power-cap limits read in the system out-of-the-box. Indeed, these values are reduced compared to the limits advertised: 35 W instead of 50 W for GTX 750 and 38 W instead of 55 W for RX 550. While the peak bandwidths reported in *Table 6* are extracted from the device datasheets, the peak performance is computed by multiplying the device frequency and the number of operations that can be computed in parallel in a cycle for each unit. For instance, the Intel i9 9900k has 8 cores that run at 4 GHz; each core can compute 32 flops (we are considering FMA as two flops) per cycle. These values are verified by using synthetic benchmarks such as clpeak [85].

We apply a similar computation for evaluating the performance of the Alveo U50. As shown in [86], we compute the theoretical performance of the Alveo U50 considering the maximum number of FMA operations (one addition and multiplication [87]) that could be theoretically be placed on this FPGA. More precisely, we count 5 DSPs for each FMA operation, and we consider the available resources as reported in *Table 7*. Then this number is multiplied by the maximum of the frequency advertised by Xilinx for the single-precision addition and multiplication IP, which is 724 MHz [87].

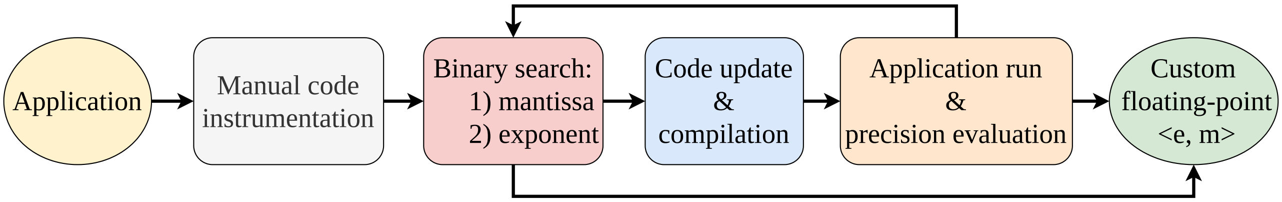

Application
Manual code instrumentation
Binary search: 1) mantissa 2) exponent
Code update & compilation
Application run & precision evaluation
Custom floating-point <e, m>

FIGURE 8. Binary search algorithm.

TABLE 6. Comparison of the employed architectures in terms of peak performance, bandwidth, TDP, energy efficiency (ratio between peak performance and TDP), and lithography process.

Architecture	Peak Performance	Bandwidth	TDP	Energy efficiency	Process
Intel i9 9900k	1.024 TFLOP/s	57.60 GB/s	95 W	10.79 GFLOP/W	14 nm Intel
NVIDIA GTX 1050 Ti	2.138 TFLOP/s	112.1 GB/s	75 W	28.50 GFLOP/W	14 nm Samsung [80]
NVIDIA GTX 750	1.164 TFLOP/s	80.19 GB/s	38 W	30.63 GFLOP/W	28 nm TSMC [81]
AMD RX 550	1.097 TFLOP/s	96.00 GB/s	35 W	31.34 GFLOP/W	14 nm GlobalFoundries [82]
Xilinx Alveo U50	**Peak Performance**	**Bandwidth**	**TDP**	**Energy efficiency**	**Process**
Theoretical (724 MHz)	1.547 TFLOP/s	316 GB/s	75 W	19.77 GFLOP/W	16 nm TSMC
Theoretical (300 MHz)	0.641 TFLOP/s	316 GB/s	75 W	8.55 GFLOP/W	16 nm TSMC
Empirical (292 MHz)	0.535 TFLOP/s	316 GB/s	75 W	6.84 GFLOP/W	16 nm TSMC

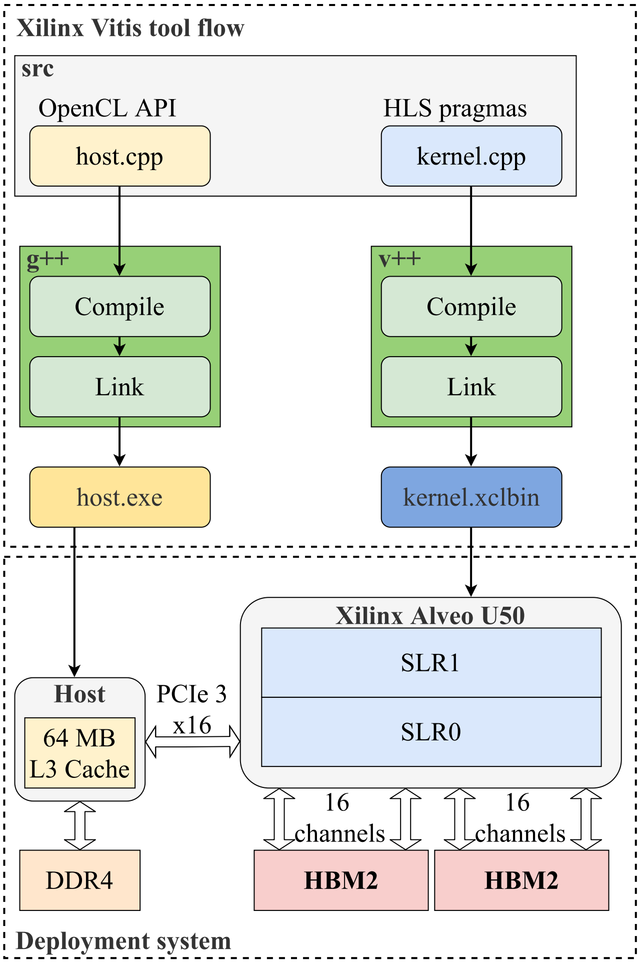


FIGURE 9. Representation of the Xilinx Vitis toolflow (top box) and how it relates to the deployment system (bottom box).

A theoretical peak of 1.547 TFLOP/s is obtained by multiplying this number by 2 since we consider the FMA as 2 flops. Since it is very difficult to achieve such frequencies as reported in [86], we employ the one Xilinx advertised for this FPGA: 300 MHz. This number (0.641 TFLOP/s) is considerably lower with respect to the theoretical one. However, this is still far from what can be realistically achieved on FPGA. Indeed, [86] shows that usually, a better upper bound is represented by employing 70% of the LUTs or 80% of the DSPs. For completeness, we also compute the peak performance using the FER (FPGA Empirical Roofline model) synthetic benchmark [86], and we obtain a value of 0.535 TFLOP/s, which is really close to what we achieve in one of our highly optimized single-precision accelerator prototypes.

V. ANALYSIS

To determine the bottleneck in the WSClean imager, we perform a bottleneck analysis (*V-A*). Then, we carry out a data types precision analysis (*V-B*) to understand the precision requirements for the identified bottleneck to be used for the accelerator design.

A. BOTTLENECK ANALYSIS

As shown in *Figure 10*, we first evaluate the execution time breakdown of the overall imaging pipeline for different datasets. The trend in the execution time breakdown is comparable for all datasets. More precisely, the most time-consuming step is inversion. Indeed, inversion needs to be run two times more than the prediction to compute the PSF and the dirty image. Typically, the deconvolution algorithm is the less critical phase.

B. DATA TYPES EXPLORATION

We evaluate custom precision floating-point data types employing software emulation since common architectures such as CPUs and GPUs usually support single- and half-precision floating points.

We notice a fundamental application property: since we are reducing the precision of the gridding kernel, it is sufficient to compare the dirty images instead of the cleaned image to evaluate the accuracy. The algorithm's intrinsic nature easily explains this: the dirty image is a sky image with added noise, and the CLEAN algorithm extracts the brightest sources at each iteration. Thus, each successive iteration needs an equal or smaller dynamic range. This feature helps evaluate the application requirements faster since we need just run the gridding algorithm once to generate the dirty image, avoiding running the degridding and any CLEAN iterations,

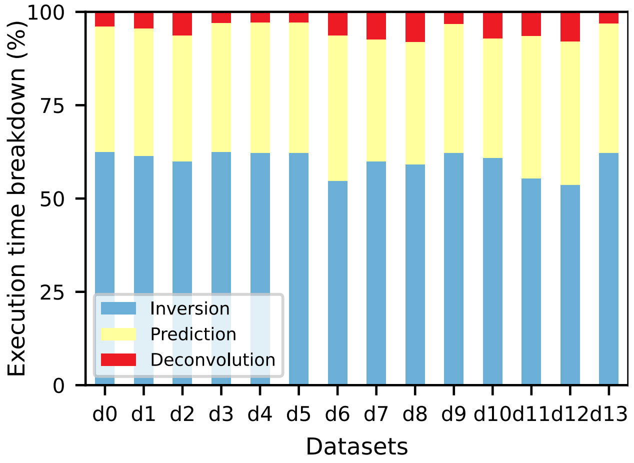


FIGURE 10. Execution time breakdown of WSClean for different datasets.

thus drastically reducing the analysis time. More precisely, we need to perform the gridding kernel once to generate the dirty image, once for the PSF, and two times during the CLEAN major iterations. Therefore, for this particular case, the analysis is about 4x faster than running the whole WSClean application.

We show in *Figure 11* how the reduced precision affects the radio-astronomical images. Noise effects can be noticed by using a mantissa of only 10 bits. Another important observation regards the relationship between dirty and clean images. In *Figure 11(a)* the images obtained emulating brain floating point contains a large quantity of noise, the same image cleaned (see *Figure 11(b)*) is empty. Therefore we can conclude that it is sufficient to analyze the dirty image to understand the precision requirements for the entire imager paying attention to the error indicators (SSIM).

In *Figure 12* we evaluate the accuracy of dirty images for different data-types for the gridding kernel in terms of SSIM and PSNR compared to the single-precision floating-point version. The precision requirements depend on the datasets employed, but the combination with 11 bits for the mantissa and 6 bits for the exponent can satisfy almost all the selected datasets. Furthermore, data types such as half-precision and fixed precision are not suitable for this field since the dynamic range is too small. Other data types with very small mantissa, such as brain floating point and NVIDIA Tensor Float, are not accurate enough to correctly represent all the faint features in the image. It is necessary to specify that the Tensor Float representation on NVIDIA GPU can only be used to perform warp matrix-to-matrix multiply and accumulate operations. In this work, we consider this data type representation for the whole kernel.

VI. CUSTOM PRECISION ACCELERATOR ARCHITECTURE

Optimizing the gridding algorithm on Xilinx FPGAs requires different steps. We first describe in *VI-A* the high-level structure of the accelerator by explaining the HLS optimization applied. Then, we explain the employed optimization for lookup tables and reduced precision respectively in *VI-B* and in *VI-C*. Finally, we discuss in *VI-D* all the different methods we explore for placing the accelerator on FPGA through Xilinx Vitis and the design points of our prototypes.

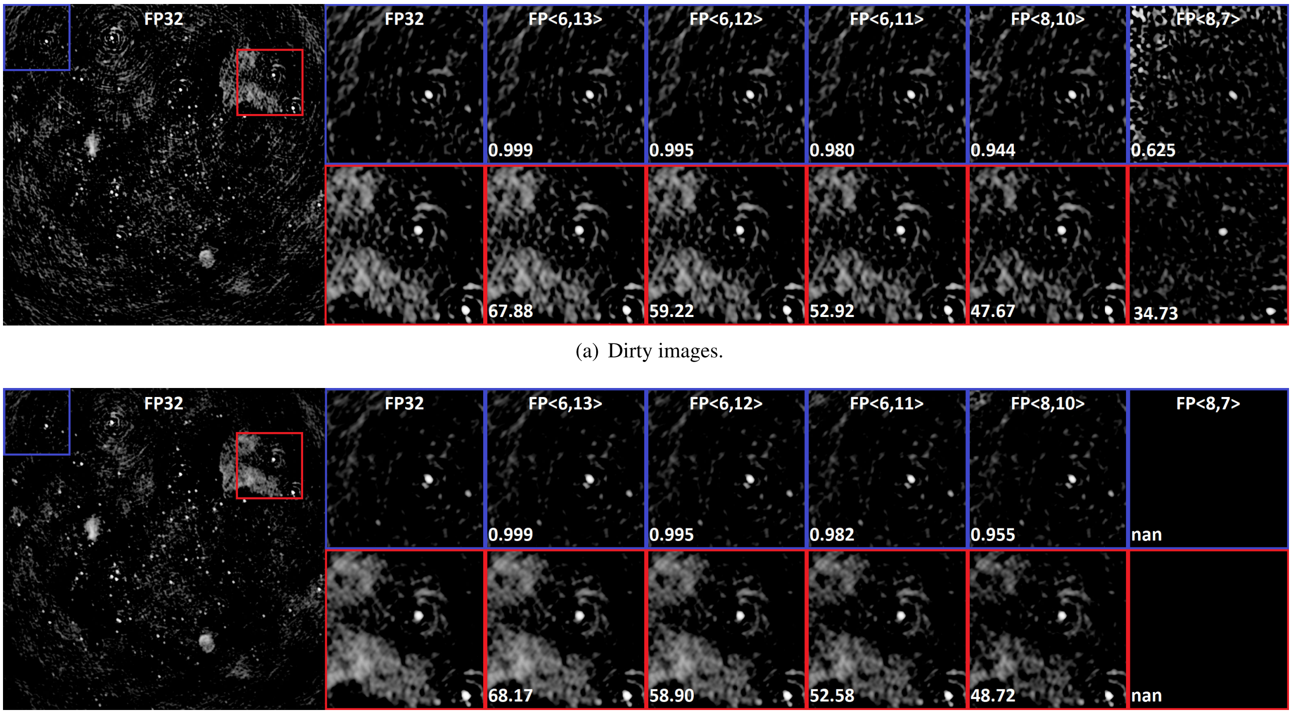
A. GRIDDING ACCELERATOR

We report the HLS optimized pseudocode of the subgrid computation in *Listing 1* and its high-level representation in *Figure 13*. We summarize the main optimization applied to get our highest performance prototypes:

Memory management: We first optimize the data accesses from the HBM2 memory by using multiple memory channels and widening the AXI (Advanced eXtensible Interface) width to 512 bits for the bus that transfers most of the data. More precisely, we use three channels per compute unit: the first one for the input of the main computation block (subgrid computation), the second one for the input of the post-processing pipeline (Aterms, tapering, reorder, and FFT), and the third one for the output subgrid (see *Figure 13*) The data is moved into local buffers to exploit reuse and improve the memory access latency.

Initiation interval of the subgrid computation: The accelerator described in HLS is implemented as a hardware pipeline where ideally, the pipeline is *stall-free*, and new data is fed into the pipeline every cycle. In this case, the *initiation interval* (II) is equal to 1. In the case of stalls, the II could be larger than 1. Depending on the design, it then takes several cycles for the operations on that data to complete. To achieve II = 1, we exchange the loop order (see *Algorithm 1*). The new loop order reduced the Read after Write (RaW) dependencies relative to the subgrid pixel update.

Parallelism: We increase the parallelism with respect to the code mentioned above at different levels. We first increase the parallelism of the subgrid computation by unrolling the channels and the pixels loops. The unrolling is implemented by employing larger local memories (BRAMs) and by unrolling the loops (see lines 5 and 7 in *Algorithm 1*) to increase the number of parallel floating-point units. Moreover, *Figure 14* reports the performance and the resource usage for different unrolling factors for channels and pixels. While the performance increases almost linearly, resources occupancy makes larger unrolling more efficient in terms of resource savings, e.g. especially for DSPs (see *Figure 14*). Indeed, unrolling the loops of a factor 2 do not imply the 2x more utilization of this resource. However, in order to be able to place more units and have a better area usage and placement, the parallelism should not be excessive, e.g., the case 4_8 is using more than 50% of resources, leaving no space for placing multiple units. Another significant observation concerns the relationship between unrolling channels over pixels: unrolling over the channels increases the BRAM usage and reduces the DSPs usage, which is the opposite behavior obtained by unrolling more over the pixels. This happens because unrolling over the pixels introduces more cosine and sine computation, which requires a larger amount of DSPs than other operations such as additions and multiplications.



(a) Dirty images.

(b) Clean images.

FIGURE 11. Comparison of (a) dirty images and (b) clean images with different data types precision. The numbers enclosed in the angular brackets represent the exponent and mantissa bits. FP<8,10> and FP<8,7> correspond respectively to NVIDIA Tensor and brain floating point (bfloat). While the dirty image a) is a noise image when using bfloat, the clean image in b) is completely dark. We can observe that the SSIM (and PSNR) does not vary significantly between dirty and clean images except for poor quality images. These values are for the whole image, and the blue and red squares are zoomed image sections. We can notice some visible differences for SSIM lower than 0.99.

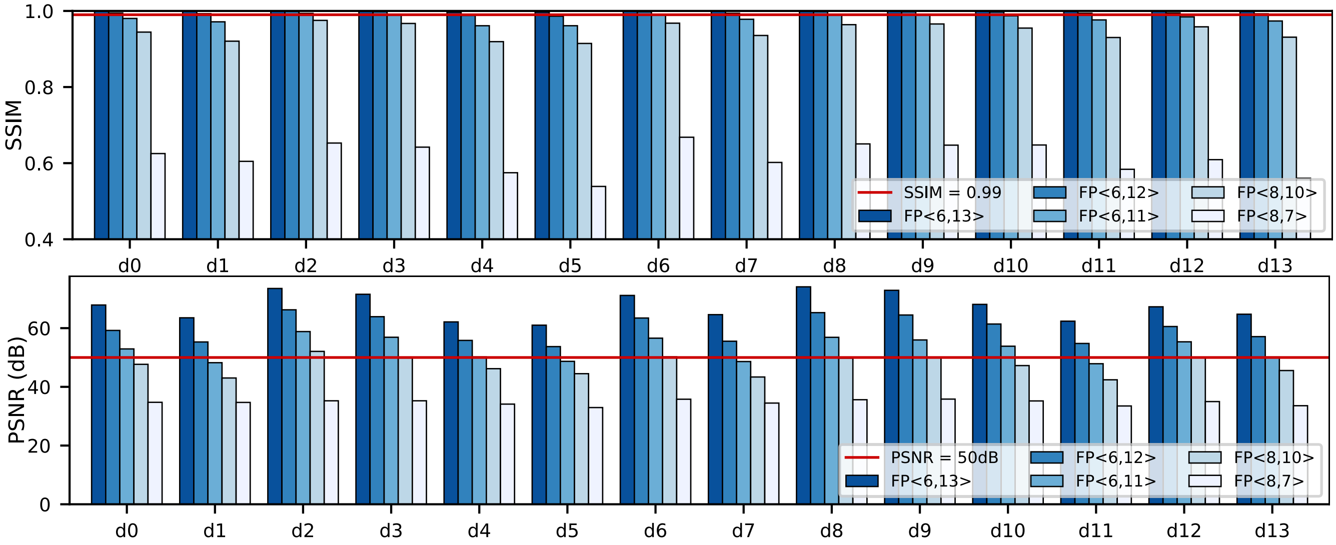


FIGURE 12. Precision accuracy for the selected 14 datasets (d0-d13) for different custom floating-point data types expressed in terms of SSIM and PSNR. The numbers enclosed in the angular brackets represent the exponent and mantissa bits. FP<8,10> and FP<8,7> correspond respectively to NVIDIA Tensor and brain floating point. High-resolution images closely similar to the original usually have an SSIM equal to 0.99 [71]. Given this threshold, reduced precision can produce high-quality images. However, it depends on the datasets. Indeed, in most cases, FP<6,11> or FP<6,12> reach the threshold, while smaller mantissa sizes do not. Such data types are not available on the standard accelerator platform such as GPU, which usually support half, single, double precision. Therefore a custom accelerator is needed. PSNR is just plotted for completeness, and the selected threshold of 50 dB is the maximum value for lossy images, and it serves only as a reference.

We conclude that the best trade-off is to have similar unrolling factors for channels and pixels to achieve balanced use of DSPs and BRAMs.

Then, we instantiate multiple subgrid computations (see *Figure 13*) that are run in parallel by using the `DATAFLOW` pragma [88]. Finally, we increase the parallelism over the

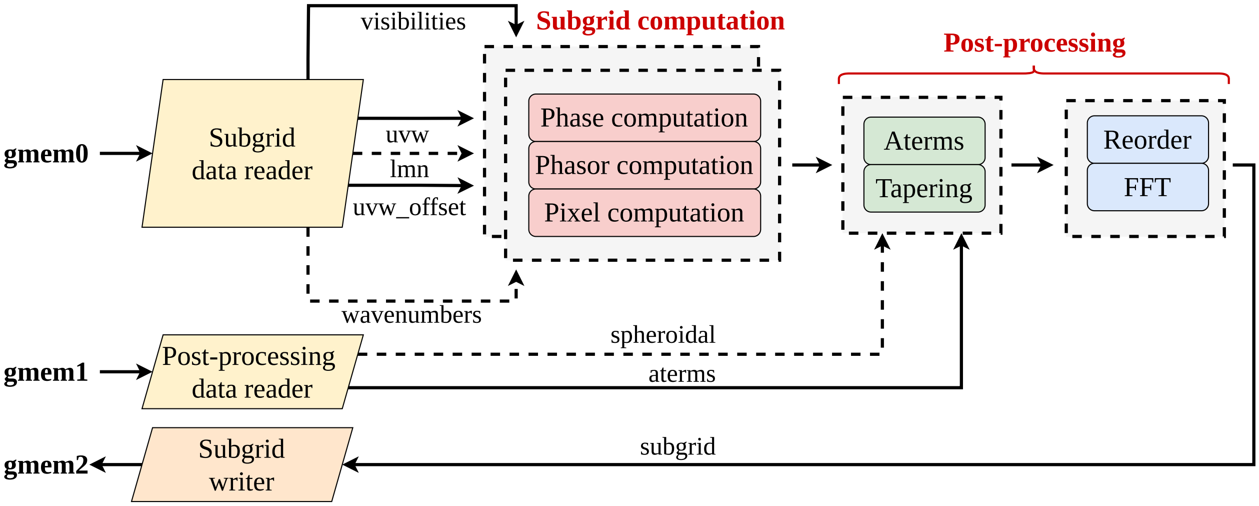


FIGURE 13. High-level scheme of one gridding compute unit: the data are read and written in parallel from different memory channels (gmem); the main part, subgrid computation, is replicated N times; finally, post-processing HW performs the Aterms, the tapering, and an FFT. Vectorized memory accesses are shown in black, while the scalar accesses are dashed.

number of instanced kernel units depending on the design time closure difficulty.

Post processing trade-offs: The post-processing computation consists of applying the A-terms, the tapering, and after a pixel reordering an FFT. As mentioned above, the subgrid computation is replicated `N` times, thus making it possible to have just a single post-processing unit that is capable of processing the data from each subgrid computation in a pipelined fashion. This will add a delay given by the execution time of the post-processing computation, but with the advantage of a constant throughput and reduced resource usage. Since the subgrid computation is much more time consuming compared to the post-processing computation, the latter is designed as cheap as possible to have just sufficient performance to balance the subgrid processing, e.g., initiation interval greater than 1 and low parallelism, and save as many resources as possible making the subgrid computation the only responsible for the resource mapping on the FPGA board. For completeness, we report that the post-processing section responsible for applying the Aterms and the tapering has an initiation interval of 2 cycles (internal computation, the data is read at II = 1 from the subgrid computation), which can be tuned up to 8 to facilitate the accelerator placing in some instances. We employ this trade off for the single-precision lookup table and custom floating-point design. The section responsible for the FFT has an initiation interval of 3 cycles.

B. COSINE/SINE LOOKUP TABLE AND REDUCED PRECISION

Similarly to [11], we employ a lookup table implementation to perform cosine and sine operations and save resources. Indeed, in Xilinx FPGAs, a `cosisin` operation, which computes the cosine and sine of a given angle, requires 11 DSPs compared with the 3 DSPs needed by the lookup implementation. We further reduce the DSPs usage to zero by not representing the phase in radiants. We move the multiplication used for the phase conversion in the outer loop and apply it when reading the `lmn` input data, which is a common factor when computing the phase offset and index and consequently the phase. The lookup table implementation consists in saving into BRAMs for pre-computed values for sine and cosine in the range of $[0; \frac{\pi}{4}]$. Then these values are used to compute the sine and cosine by employing the symmetry properties of trigonometric functions. Then, we have to tune the performance of the post-processing pipeline to let the accelerator achieve better frequencies (about 10 MHz higher).

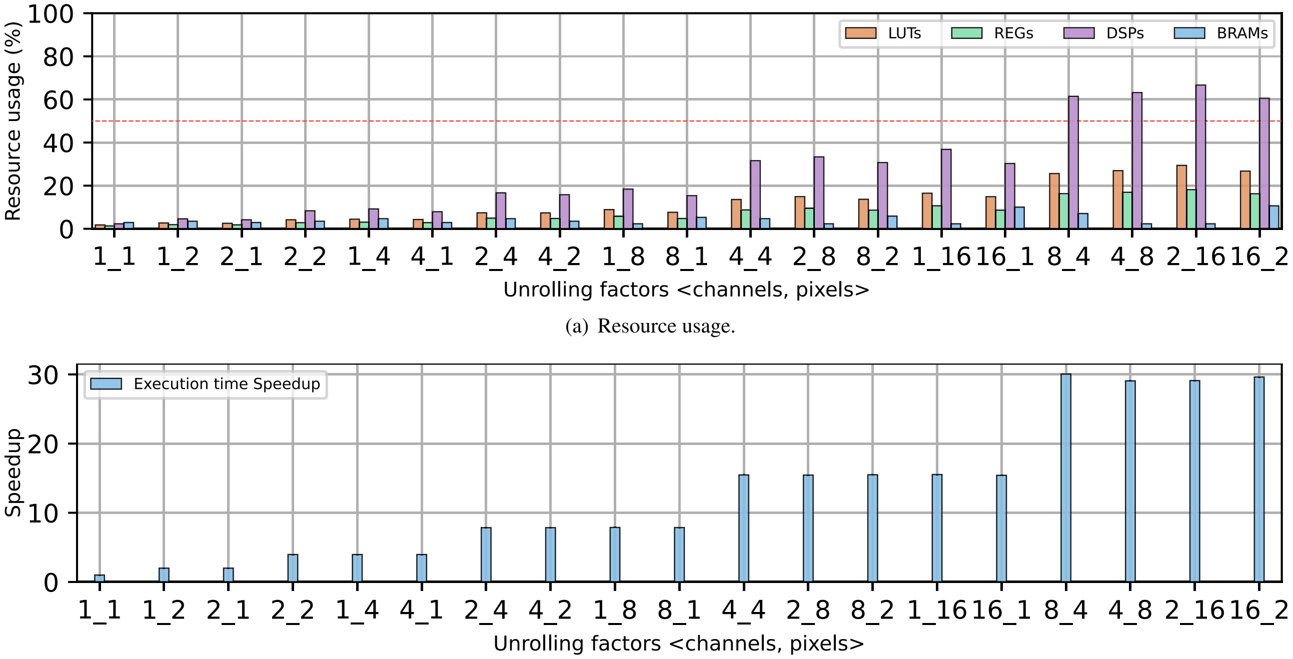
C. REDUCED PRECISION

The reduced-precision accelerators are obtained by employing the Templatized Floating-Point HLS library [60] for replacing the floating-point operations. Even if we select a global minimum precision for the gridding kernel (homogenous custom floating-point operations), we decide to employ this library instead of the CPFP [59] since it is the most resource-efficient (see *Appendix A*).

In order to have a portable accelerator, we use the same host-kernel interfaces as for the single-precision floating-point accelerators. Then, we add some conversion steps to the input (from single precision to custom precision) and to the output (from custom precision to single precision) of the subgrid computation. This design choice is supported by the fact that the application is purely compute bound, and the conversion does not affect the performance significantly. Since sine and cosine are not available in the library mentioned above, we adapt the lookup implementation used for single-precision by adding a custom floating-point round to integer method that is used to determine the index of the lookup table. As mentioned above, we have to tune the performance of the post-processing pipeline to achieve the accelerator placement at a reasonable frequency (greater than 250 MHz), which means getting a significant speedup. Indeed, when using

Algorithm 1 Subgrid Computation HLS Pseudocode

```
Input: visibilities, wavenumbers, uvw, uvw_offset, lmn
Result: subgrids
subgrids ⟵ 0;
for s in subgrids_per_cu do
    for t in timesteps do
        for c in channels do
            #pragma unroll factor = UNROLL_CHANNELS
            for p in pixels do
                #pragma unroll factor = UNROLL_PIXELS
                complex<float> pixel[pol]
                float lmn [3] ⟵ lmn[p]
                float phase_offset ⟵ compute_phase_offset(uvw_offsets, lmn)
                float phase_index ⟵ compute_phase_index(uvw, lmn)
                float phase ⟵ compute_phase(phase_index, phase_offset, wavenumbers)
                float phasor [2] ⟵ cosisin(phase)
                for pol in polarizations do
                    #pragma unroll
                    complex<float> pixel[pol] += visibilities[t][c][pol] * phasor
                end
            end
        end
    end
end
```



(a) Resource usage.

(b) Speedup.

FIGURE 14. Resource usage and speedup of the subgrid computation (not including the post-processing pipeline since it is not a critical computation) for different channels and pixels unrolling factors. The resources values refer to the kernel only, therefore without considering the overhead used by Xilinx Vitis, and are normalized by the maximum number of resources available in the device. The execution time is normalized by the largest values, which is the subgrid computation without unrolling (<1,1>).

more resources, the achieved frequency becomes considerably low, e.g., 200 MHz, thus not being beneficial to place multiple units.

D. DEVICE-SPECIFIC CONSIDERATIONS

After optimizing the high-level synthesis code, one of the main tasks to get the best performance from an FPGA device is the accelerator placement. While it is possible to let the Xilinx Vitis tool flow map the accelerator automatically on the device, it is more efficient in terms of achieved frequency and, consequently performance to fully customize the accelerator mapping. We report the most challenging tasks we face during the FPGA placement:

Super Logic Regions: *Figure 15* shows an example of two accelerators with two compute units each placed on the Xilinx Alveo U50. The Xilinx Alveo U50 consists of two Super Logic Regions (SLRs), and it is recommended to map an accelerator on a single SLR. Crossing two SLRs can cause a critical path even if the SLRs are connected with special registers that try to mitigate this problem. During the placement of our accelerators, we find out that each SLR is divided into two subregions [73], and in the middle of them, there is a region consisting of units responsible for managing the input clock. Traversing the two regions as shown in *Figure 15* can introduce critical paths that will negatively affect the maximum clock frequency of the accelerator. In our most resource-demanding design, the single-precision lookup table and the reduced-precision ones, we have to instantiate multiple units at different levels to meet timing requirements: 1) one accelerator/compute unit per SLR to avoid SLR crossing, and 2) multiple subgrid computation in each compute units to avoid clock region crossing.

Static region overhead: When mapping application by using HLS a static region (blue area in *Figure 15*) is flashed on the FPGA for supporting (accelerated) applications by using HLS (see *Table 7*). This static region has the task of managing interconnections such as AXI. It also makes the programmers able to only place the accelerator in the dynamic region, which results, as depicted by *Figure 15*, in two asymmetric SLR sub-regions, thus being important to place, for a large design, more compute units on the left side than on the right side. This observation leads to placing multiple compute elements, whenever possible, in the same accelerator to help the tool find a better hardware placement.

HBM2 memory channels: HBM2 memory can be employed efficiently for both memory-bound and compute-bound applications. In the first case, the large memory bandwidth will improve the application runtime by speeding up the memory transfers. In the second case, which coincides with our case, the multiple channels allow instantiating multiple accelerators that access independent memory spaces. Related to the HBM2 channels, there is another key issue: the Alveo U50 HBM2 memory channels are connected directly to the SLR0 [89], which means that additional logic is needed to transfer the data to the SLR1. When placing multiple accelerators and/or large sub-units that access the memory, connecting the AXI interfaces to distant memory channels is beneficial to avoid a long critical path between memory and processing. We carefully select the placement of HBM2 memory channels to avoid critical paths caused by area congestion, e.g., between two employed HBM2 channels, we decide to leave at least three channels unused.

TABLE 7. Xilinx Alveo U50 resources: total indicates the overall number of resources of the FPGA, while the dynamic region reports the available resources for accelerator deployment using Xilinx Vitis HLS, which is also reported in percentage.

Resources type	Total	Dynamic region	Available (%)
LUTs	872 K	731 K	83.83%
REGs	1743 K	1462 K	83.88%
DSPs	5952	5340	89.72%
BRAMs	1344	1128	83.93%
URAMs	640	608	95.00%

More precisely, the Alveo U50 has 32 HBM2 channels, of which only 28 are usable due to power budget limitations (the peak bandwidth of 316 GB/s can be achieved by employing 24 channels) [90]. As shown in *Figure 3* we use 3 HBM2 channels per compute unit. Moreover, we employ 6 channels in the lookup and reduced precision implementation because we instantiate two compute units. Furthermore, as shown in *Figure 16*, the application is not memory bound for any of the accelerator designs that we considered. Therefore, the bandwidth of a single HBM2 channel is sufficient to satisfy the bandwidth requirements for this application. The choice of using multiple HBM2 channels is determined by the mentioned above congestion area issues.

Vivado strategies [91]: Xilinx Vitis [75] is built on top of Xilinx Vivado, which is responsible for placing and routing the design on the FPGA board. Differently from the CPUs and GPUs programming model, the user can fully customize placement and routing strategies. The choice is between predefined implementation strategies or fully customizable strategies that reduce power consumption, area usage, improved performance, re-timing, etc. However, the main strategies are accessible by the user simply changing the Vitis Compiler optimization flags. We observe that both approaches lead to similar HW results. Indeed, when the frequency of the default Vivado strategy (-O0) is not able to meet the power constraints, we employ the PowerOpt strategy (-O1), e.g., single-precision lookup table, or when it is not able to meet the timing constraint, we use the ExtraTimingOpt (-O3), e.g., to achieve higher frequencies in the reduced-precision accelerators.

Pblock placement: This is a technique that can be applied after the design is implemented for guiding Xilinx Vivado towards a better design placement. With this option, the user can visually select from the GUI where to place certain units by creating a physical constrained region, the so-called pblock. This usually helps in cases of a desired higher frequency. However, we did not notice a significant improvement in our accelerators since they were already close to the highest advertised frequency.

Frequency overclock: Although Xilinx advertises 300 MHz as the maximum frequency for the HLS kernels, in reality, it is possible to achieve higher frequencies. Indeed, in [92] the authors match the same HBM2 frequency of

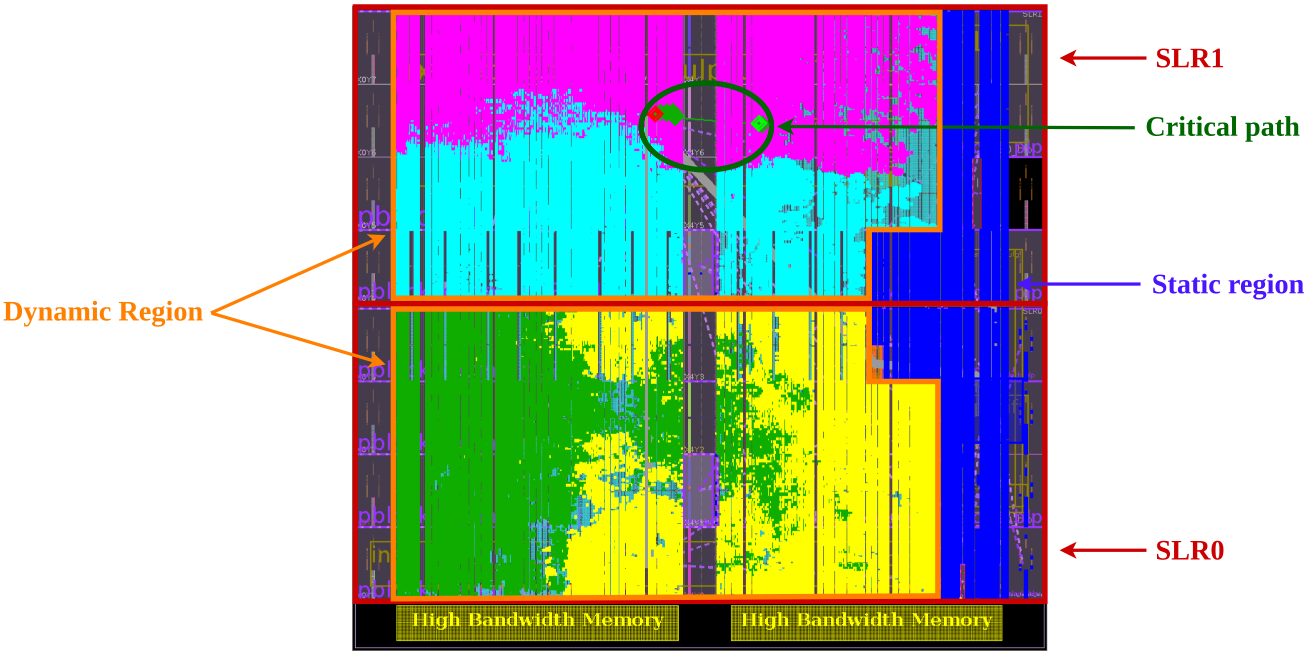


FIGURE 15. Example of an accelerator placement on Xilinx Alveo U50. In blue is reported the static region for deploying accelerators with Xilinx Vitis. The dynamic regions, which are the areas where the accelerators (yellow, purple, green, light blue) are mapped, are highlighted in orange. The FPGA is divided into multiple Super Logic Regions, FPGA die slices that compose large FPGA boards. Critical paths can usually occur when crossing multiple SLRs. However, critical paths can occur in the same SLR when connecting physically distant components and can be due to traversing the clock region like in the highlighted case.

450 MHz for small accelerators. However, this is not achievable for larger accelerators that usually reach frequency in the range of 200-300 MHz [86], [93]. We manage to increase the frequency of the baseline single-precision floating-point accelerator since the resource usage is not so close to the maximum, and it is able to reach the standard frequency (300 MHz) by just employing the Vivado Default implementation strategy.

E. DESIGN POINTS

Here we report the main design points of our accelerators:

Single-precision: Our baseline accelerator in single-precision (`FP32`) is obtained by simply using 2 subgrid computations with unrolling factor <4,4> and a single post-processing pipeline. In this case, due to the modest use of LUTs and FFs, it is not necessary to place the computation over different SLRs or apply any particular strategy. We are also able to achieve higher frequencies (`FP32_OC` in *Table 8*). Unfortunately, it is impossible to use more resources on this FPGA without violating timings (slow clocks that make the accelerator inefficient) or power constraints (high power budget).

Single-precision lookup-tables: The cosine/sine lookup table implementation can significantly reduce the number of DSPs employed. We notice only a small, but still significant reduction of DSPs since the main computation consists of FMA operations. We manage to improve the performance by placing 50% more computations. This is achieved by using an unroll factor of <2,4> and three subgrid computation units. This accelerator is then instantiated in each SLR, thus having two post-processing pipelines. In this case, to be able to place the accelerator, we have to reduce the number of DSPs by employing the `config_op` option during the HLS compilation to implement all the floating-point additions and multiplication with LUTs. To achieve better frequency, we employ the ExtraTiming_Opt strategy and reduced the post-processing computation's performance. We notice through analysis, a lookup table of 2048 is sufficient to keep the SSIM close to 1. The reported accelerator `FP32_LT` uses a lookup table with 2048 entries.

Reduced-precision: we manage to place 100% more computation with reduced precision compared to the single-precision baseline accelerator and 50% more than the single-precision lookup table implementation. This accelerator consists of 2 subgrid computations with unroll factors <4,4> placed in each SLR. To achieve better frequency, we employ the ExtraTiming_Opt strategy. For the reduced-precision prototypes, we use a lookup table with 2048 entries. In order to achieve the 300 MHz frequency, we employ the reduced performance post-processing computation. Higher frequencies do not meet timing constraints and the power budget (or TDP).

VII. EVALUATION AND DISCUSSION

In *VII-A* we report the area usage of the proposed accelerators, and we assess our accelerators performance in *VII-B* by employing the roofline model [95] and by measuring the throughput and the energy efficiency. Then, we highlight significant lessons learned during this work in *VII-C*.

TABLE 8. Resource utilization for the highest performance gridding accelerators. FP32_OC is the single-precision floating-point prototype at higher frequency and FP32_LT is the single-precision floating-point prototypes the lookup implementation.

Version	LUTs	FFs	DSPs	BRAMs	Frequency
FP32	434 k (49.88%)	604 k (34.72%)	4114 (69.12%)	454 (33.74%)	300 MHz
FP32_OC	435 k (49.99%)	640 k (36.78%)	4114 (69.12%)	454 (33.74%)	346 MHz
FP32_LT	649 k (74.58%)	614 k (35.29%)	3142 (52.71%)	1045 (77.75%)	296 MHz
FPX_6_11	642 k (74.81%)	754 k (43.33%)	1956 (32.81%)	818 (60.86%)	300 MHz
FPX_6_12	656 k (75.39%)	767 k (44.10%)	1956 (32.81%)	818 (60.86%)	300 MHz

A. AREA USAGE

In *Table 8* is presented the area usage of the highest performance accelerators here designed. More precisely, we report as FP32 the baseline single-precision floating-point accelerator, which can reach the advertised frequency of 300 MHz. Moreover, we show the same design with higher frequency (346 MHz) as FP32_OC. Resource usage does not vary significantly. Compared to [11] our DSP usage is significantly lower due to the mentioned above issues regarding the static region and timing closure.

The reported single-precision lookup-table (FP32_LT) accelerator has a higher overall resource usage because we manage to place more units compared to FP32. DSPs usage is lower since the lookup-table sine and cosine computation uses fewer DSPs with regards to the baseline design, and we implement addition and subtraction without DSPs. The achieved frequency of 295 MHz is close to the advertised one, which is difficult to achieve due to power-budget constraints.

The reduced-precision accelerators (FPX_6_11 and FPX_6_12) consume more resources than the baseline, but we can place two times more compute units. Furthermore, this design uses fewer resources than the baseline and consumes less power than the design with a single-precision lookup table. As already mentioned FPX_6_11 consumes slightly fewer resources (LUTs and FFs) than FPX_6_12 because of the smaller mantissa.

B. PERFORMANCE

We assess the performance of our accelerators against CPU and GPUs with similar peak performance and manufacturing technology. We first evaluate the performance achieved by each platform with the roofline model [95], [96] in *Figure 16*. The roofline model shows the performance obtained (TFLOP/s) and the analyzed kernel's arithmetic intensity (FLOP/Byte). The roofline defined in this way is not a measure of throughput but an indicator of how the application can be optimized for a certain architecture. Indeed, we report that only the NVIDIA GTX 750 can reach almost peak performance. This is because cosine and sine operations are offloaded to special units and indeed do not create bottlenecks. Differently, the AMD RX 550 does not have these special units, and the cosine and sine operations run at a quarter speed [97] compared to single-precision floating-point operations. We further notice that each cosine and sine function introduces three floating-point operations. Indeed, in *Figure 16* the RX 550 performance takes into account the largest number of instructions. We similarly reported the performance of the proposed accelerators showing how they are performing better than CPU. In particular, our best reduced-precision design is close to the AMD GPU in terms of TFLOP/s. Compared to [11] our baseline design has lower performance due to not having single-precision floating-point DSPs support.

Moreover, *Figure 16* shows that the application is compute-bound on all the architectures, and its high arithmetic intensity value highlights low memory bandwidth requirements. For instance, considering the RX 550, the roofline shows that the bandwidth required is lower than 3.7 GB/s. Similar results applies to the other architectures in *Figure 16*. Therefore, a single HBM2 channel (per compute unit) would satisfy the bandwidth requirement of the application on FPGA. Note, however, that we employed more channels to facilitate timing closure, as previously mentioned. Being compute bound can also be explained at the algorithm level: the application reads cachable data (e.g., visibilities). It produces the subgrid pixels by performing many operations over this cached data. Thus, the application would be memory bound only in the case of architecture with a lower DRAM bandwidth. Higher bandwidth requirements would be needed if next-generation FPGAs would enormously increase the computing capability, e.g., with more sophisticated DSPs.

In *Figure 17* we evaluate the throughput and the energy efficiency of the accelerators. More precisely, in *Figure 17(a)* we report the throughput in terms of Mega Visibilities per second (Mvis). *Figure 17(a)* evaluates the performance of our accelerators. All our designs have higher throughput compared to the i9 9900k up to 2.12x. Our best reduced-precision design is close to the AMD GPU performance. Moreover, the reduced-precision prototype is 1.84x faster than the single-precision baseline. However, the GTX 750 is the faster architecture due to their special function units mentioned above, and our baseline architecture is outperformed by the Intel counterpart proposed in [11] because of the single-precision DSPs support of Intel FPGAs.

As shown in *Figure 17(b) and 17(c)* our accelerators outperform up to 3.46x in terms of energy-efficient the CPU. The single-precision lookup table implementation reaches 88.97% of the empirical peak performance. Our best reduced-precision design is 2.03x more energy efficient than the single-precision baseline design. However, it is 78.77% and 63.29% less energy-efficient than AMD and

NVIDIA GPUs. We also observe that the AMD RX 550 is more energy-efficient than the NVIDIA GTX 750. This is mainly motivated by the different lithography technology. Scaling the performance of the NVIDIA GTX 750 chip to a 14 nm [84] the power consumption would be ∼2x reduced, thus being more efficient than the AMD counterpart.

Summarizing, the key observations are:

1) Our single-precision floating-point accelerators are more energy-efficient compared to CPUs with similar technology because of the better hardware utilization.
2) Reduced precision improves the accelerator's overall performance being much faster than CPU and achieving comparable performance compared to AMD GPUs thanks to the higher density of operations that we placed on the FPGA.
3) NVIDIA GPUs reach the highest percentage of peak performance exploited (∼88%) compared to the other architectures, especially against AMD GPUs (∼73%) thanks to the special units for sine and cosine. Overall, GPUs are the more energy-efficient architecture (see *Figure 16* and *17*) compared to CPUs and FPGAs with similar features due to their energy-efficient architecture (see *Table 6*).

C. LESSONS LEARNED

Radio-astronomical imaging applications usually employ single-precision or double-precision floating-point data types. We evaluate the use of reduced-precision data types for the gridding kernel and make the following observations:

Reduced precision applicability in radio-astronomical imaging: Different from artificial intelligence applications, where it is possible to highly reduce the data size, e.g. 1 or 8 bits [7], [101], radio-astronomical imaging needs higher precision for reconstructing sky images. Indeed, from our analysis, we observe that reduced precision can be applied in the state-of-the-art radio-astronomical imager. However, compared to AI tasks, the required precision is higher to avoid image artifacts. Tensor-Float floating-point numbers have a sufficiently high dynamic range for radio-astronomical imaging kernels, but the number of bits is too low to accurately represent the application values, such as visibilities and subgrids. Moreover, this format can only be used for specific warp matrix-to-matrix multiply and accumulate operations.

Benefits of applying reduced-precision in radio-astronomical imaging: Reduced precision is a well-known technique for improving performance and energy efficiency [12]. It is a technique that can be applied to compute-bound applications to reduce the compute unit size and memory-bound applications to decrease memory bandwidth requirements. We observe a significant improvement in GFLOP/s and energy efficiency, respectively 81.96% and 84.71% compared to the standard single-precision accelerator and 25.75% and 33.02% compared to the lookup-table implementation.

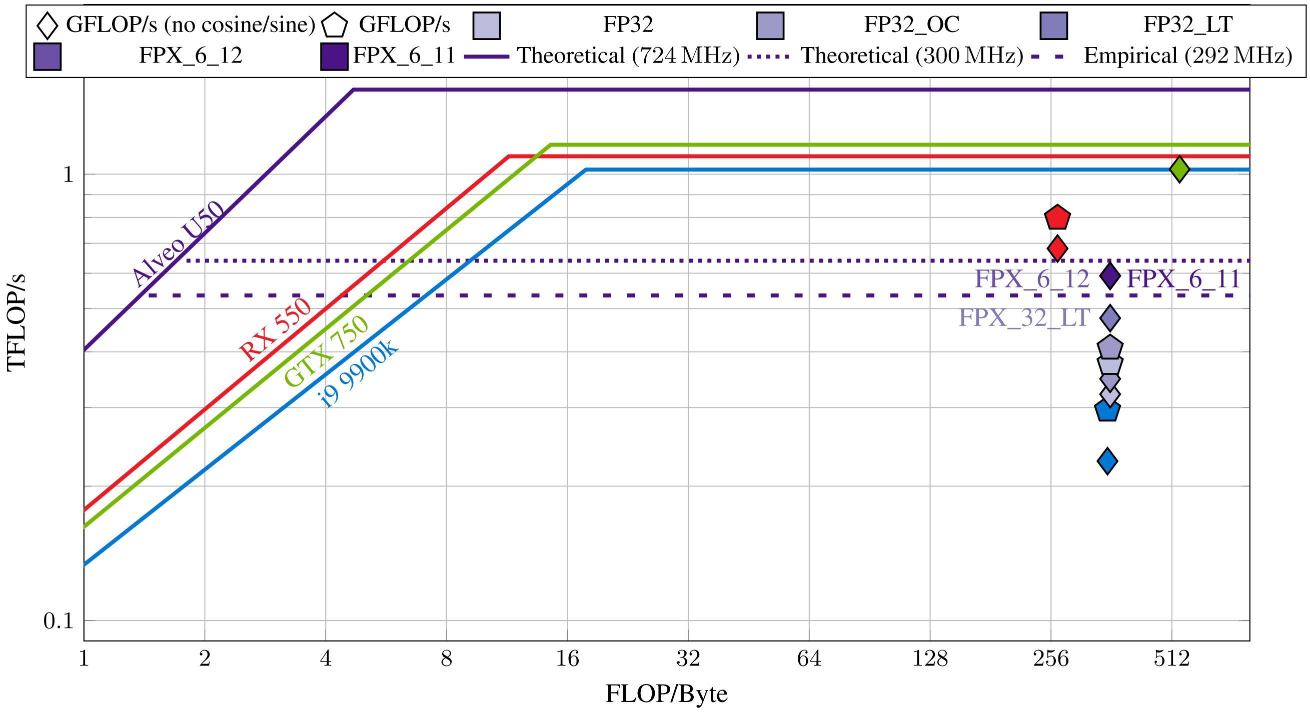
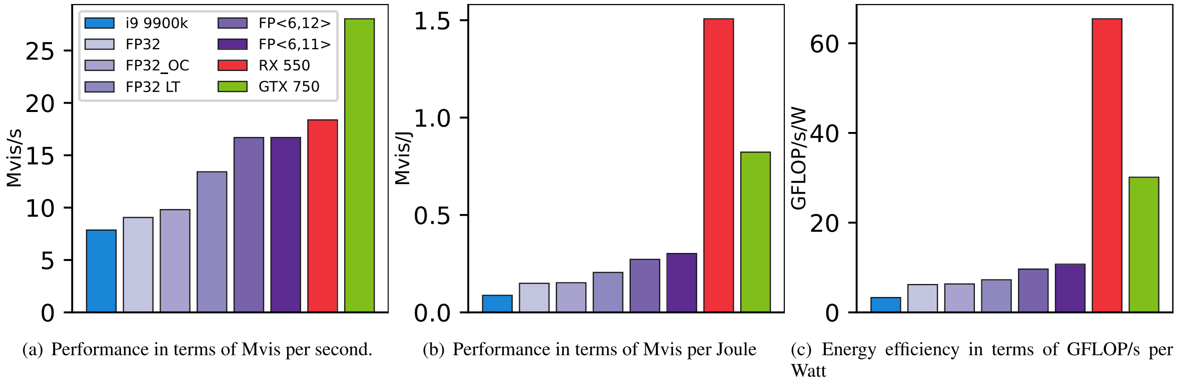


FIGURE 16. Roofline model of the gridding kernel mapped to different architectures. The pentagon shape represented the TFLOP/s achieved by each platform measured using performance counters. The diamond shape shows the TFLOP/s achieved without including sine and cosine since specific architectures such as the AMD RX 550 uses multiple floating-point instructions to compute sine and cosine compared to the NVIDIA GTX 750 that has special function units for transcendental math operations [94]. We show only one point for the GTX 750 and the lookup table accelerators (including the custom floating-point prototypes) since the sine and cosine operations are not executed as floating-point operations. Thus the two points have the same value. We report different horizontal roofs for the Xilinx Alveo U50 based on the discussion regarding peak performance in IV-D.



(a) Performance in terms of Mvis per second.

(b) Performance in terms of Mvis per Joule

(c) Energy efficiency in terms of GFLOP/s per Watt

FIGURE 17. Performance and energy efficiency evaluation on different architectures.

FPGAs vs CPUs and GPUs: The state of the art confirms that FPGAs and GPUs are more energy-efficient than CPUs except in rare cases [102]. The major part of past works shows that FPGAs are more energy-efficient than GPUs. However, these works compare GPUs with higher performance and power consumption. In this work, we show, as presented in [11], that for this particular application domain, GPUs with similar peak performance, thermal design power and manufacturing process are faster and have better energy efficiency. GPUs are also easier to program compared to FPGAs. Moreover, FPGA design requires hours of compilation to generate a single bitstream [103]. However, FPGAs have the flexibility to synthesize custom data types on hardware for assessing performance improvement against standard data types. In this work, we first evaluate the applicability of reduced precision for radio-astronomical imaging. Then, we design a reduced precision accelerator for radio-astronomical imaging on FPGAs reporting similar performance to GPUs without special sine/cosine units and improved performance and energy efficiency compared to its baseline.

Xilinx Alveo U50: We make the following observations four about the Xilinx Alveo U50 FPGA:

1) The static region consumes many resources (∼17% LUTs and ∼13% DSPs) reducing the maximum performance attainable with HLS accelerators.
2) The power budget and timing closure constraints, which can be improved with Vivado strategies and overclocking, make it impossible to fit more compute units (see *Table 8*) in the designs reaching a maximum of 69% of DSPs (77% counting the static-region effect).
3) Large accelerators are difficult to place as discussed in VI-D due to SLR placement. However, as previously discussed Xilinx FPGAs are appealing candidates due to their DSPs structure (see *IV-C*).
4) A positive feature in HBM-based FPGAs is the possibility to map AXI interfaces to a larger set of memory channels, alleviating congestion issues arising from small numbers of memory channels.

VIII. RELATED WORK

Related work on radio-astronomical imaging acceleration on modern computing systems is described in *VIII-A*; moreover, we report related work on accelerating domain-specific-application by using Xilinx Vitis in *VIII-B*.

A. RADIO-ASTRONOMY ACCELERATION

Over the past couple of years, hardware technologies have significantly improved, and a fair amount of research has been done on optimizing radio-astronomy algorithms to satisfy the Square Kilometre Array requirements [104].

The current state-of-the-art full imager, WSClean, is proposed by Offringa *et al.* [20]. It consists of an entire radio-astronomical imaging pipeline, including W-stacking, which is an extension of the previous gridding and degridding algorithm, the so-called W-projection, and different CLEAN algorithms such as Högbom, Cotton-Schwab and Multiscale CLEAN [35].

Veenboer and Romein [6], [98] optimize the Image-Domain Gridding [19], the current state-of-the-art fast algorithm for gridding and degridding for radio-astronomical imaging. They show how GPUs could reach almost peak performance and deliver better execution time and energy efficiency than CPUs. Image-Domain Gridding is now part of the WSClean imager. In [11] the authors also accelerate the gridding and degridding kernels on FPGA using a high-level-synthesis methodology based on OpenCL. Their FPGA implementation outperforms CPUs, but the GPU one is more energy efficient. Our single-precision accelerator baseline is outperformed by the one presented in [11] due to the single-precision floating-point DSPs support of Intel Arria FPGA [9]. However, we employ Xilinx FPGAs to assess the performance of a reduced precision accelerator for radio-astronomical imaging.

Hou *et al.* [100] implement an optimized prototype for the degridding algorithm on FPGA, outperforming both CPU and GPU by respectively 2.74x and 2.03x in terms of energy-delay product (EDP). However, they employ an outdated version of the degridding algorithm called W-projection, which does not reach the performance of IDG and does not include DDE corrections [28].

Corda *et al.* [4] focus on estimating the benefit of near-memory computing for huge images in IDG. They show that FFT is a critical bottleneck, which can be alleviated using architectures exploiting High-Bandwidth Memory.

TABLE 9. Radio-astronomy related work.

Work	Date	Application	Platform	Optimization
Offringa [20, 35]	2014-2017	W-Stacking, CLEAN (Högbom, Cotton-Schwab, Multiscale)	CPU	Optimized full imager (WSClean)
Veenboer [6, 98]	2017-2020	Image-Domain Gridding	CPU/GPU	code optimization (added to WSClean)
Grel [99]	2018	Högbom CLEAN	FPGA	Custom accelerator of Högbom CLEAN formulated as a Compressive Sensing problem
Veenboer [11]	2019	Image-Domain Gridding	FPGA	custom accelerator
Seznec [15]	2019	Generic deconvolution	GPU	half-precision deconvolution
Hou [100]	2020	W-Projection	FPGA	custom accelerator
Corda [4]	2020	Large 2D FFT	FPGA	NMC acceleration
This work	2021	WSClean (Image-Domain Gridding)	FPGA	Reduced precision analysis and acceleration

More precisely, they demonstrate how an FPGA design could reach a similar performance compared to a GPU with smaller memory and less memory bandwidth.

Seznec *et al.* [15] propose a simple deconvolution GPU implementation using half-precision data type. While half-precision floating-points speed up the algorithm significantly, this data type brings output degradation based on the dataset used. Furthermore, they focus on small images (2048 × 2048 pixels), while radio astronomical images are typically bigger, e.g., 5000 × 5000, 16000 × 16000 and larger [6].

Gürel *et al.* [99] formulate the Hogbom Clean as an Iterative Hard Thresholding (IHT), a compressive sensing technique, to reduce the data dimensions. While this work provides compelling insights, they optimize the most simple Clean algorithm and use very low resoluted images (256 × 256 pixels).

To the best of our knowledge, our work is the first demonstration in proposing a custom floating-point analysis and architecture for radio-astronomical imaging by focusing on the main bottleneck (gridding kernel).

B. XILINX VITIS

Recent high-level synthesis work on Xilinx FPGAs is based on Xilinx's new programming tool-flow: Xilinx Vitis. Brown [92] show the steps to take when optimizing an application for Xilinx FPGAs by employing Xilinx Vitis. Although some observations in their work have been helpful, they focus on very small FPGA designs (using about 2% of the available resources), while we try to put as many resources to good use as possible. Indeed, for a small accelerator, it is possible to reach a frequency of 450 MHz, while in a realistic use case, it turns out to be in a range of 200-300 MHz.

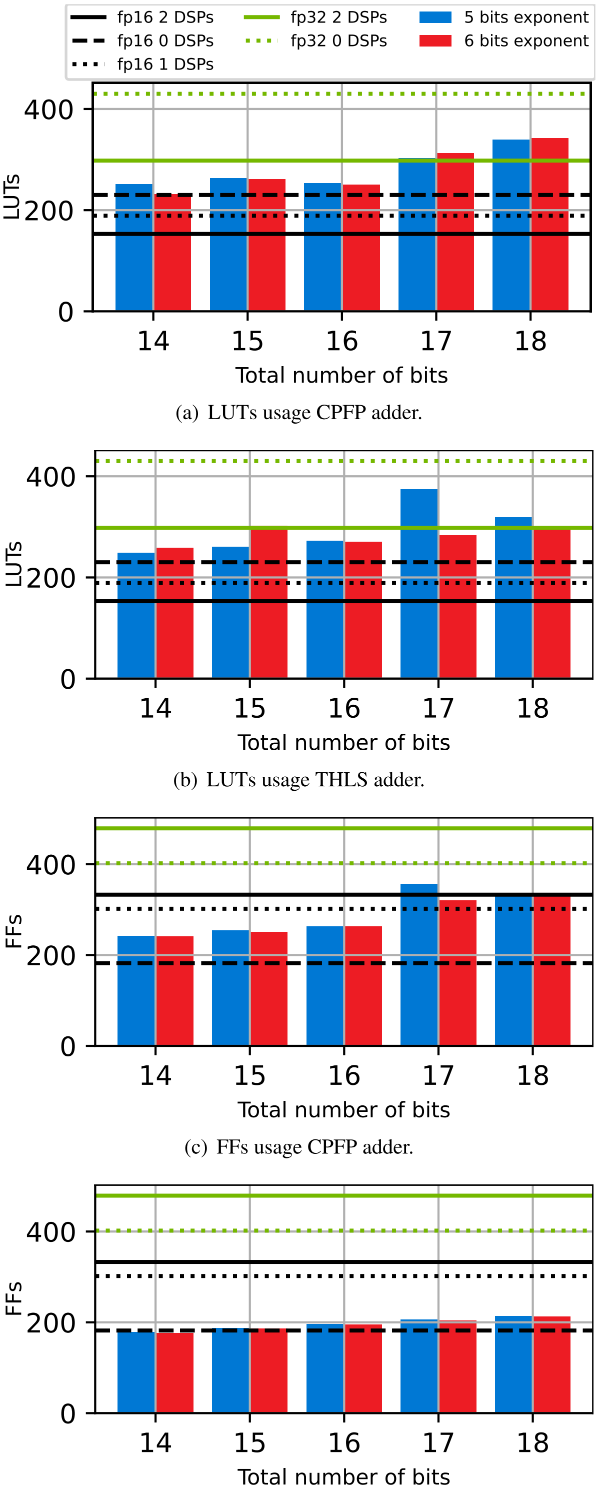
Calore and Schifano [86] benchmark the Alveo U250 with the FPGA Empirical Roofline model (FER), showing that there is a considerable difference between theoretical and attainable performance on FPGA, which usually is not so high in different architectures such as CPUs and GPUs. In our work, we employ the FER benchmark to determine the single-precision horizontal roof for the Alveo U50.

Nguyen *et al.* [103] evaluate FPGAs from different vendors with GPUs showing that modern GPUs are easier to program and have better energy efficient except for rare cases such as fixed-point precision computations.

Choi *et al.* [105] benchmark HBM-based FPGAs from different vendors to compare the memory performance. They present insightful observations regarding how the HBM2 memory channels must be mapped on Xilinx Vitis, e.g. Alveo FPGAs usually has 30 channels available over a total of 32. Related to this work, we notice that it is crucial to carefully select the HBM2 channel when implementing the accelerator to avoid critical paths that can significantly reduce the achieved frequency.

IX. CONCLUSION

We present the first reduced-precision custom floating-point analysis and accelerator for radio-astronomical imaging. More precisely, we evaluate the bottleneck of WSClean, the state-of-the-art radio-astronomical imager. Then, we analyze for the first time the impact of reduced precision on radio-astronomical imaging by employing custom floating-point for the main kernel, the so-called gridding, from the state-of-the-art Image-Domain Gridding algorithm. We demonstrate that reduced precision could be applied to radio-astronomical imaging, but data types must be selected carefully based on the dataset to avoid precision loss. Indeed, our analysis excludes standard low-precision data types supported in modern GPUs and highlights insightful observations regarding how the analysis can be evaluated in a shorter time. We port the gridding algorithm (from Image-Domain Gridding) on Xilinx FPGAs in single-precision floating-point to serve as a baseline. Moreover, we map the first reduced precision prototype of gridding kernel on the same hardware by employing custom floating-point data types. We find that the reduced precision accelerator is up to 1.84x faster and 2.03x more energy-efficient than its single-precision version. Compared to the CPU, our best accelerator prototype is 2.12x faster and 3.46x more energy-efficient than the CPU. However, it is 9.16% and 40.44% slower and 78.77% and 63.29% less energy-efficient than AMD and NVIDIA GPUs.



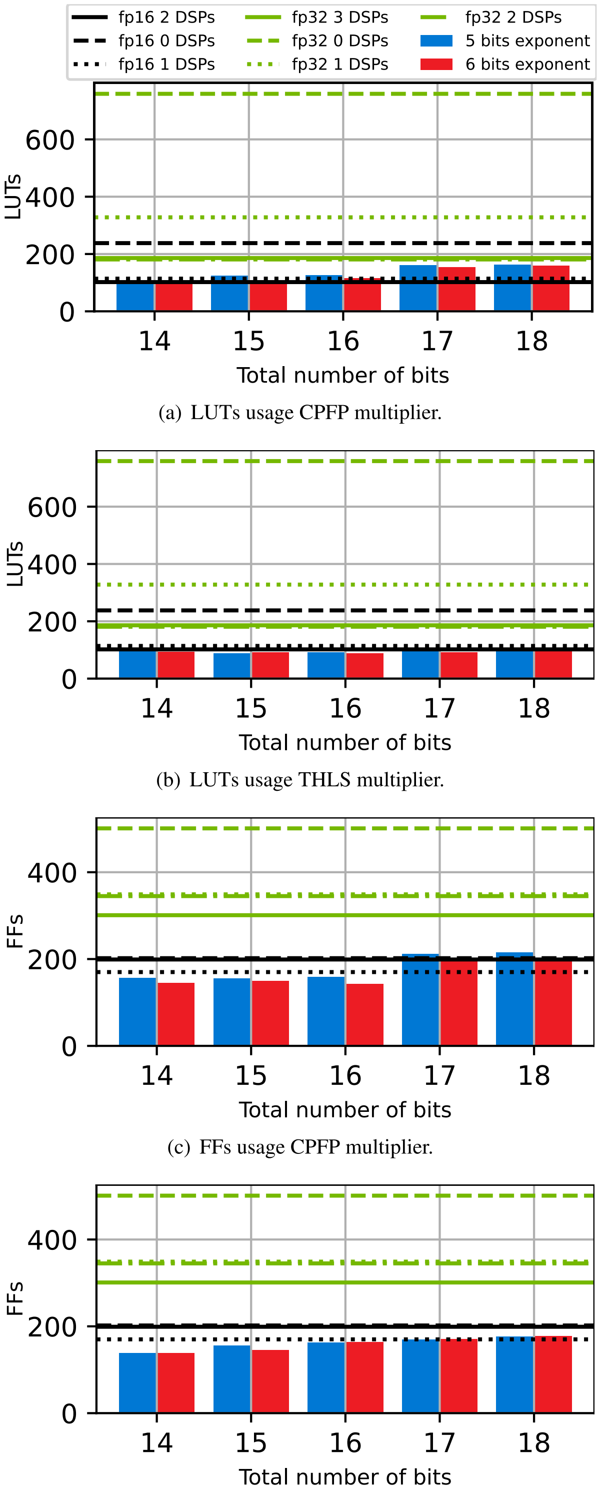
(a) LUTs usage CPFP adder.

(b) LUTs usage THLS adder.

(c) FFs usage CPFP adder.

(d) FFs usage THLS adder.

FIGURE 18. Custom floating-point adder (0 DSPs) resource usage for the CPFP and THLS libraries.



(a) LUTs usage CPFP multiplier.

(b) LUTs usage THLS multiplier.

(c) FFs usage CPFP multiplier.

(d) FFs usage THLS multiplier.

FIGURE 19. Custom floating-point multiplier (1 DSPs) resource usage for the CPFP and THLS libraries.

Thus we corroborate with the earlier observation that currently GPUs are the most energy-efficient architecture for radio-astronomical imaging, and custom precision support in the next generation of GPUs could further improve the performance/watt for radio astronomy imaging. Even though FPGA design support improved in the past years, FPGAs would need improved lithography technology [106], more DSPs with improved performance, smarter HLS compilers, and faster placement tools to be competitive against GPU in this application domain.

This work paves the way for future research in this specific application domain. Indeed, it would be interesting to explore reduced precision at runtime, e.g., by reconfiguring the accelerators based on dataset requirements and/or user parameters, to further improve performance. Fine-grained reduced precision, which consists of differentiating the data types for different instruction or kernel sections, would be a viable option for higher performance and energy efficiency in future works. The presented work also show that FPGA technology can be improved to make them more competitive compared to GPU in radio astronomy applications.

.

APPENDIX A HLS CUSTOM FLOATING-POINT LIBRARY COMPARISON

The main custom floating-point libraries for High-Level-Synthesis are the custom-precision floating-point (CPFP) [59], and the templatised soft floating-point for high-level synthesis (THLS) [60]. Both the libraries are based on the Xilinx arbitrary precision library [107]. While CPFP support homogeneous custom floating-point, THLS enables heterogeneous custom floating-point, which may be helpful for fine-grained mixed-precision or compute operations at higher precision and then truncate the result. Since we do not not need heterogeneous operators for this work the CPFP library would be sufficient. However,we evaluate the resource usage of the two libraries for the adder (see *Figure 18*) and multiplier operator (see *Figure 19*) to motivate our choice of THLS over CPFP.

Both the operators use less resources compared to single-precision floating-point for the presented precision, including the one employed in our highest performing accelerator (6 bits exponent and 11 bits mantissa, with a total of 18 bits). Unlike the adder operator, the multiplier operator, considering the same precision and number of DSPs (1), has similar resource usage (less in the case of THLS) than half-precision. Overall THLS has reduced LUTs, e.g. *Figure 19(b)* vs *Figure 19(a)*, and FFs, e.g. *Figure 19(d)* vs *Figure 19(c)*, usage compared to CPFP. This analysis thus motivates the usage of THLS over CPFP.

ACKNOWLEDGMENT

The authors would like to thank Remi Robert from Ericsson Research for reviewing the first draft.

REFERENCES

[1] CSP Consortium, "Square kilometre array phase 1 central signal processor software: An overview," in *Proc. 2nd URSI Atlantic Radio Sci. Meeting (AT-RASC)*, 2018, p. 1.

[2] R. V. van Nieuwpoort and J. W. Romein, "Correlating radio astronomy signals with many-core hardware," *Int. J. Parallel Program.*, vol. 39, no. 1, pp. 88–114, Jun. 2010, doi: 10.1007/s10766-010-0144-3.

[3] R. Bolton, "Parametric models of SDP compute requirements," SKA Office, Manchester, U.K., Tech. Rep. SKA-TEL-SDP-0000040, 2016, vol. 21.

[4] S. Corda, B. Veenboer, A. J. Awan, A. Kumar, R. Jordans, and H. Corporaal, "Near memory acceleration on high resolution radio astronomy imaging," in *Proc. 9th Medit. Conf. Embedded Comput. (MECO)*, Jun. 2020, pp. 1–6.

[5] R. Jongerius, S. Wijnholds, R. Nijboer, and H. Corporaal, "An end-to-end computing model for the square kilometre array," *Computer*, vol. 47, no. 9, pp. 48–54, Sep. 2014.

[6] B. Veenboer and J. W. Romein, "Radio-astronomical imaging on graphics processors," *Astron. Comput.*, vol. 32, Jul. 2020, Art. no. 100386. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2213133720300408

[7] C. Wu, M. Wang, X. Chu, K. Wang, and L. He, "Low precision floating-point arithmetic for high performance FPGA-based CNN acceleration," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 15, no. 1, pp. 1–21, 2021.

[8] J. G. Pandey, A. Karmakar, C. Shekhar, and S. Gurunarayanan, "An FPGA-based fixed-point architecture for binary logarithmic computation," in *Proc. IEEE 2nd Int. Conf. Image Inf. Process.*, Dec. 2013, pp. 383–388.

[9] Intel. (2020). *Enabling High-Performance Floating-Point Designs*. [Online]. Available: https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/wp-01267-fpgas-enable-high-performance-floating-point.pdf

[10] R. Nane, V.-M. Sima, C. Pilato, J. Choi, B. Fort, A. Canis, and Y. T. Chen, "A survey and evaluation of FPGA high-level synthesis tools," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 35, no. 10, pp. 1591–1604, Oct. 2016.

[11] B. Veenboer and J. W. Romein, "Radio-astronomical imaging: FPGAs vs GPUs," in *Parallel Processing*, R. Yahyapour, Ed. Cham, Switzerland: Springer, 2019, pp. 509–521.

[12] S. Cherubin and G. Agosta, "Tools for reduced precision computation: A survey," *ACM Comput. Surveys*, vol. 53, no. 2, pp. 1–35, Mar. 2021, doi: 10.1145/3381039.

[13] Z. Carmichael, H. F. Langroudi, C. Khazanov, J. Lillie, J. L. Gustafson, and D. Kudithipudi, "Performance-efficiency trade-off of low-precision numerical formats in deep neural networks," in *Proc. Conf. Next Gener. Arithmetic*, New York, NY, USA, Mar. 2019, pp. 1–9, doi: 10.1145/3316279.3316282.

[14] M. M. Thomas, K. Vaidyanathan, G. Liktor, and A. G. Forbes, "A reduced-precision network for image reconstruction," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–12, Dec. 2020, doi: 10.1145/3414685.3417786.

[15] M. Seznec, N. Gac, A. Ferrari, and F. Orieux, "A study on convolution using half-precision floating-point numbers on GPU for radio astronomy deconvolution," in *Proc. IEEE Int. Workshop Signal Process. Syst.*, Oct. 2018, pp. 170–175.

[16] J. W. Romein, "The tensor-core correlator," *Astron. Astrophys.*, vol. 656, p. A52, Dec. 2021.

[17] N. Doucet, H. Ltaief, D. Gratadour, and D. Keyes, "Mixed-precision tomographic reconstructor computations on hardware accelerators," in *Proc. IEEE/ACM 9th Workshop Irregular Appl., Archit. Algorithms (IA3)*, Nov. 2019, pp. 31–38.

[18] *UltraScale Architecture DSP Slice*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.xilinx.com/support/documentation/user_guides/ug579-ultrascale-dsp.pdf

[19] S. van der Tol, B. Veenboer, and A. R. Offringa, "Image domain gridding: A fast method for convolutional resampling of visibilities," *Astron. Astrophys.*, vol. 616, p. A27, Aug. 2018, doi: 10.1051/0004-6361/201832858.

[20] A. R. Offringa, B. McKinley, N. Hurley-Walker, F. H. Briggs, R. B. Wayth, D. L. Kaplan, M. E. Bell, L. Feng, A. R. Neben, J. D. Hughes, and J. Rhee, "WSCLEAN: An implementation of a fast, generic wide-field imager for radio astronomy," *Monthly Notices Roy. Astron. Soc.*, vol. 444, no. 1, pp. 606–619, Oct. 2014, doi: 10.1093/mnras/stu1368.

[21] M. P. van Haarlem, M. W. Wise, A. W. Gunst, G. Heald, J. P. McKean, J. W. T. Hessels, and A. Ger de Bruyn, "LOFAR: The low-frequency array," *Astron. Astrophys.*, vol. 556, p. A2, Aug. 2013, doi: 10.1051/0004-6361/201220873.

[22] M. G. Labate, R. Braun, P. Dewdney, M. Waterson, and J. Wagg, "SKA1-LOW: Design and scientific objectives," in *Proc. XXXIInd Gen. Assem. Sci. Symp. Int. Union Radio Sci.*, Aug. 2017, pp. 1–4.

[23] R. J. Selina, E. J. Murphy, M. McKinnon, A. Beasley, B. Butler, C. Carilli, and B. Clark, "The next-generation very large array: A technical overview," *Proc. SPIE*, vol. 10700, Jul. 2018, Art. no. 107001O.

[24] J. L. Jonas, "MeerKAT," in *Proc. URSI Asia–Pacific Radio Sci. Conf. (AP-RASC)*, Mar. 2019, p. 1.

[25] S. W. Schediwy, D. R. Gozzard, S. Stobie, C. Gravestock, T. Pulbrook, R. Whitaker, B. Alachkar, J. A. Malan, P. Boven, and K. Grainge, "Phase synchronization for the mid-frequency square kilometre array telescope," in *Proc. 2nd URSI Atlantic Radio Sci. Meeting (AT-RASC)*, May 2018, p. 1.

[26] O. M. Smirnov, "Revisiting the radio interferometer measurement equation: I. A full-sky Jones formalism," *Astron. Astrophys.*, vol. 527, p. A106, Mar. 2011, doi: 10.1051/0004-6361/201016082.

[27] U. Rau, S. Bhatnagar, M. A. Voronkov, and T. J. Cornwell, "Advances in calibration and imaging techniques in radio interferometry," *Proc. IEEE*, vol. 97, no. 8, pp. 1472–1481, Aug. 2009.

[28] A. R. Offringa, F. Mertens, S. van der Tol, B. Veenboer, B. K. Gehlot, L. V. E. Koopmans, and M. Mevius, "Precision requirements for interferometric gridding in the analysis of a 21 cm power spectrum," *Astron. Astrophys.*, vol. 631, p. A12, Oct. 2019, doi: 10.1051/0004-6361/201935722.

[29] T. J. Cornwell, K. Golap, and S. Bhatnagar, "The noncoplanar baselines effect in radio interferometry: The W-projection algorithm," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 5, pp. 647–657, Oct. 2008.

[30] S. Bhatnagar, T. J. Cornwell, K. Golap, and J. M. Uson, "Correcting direction-dependent gains in the deconvolution of radio interferometric images," *Astron. Astrophys.*, vol. 487, no. 1, pp. 419–429, Jun. 2008, doi: 10.1051/0004-6361:20079284.

[31] J. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astron. Astrophys. Suppl. Ser.*, vol. 15, p. 417, Jun. 1974.

[32] B. G. Clark, "An efficient implementation of the algorithm 'CLEAN'," *Astron. Astrophys.*, vol. 89, no. 3, p. 377, Sep. 1980.

[33] F. Schwab, "Relaxing the isoplanatism assumption in self-calibration; Applications to low-frequency radio interferometry," *Astron. J.*, vol. 89, pp. 1076–1081, Jul. 1984.

[34] T. J. Cornwell, "Multiscale CLEAN deconvolution of radio synthesis images," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 5, pp. 793–801, Oct. 2008, doi: 10.1109/JSTSP.2008.2006388.

[35] A. R. Offringa and O. Smirnov, "An optimized algorithm for multiscale wideband deconvolution of radio astronomical images," *Monthly Notices Roy. Astron. Soc.*, vol. 471, no. 1, pp. 301–316, Jun. 2017, doi: 10.1093/mnras/stx1547.

[36] J. W. Rich, W. J. G. de Blok, T. J. Cornwell, E. Brinks, F. Walter, I. Bagetakos, and R. C. Kennicutt, "Multi-scale CLEAN: A comparison of its performance against classical CLEAN in galaxies using THINGS," *Astron. J.*, vol. 136, no. 6, pp. 2897–2920, Nov. 2008, doi: 10.1088/0004-6256/136/6/2897.

[37] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "NVIDIA A100 tensor core GPU: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, Mar. 2021.

[38] L. Polok and P. Smrz, "Increasing double precision throughput on NVIDIA Maxwell GPUs," in *Proc. 24th High Perform. Comput. Symp.* San Diego, CA, USA: Soc. Comput. Simul. Int., 2016, pp. 1–8, doi: 10.22360/SpringSim.2016.HPC.032.

[39] *IEEE Standard for Binary Floating-Point Arithmetic*, ANSI/IEEE Standard 754-1985, I. C. Society, 1985, pp. 1–20.

[40] *IEEE Standard for Binary Floating-Point Arithmetic (Revision 2008)*, ANSI/IEEE S 754-1985, 2008, pp. 1–20.

[41] G. Tagliavini, S. Mach, D. Rossi, A. Marongiu, and L. Benini, "A transprecision floating-point platform for ultra-low power computing," in *Proc. Des., Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 1051–1056.

[42] A. V. Cueva. (2020). *Code Sample: Intel Deep Learning Boost New Deep Learning Instruction BFloat16—Intrinsic Functions*. [Online]. Available: https://software.intel.com/content/www/us/en/develop/articles/intel-deep-learning-boost-new-instruction-bfloat16.html?wapkw=bfloat16

[43] S. Wang and P. Kanwar. (2020). *BFloat16: The secret to high performance on Cloud TPUs*. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus

[44] S. Kim, K.-I. Kum, and W. Sung, "Fixed-point optimization utility for C and C++ based digital signal processing programs," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 45, no. 11, pp. 1455–1464, Nov. 1998.

[45] M. Nguyen, N. Serafin, and J. C. Hoe, "Partial reconfiguration for design optimization," in *Proc. 30th Int. Conf. Field-Program. Log. Appl. (FPL)*, Aug. 2020, pp. 328–334.

[46] L. Qiao, G. Luo, W. Zhang, and M. Jiang, "FPGA acceleration of ray-based iterative algorithm for 3D low-dose CT reconstruction," in *Proc. 30th Int. Conf. Field-Program. Log. Appl. (FPL)*, Aug. 2020, pp. 98–102.

[47] G. Flegar, F. Scheidegger, V. Novaković, G. Mariani, A. E. Tomás, A. C. I. Malossi, and E. S. Quintana-Ortí, "FloatX: A C++ library for customized floating-point arithmetic," *ACM Trans. Math. Softw.*, vol. 45, no. 4, pp. 1–23, Dec. 2019, doi: 10.1145/3368086.

[48] J. L. Gustafson and I. T. Yonemoto, "Beating floating point at its own game: Posit arithmetic," *Supercomput. Frontiers Innov.*, vol. 4, no. 2, pp. 71–86, 2017, doi: 10.14529/jsfi170206.

[49] L. Forget, Y. Uguen, and F. De Dinechin, "Hardware cost evaluation of the posit number system," in *Proc. Conf. Informatique Parallélisme, Archit. Syst.*, Anglet, France, Jun. 2019, pp. 1–7. [Online]. Available: https://hal.inria.fr/hal-02131982

[50] E. Theodore L. Omtzigt, P. Gottschling, M. Seligman, and W. Zorn, "Universal numbers library: Design and implementation of a high-performance reproducible number systems library," 2020, *arXiv:2012.11011*.

[51] *Universal: A Header-Only C++ Template Library for Universal Number Arithmetic*. Accessed: Nov. 30, 2021. [Online]. Available: https://github.com/stillwater-sc/universal

[52] W.-F. Chiang, M. Baranowski, I. Briggs, A. Solovyev, G. Gopalakrishnan, and Z. Rakamarić, "Rigorous floating-point mixed-precision tuning," *ACM SIGPLAN Notices*, vol. 52, no. 1, pp. 300–315, May 2017, doi: 10.1145/3093333.3009846.

[53] M. O. Lam and J. K. Hollingsworth, "Fine-grained floating-point precision analysis," *Int. J. High Perform. Comput. Appl.*, vol. 32, no. 2, pp. 231–245, Mar. 2018.

[54] A. Borghesi, G. Tagliavini, M. Lombardi, L. Benini, and M. Milano, "Combining learning and optimization for transprecision computing," in *Proc. 17th ACM Int. Conf. Comput. Frontiers*, New York, NY, USA, May 2020, pp. 10–18, doi: 10.1145/3387902.3392615.

[55] S. Cherubin, D. Cattaneo, M. Chiari, A. D. Bello, and G. Agosta, "TAFFO: Tuning assistant for floating to fixed point optimization," *IEEE Embedded Syst. Lett.*, vol. 12, no. 1, pp. 5–8, Mar. 2020.

[56] R. DiCecco, L. Sun, and P. Chow, "FPGA-based training of convolutional neural networks with a reduced precision floating-point library," in *Proc. Int. Conf. Field Program. Technol. (ICFPT)*, Dec. 2017, pp. 239–242.

[57] D. B. Thomas, "Templatised soft floating-point for high-level synthesis," in *Proc. IEEE 27th Annu. Int. Symp. Field-Program. Custom Comput. Mach. (FCCM)*, Apr. 2019, pp. 227–235.

[58] D. B. Thomas, "Compile-time generation of custom-precision floating-point IP using HLS tools," in *Proc. IEEE 26th Symp. Comput. Arithmetic*, Jun. 2019, pp. 192–193.

[59] *HLS Custom-Precision Floating-Point Library*. Accessed: Nov. 30, 2021. [Online]. Available: https://github.com/dicecco1/fpga_cpfp

[60] D. B. Thomas. *Templatised Soft Floating-Point for High-Level Synthesis*. [Online]. Available: https://github.com/template-hls/template-hls-float

[61] F. de Dinechin, "Reflections on 10 years of FloPoCo," in *Proc. IEEE 26th Symp. Comput. Arithmetic*, Jun. 2019, pp. 187–189.

[62] *Casacore*. Accessed: Nov. 30, 2021. [Online]. Available: https://github.com/casacore/casacore

[63] *Dysco*. Accessed: Nov. 30, 2021. [Online]. Available: https://github.com/aroffringa/dysco

[64] *Image Domain Gridding*. Accessed: Nov. 30, 2021. [Online]. Available: https://git.astron.nl/RD/idg

[65] *WSClean*. Accessed: Nov. 30, 2021. [Online]. Available: https://gitlab.com/aroffringa/wsclean

[66] *ASTRON*. Accessed: Nov. 30, 2021. [Online]. Available: https://support.astron.nl/LOFARImagingCookbook/tutorial.html

[67] *LOFAR Long Term Archive*. Accessed: Nov. 30, 2021. [Online]. Available: https://lta.lofar.eu/

[68] *Kstars*. Accessed: Nov. 30, 2021. [Online]. Available: https://docs.kde.org/trunk5/en/extragear-edu/kstars/tool-fitsviewer.html

[69] *FloatX*. Accessed: Nov. 30, 2021. [Online]. Available: https://github.com/oprecomp/FloatX

[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[71] O. S. Faragallah, H. El-Hoseny, W. El-Shafai, W. A. El-Rahman, H. S. El-Sayed, E.-S.-M. El-Rabaie, F. E. A. El-Samie, and G. G. N. Geweid, "A comprehensive survey analysis for present solutions of medical image fusion and future directions," *IEEE Access*, vol. 9, pp. 11358–11371, 2021.

[72] Intel. *Intel Stratix 10 Variable Precision DSP Blocks Overview*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/wp-01267-fpgas-enable-high-performance-floating-point.pdf

[73] Xilinx. *Large FPGA Methodology Guide*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.xilinx.com/support/documentation/sw_manuals/xilinx14_7/ug872_largefpga.pdf

[74] *Xilinx Alveo U50*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.xilinx.com/products/boards-and-kits/alveo/u50.html

[75] *Vitis Unified Software Development Platform 2020.2 Documentation*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.xilinx.com/html_docs/xilinx2020_2/vitis_doc/devckernels.html

[76] B. Gregg. *Performance Counters*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.brendangregg.com/perf.html

[77] NVIDIA. *CUDA Toolkit Documentation—Nvprof*. Accessed: Nov. 30, 2021. [Online]. Available: https://docs.nvidia.com/cuda/profiler-users-guide/index.html

[78] AMD. *AMD CodeXL*. Accessed: Nov. 30, 2021. [Online]. Available: https://github.com/GPUOpen-Archive/CodeXL

[79] *PowerSensor Library*. Accessed: Nov. 30, 2021. [Online]. Available: https://gitlab.com/astron-misc/libpowersensor

[80] *Samsung*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.samsung.com/us/sas/Business/Foundry14nm

[81] *TSMC*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.tsmc.com/english

[82] *GlobalFoundries*. Accessed: Nov. 30, 2021. [Online]. Available: https://gf.com/

[83] *NVIDIA 387.12 Breaks Power Reading in NVIDIA-SMI*. Accessed: Nov. 30, 2021. [Online]. Available: https://forums.developer.nvidia.com/t/nvidia-387-12-breaks-power-reading-in-nvidia-smi/53946/21

[84] S. Sarangi and B. Baas, "DeepScaleTool: A tool for the accurate estimation of technology scaling in the deep-submicron era," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2021, pp. 1–5.

[85] *Clpeak*. Accessed: Nov. 30, 2021. [Online]. Available: https://github.com/krrishnarraj/clpeak

[86] E. Calore and S. F. Schifano, "Performance assessment of FPGAs as HPC accelerators using the FPGA empirical roofline," in *Proc. 31st Int. Conf. Field-Program. Log. Appl. (FPL)*, Aug. 2021, pp. 83–90.

[87] Xilinx. *Performance and Resource Utilization for Floating-Point V7.1*. Accessed: Nov. 30, 2021. [Online]. Available: https://china.xilinx.com/html_docs/ip_docs/pru_files/floating-point.html

[88] *HLS Pragmas*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.xilinx.com/html_docs/xilinx2021_1/vitis_doc/hls_pragmas.html#sxx1504034358866

[89] Y. Hu, Y. Du, E. Ustun, and Z. Zhang, "GraphLily: Accelerating graph linear algebra on HBM-equipped FPGAs," *Power*, vol. 268, no. 182, p. 49, 2021.

[90] *75220—Alveo U50—Limited Number of HBM2 Pseudo-Channel Ports*. Accessed: Nov. 30, 2021. [Online]. Available: https://support.xilinx.com/s/article/75220?language=en_US

[91] G. Daughtry. *Top 5 Timing Closure Techniques*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.xilinx.com/publications/prod_mktg/club_vivado/presentation-2015/paris/Xilinx-TimingClosure.pdf

[92] N. Brown, "Weighing up the new kid on the block: Impressions of using Vitis for HPC software development," in *Proc. 30th Int. Conf. Field-Program. Log. Appl. (FPL)*, Aug. 2020, pp. 335–340.

[93] N. Zhang, X. Chen, and N. Kapre, "RapidLayout: Fast hard block placement of FPGA-optimized systolic arrays using evolutionary algorithms," in *Proc. 30th Int. Conf. Field-Program. Log. Appl.*, Aug. 2020, pp. 145–152.

[94] A. Li, S. L. Song, M. Wijtvliet, A. Kumar, and H. Corporaal, "SFU-driven transparent approximation acceleration on GPUs," in *Proc. Int. Conf. Supercomput.*, New York, NY, USA, Jun. 2016, pp. 1–14, doi: 10.1145/2925426.2926255.

[95] S. Williams, A. Waterman, and D. Patterson, "Roofline: An insightful visual performance model for multicore architectures," *Commun. ACM*, vol. 52, no. 4, pp. 65–76, 2009.

[96] G. Ofenbeck, R. Steinmann, V. Caparros, D. G. Spampinato, and M. Puschel, "Applying the roofline model," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw.*, Mar. 2014, pp. 76–85.

[97] *AMD RDNA 2 Performance Guide—GPUOpen*, GPUOpen, Adv. Micro Devices, Santa Clara, CA, USA, 2021.

[98] B. Veenboer, M. Petschow, and J. W. Romein, "Image-domain gridding on graphics processors," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. (IPDPS)*, May 2017, pp. 545–554.

[99] N. M. Gurel, K. Kara, A. Stojanov, T. Smith, T. Lemmin, D. Alistarh, M. Puschel, and C. Zhang, "Compressive sensing using iterative hard thresholding with low precision data representation: Theory and applications," *IEEE Trans. Signal Process.*, vol. 68, pp. 4268–4282, 2020.

[100] J. Hou, Y. Zhu, S. Du, S. Song, and Y. Song, "FPGA-based scale-out prototyping of degridding algorithm for accelerating square kilometre array telescope data processing," *IEEE Access*, vol. 8, pp. 15586–15597, 2020.

[101] P. Colangelo, N. Nasiri, E. Nurvitadhi, A. Mishra, M. Margala, and K. Nealis, "Exploration of low numeric precision deep learning inference using Intel FPGAs," in *Proc. IEEE 26th Annu. Int. Symp. Field-Program. Custom Comput. Mach. (FCCM)*, Apr. 2018, pp. 73–80.

[102] S. Daghaghi, N. Meisburger, M. Zhao, Y. Wu, S. Gobriel, C. Tai, and A. Shrivastava, "Accelerating SLIDE deep learning on modern CPUs: Vectorization, quantizations, memory optimizations, and more," in *Proc. Mach. Learn. Syst.*, 2021, pp. 156–166.

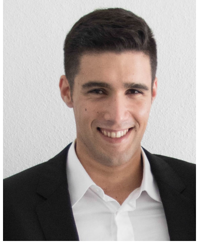
[103] T. Nguyen, S. Williams, M. Siracusa, C. MacLean, D. Doerfler, and N. J. Wright, "The performance and energy efficiency potential of FPGAs in scientific computing," in *Proc. IEEE/ACM Perform. Modeling, Benchmarking Simulation High Perform. Comput. Syst. (PMBS)*, Nov. 2020, pp. 8–19.

[104] E. Vermij, L. Fiorin, R. Jongerius, C. Hagleitner, and K. Bertels, "Challenges in exascale radio astronomy: Can the SKA ride the technology wave?" *Int. J. High Perform. Comput. Appl.*, vol. 29, no. 1, pp. 37–50, Feb. 2015, doi: 10.1177/1094342014549059.

[105] Y.-K. Choi, Y. Chi, W. Qiao, N. Samardzic, and J. Cong, "HBM connect: high-performance HLS interconnect for FPGA HBM," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, New York, NY, USA, Feb. 2021, pp. 116–126, doi: 10.1145/3431920.3439301.

[106] B. Gaide, D. Gaitonde, C. Ravishankar, and T. Bauer, "Xilinx adaptive compute acceleration platform: Versal TM architecture," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, New York, NY, USA, Feb. 2019, pp. 84–93, doi: 10.1145/3289602.3293906.

[107] *Xilinx's Arbitrary Precision Data Types Library*. Accessed: Nov. 30, 2021. [Online]. Available: https://www.xilinx.com/html_docs/xilinx2020_2/vitis_doc/ogi1585574074269.html



STEFANO CORDA received the B.Sc. degree in electrical and electronic engineering and the M.Sc. degree in electronic engineering from the University of Cagliari, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Electronic Systems Group, Department of Electrical Engineering, Eindhoven University of Technology. He is also a Guest Ph.D. Student at the Chair of Processor Design, Dresden Technical University. During his Ph.D., he visited IBM Zürich and Dresden Technical University as a Guest Researcher. His research interests include high-performance computing, FPGA design, and application profiling.

BRAM VEENBOER received the Ph.D. degree in computer science from Vrije Universiteit, Amsterdam, with a focus on radio-astronomical imaging using accelerator hardware. He is currently a Researcher with the HPC Group, Innovation and Systems Department, ASTRON, Netherlands Institute for Radio Astronomy. His current research interests include high-performance imaging for the low frequency array (LOFAR) and the upcoming square kilometre array (SKA), the development of novel algorithms (for imaging, but also for time-domain radio-astronomy), and the development of a real-time processing pipeline for all-sky imaging.

AHSAN JAVED AWAN received the joint master's degree from TU Kaiserslautern, Germany, the University of Southampton, U.K., and NTNU, Norway, and the joint Ph.D. degree from the KTH Royal Institute of Technology, Stockholm, Sweden, and the Universitat Politècnica de Catalunya, Barcelona, Spain, with a focus on performance characterization and optimization of in-memory data analytics on a scale-up server. He is currently a Senior Researcher of future computing platforms at Ericsson Research. Prior to joining Ericsson in 2018, he worked at Imperial College London, IBM Research—Tokyo, Japan, and the Barcelona Supercomputing Center, Spain. He holds the Erasmus Mundus Scholarship for his master's and Ph.D. degrees. He has written more than 20 scientific papers (including two best paper awards), filed 14 patent applications, and served as a TPC member of international conferences.

JOHN W. ROMEIN received the Ph.D. degree in computer science from Vrije Universiteit, Amsterdam, with a focus on distributed game-tree search. He is currently a Researcher with ASTRON, where he leads several projects on HPC research for radio-astronomical applications. His research interests include accelerated computing, using GPUs, FPGAs, and other accelerators.

ROEL JORDANS is currently an Assistant Professor with the Electronic Systems Group, Electrical Engineering Department, Eindhoven University of Technology (TU/e), and a Research and Development Manager with the Radboud Radio Laboratory, Radboud University Nijmegen. His research interests include developing hardware architectures for reliable, high-performance systems, and compilation techniques for and design automation applied to application-specific processors.

AKASH KUMAR (Senior Member, IEEE) received the joint Ph.D. degree in electrical engineering and embedded systems from the Eindhoven University of Technology, Eindhoven, The Netherlands, and the National University of Singapore (NUS), Singapore, in 2009. From 2009 to 2015, he was with NUS. He is currently a Professor with Technische Universität Dresden, Dresden, Germany, where he is directing the Chair for processor design. His current research interests include the design, analysis, and resource management of low-power and fault-tolerant embedded multiprocessor systems.

ALBERT-JAN BOONSTRA received the M.Sc. degree in applied physics from Groningen University, The Netherlands, in 1987, and the Ph.D. degree from the Delft University of Technology, The Netherlands, in 2005. He started his career at SRON Groningen, where he conducted cryogenic optical and infrared tests on the SWS spectrometer of the infrared astronomical satellite ISO. In 1992, he joined ASTRON, Netherlands Institute for Radio Astronomy, Dwingeloo. He coordinated the upgrade of the westerbork synthesis radio telescope (WSRT), after which he investigated temporal and spatial filtering techniques for suppressing radio interference at WSRT. From 2012 to 2017, he led the Dome Project at ASTRON, aimed at developing exascale technologies for the square kilometre array. He has led several groups and projects at ASTRON. He is currently a Program Manager of technical research at the Innovation and Systems Department, ASTRON. His research interests include signal processing, scientific computing, and artificial intelligence.

HENK CORPORAAL received the M.Sc. degree in theoretical physics from the University of Groningen, and the Ph.D. degree in electrical engineering (in the area of computer architecture) from the Delft University of Technology. He is currently a Professor in embedded system architectures with the Eindhoven University of Technology (TU/e), The Netherlands. He has coauthored over 500 journals and conference papers. Furthermore, he invented a new class of VLIW architectures, the transport triggered architectures, which is used in several commercial products, and by many research groups. He also initatiated and leads the Dutch NWO Perspectief Program on efficient deep learning in this program many research institutes and over 30 companies participate. His research interests include low power multi-processor, heterogenous processing architectures, their programmability, and the predictable design of soft- and hard real-time systems. This includes research and design of embedded system architectures, including CGRAs, SIMD, VLIW, and GPUs, on accelerators, the exploitation of all kinds of parallelism, fault-tolerance, approximate computing, architectures for machine and deep learning, optimizations and mapping of deep learning networks, and the (semi-) automated mapping of applications to these architectures. (For further details see corporaal.org).