Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **486,** 5052–5060 (2019) doi:10.1093/mnras/stz1206
Advance Access publication 2019 May 2

A real-time, all-sky, high time resolution, direct imager for the long wavelength array

James Kent [1]★ Jayce Dowell [2] Adam Beardsley,[3] Nithyanandan Thyagarajan,[3,4]† Greg Taylor[2] and Judd Bowman[3]

[1]*Cavendish Laboratory, University of Cambridge, Cambridge, UK*
[2]*Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM, USA*
[3]*School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA*
[4]*National Radio Astronomy Observatory, Socorro, NM, USA*

Accepted 2019 April 20. Received 2019 April 15; in original form 2019 February 2

ABSTRACT

The future of radio astronomy will require instruments with large collecting areas for higher sensitivity, wide fields of view for faster survey speeds, and efficient computing and data rates relative to current capabilities. We describe the first successful deployment of the E-field Parallel Imaging Correlator (EPIC) on the LWA station in Sevilleta, New Mexico, USA (LWA-SV). EPIC is a solution to the computational problem of large interferometers. By gridding and spatially Fourier transforming channelized electric fields from the antennas in real time, EPIC removes the explicit cross-multiplication of all pairs of antenna voltages to synthesize an aperture, reducing the computational scaling from $\mathcal{O}(n_a^2)$ to $\mathcal{O}(n_g \log_2 n_g)$, where n_a is the number of antennas and n_g is the number of grid points. Not only does this save computational costs for dense arrays but it produces very high time resolution images in real time. The GPU-based implementation uses existing LWA-SV hardware and the high performance streaming framework, Bifrost. We examine the practical details of the EPIC deployment and verify the imaging performance by detecting a meteor impact on the atmosphere using continuous all-sky imaging at 50 ms time resolution.

Key words: instrumentation: interferometers – techniques: interferometric – telescopes.

1 INTRODUCTION

Radio astronomy has been undergoing a renaissance in recent years, with a number of new instruments, both built and in the proposal and design phases. Future instruments such as the Square Kilometre Array (SKA; Dewdney et al. 2009), and current instruments such as the Long Wavelength Array (LWA; Taylor et al. 2012), Canadian Hydrogen Intensity Mapping Experiment (CHIME; The CHIME/FRB Collaboration 2018), and the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017), are looking at using high-density interferometric arrays with hundreds or thousands of individual antennas to facilitate wide-field, high-sensitivity, and angular resolution imaging of the sky.

There has also been a renewed focus on observations of transient phenomena such as Fast Radio Bursts (FRBs), where the origins and physical mechanisms are an active area of research. Therefore the capability to detect and image these in real time is of key scientific importance. Interferometric measurements of FRBs have been previously achieved (Caleb et al. 2017), including by CHIME (Amiri et al. 2019). High time resolution imaging of such phenomena would provide a significant new capability, by allowing dragnet surveys of the sky with wide field-of-view instruments.

Together, these two developments present a significant computational challenge for future interferometers, especially for the correlator. The standard FX correlator, where the signal from each antenna is multiplied with the signals from every other antenna to produce 'visibilities', mathematically defined as an outer product, scales as $\mathcal{O}(n_a^2)$, where n_a is the number of antennas (Romney 1985). This scaling becomes problematic as proposed arrays will contain thousands of dipole elements. All n_a^2 visibilities must be generated at the time resolution desired and subsequently gridded and then Fourier transformed to produce images, typically creating a bottleneck for high time resolution studies.

For some array geometries, the number of visibilities calculated can be reduced by omitting short baselines with little impact on point source imaging performance. Fast convolution algorithms may also be used for correlation (Bunton 2011) to further reduce the computational costs to $\mathcal{O}(n_a^{3/2})$.

★ E-mail: jck42@cam.ac.uk
† Nithyanandan Thyagarajan is a Jansky Fellow of the National Radio Astronomy Observatory.

© 2019 The Author(s)
Published by Oxford University Press on behalf of the Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

An alternative to full-field imaging with an FX correlator is to use a beamformer that provides the telescope's response to only a few chosen locations on the sky by summing over the voltages from all antennas with appropriate delays to direct the response in a particular direction. The computational costs of a beamformer generally scale as $\mathcal{O}(n_a)$ per calculated beam and the output data volume is proportional only to the number of beams calculated. This avoids the challenges of full-field imaging with an FX correlator, but with an associated compromise of limited sky coverage.

Direct Fourier transform imagers (Daishido et al. 1991; Foster et al. 2014) provide another alternative to both of the above approaches. Direct imaging forgoes the calculation of antenna cross-products. Instead, the antenna electric fields are gridded directly on to an aperture plane and Fourier transformed into an image plane. These images can be accumulated for noise reduction, in the same way visibilities are accumulated in FX correlators.

Theoretically they can provide significant potential scaling improvement by scaling as $\mathcal{O}(n_g \log n_g)$ where n_g is the number of grid points in the aperture, yielding a significant potential scaling advantage for high-density arrays (Morales 2011; Thyagarajan et al. 2017). Direct imagers facilitate full-field imaging at a high time resolution also because the output data volume can be much lower than for an FX correlator, scaling only as $n_g \approx n_a$ for a dense array.

Previous direct imagers such as Daishido et al. (1991) and Foster et al. (2014) have relied on antennas being on a regular grid, which limits their application from a scientific standpoint. For example, their uniform layouts yield point spread functions (PSFs) that contain periodic grating responses that are not ideal for imaging applications. Further inherent assumptions about identical behaviour of antenna elements have to be made. As well as this, calibration still relies on using cross-correlated data products. Morales (2011) proposed the Modular Optimal Frequency Fourier (MOFF) formalism as a flexible generalization of the direct imaging approach. A framework is described which exploits the computational advantages provided by direct Fourier transform imaging but with no limitations placed on the mixture of antenna elements or their placement, as well as producing fully calibrated images. In addition, provision is made for adaptive Fourier optics which can correct for non-coplanar array effects as well as antenna dependent terms. Visibilities from an FX correlator can be stored and calibrated offline due to explicit cross-correlations between all antenna pairs, which is not the case for gridded electric fields. Thus, direct imagers have the added requirement to calibrate in real time since individual antenna information is not retained after gridding. Beardsley et al. (2017) have successfully demonstrated an algorithm for this purpose.

The E-Field Parallel Imaging Correlator (EPIC), a generic implementation and simulation of this imaging approach in PYTHON, was described by Thyagarajan (Thyagarajan et al. 2017). As a streaming, direct imaging correlator, it can be thought of as a generic, flexible real-time camera of the radio sky for large interferometer arrays.

Here, we report a GPU-accelerated implementation of EPIC, built on Bifrost, a high-performance streaming framework. The implementation has been deployed on the LWA station located on the Sevilleta National Wildlife Refuge in New Mexico, USA. First light observations are shown, demonstrating its capability for transient detection.

The theory of the MOFF formalism underlying the EPIC imager is reviewed in Section 2, and a technical description of the implementation and development is discussed in Section 3. First light observations and an initial meteor transient detection are shown in Section 4, with benchmarks characterizing the performance on the LWA Sevilleta (LWA-SV) site discussed in Section 5. We summarize future work and conclude in Section 6.

2 THEORY

The interferometry formulation is based on the optical theory of partially coherent quasi-monochromatic light, by the van Cittert–Zernike theorem (Zernike 1938; Born & Wolf 1999). From this a relationship can be derived between the radiation pattern on the celestial sphere (in the far field) and a spatial coherence function measured on some plane between two points sampling the radiation pattern from the celestial sphere. This coherence function is the cornerstone of radio interferometry and is known as a 'visibility'.

A modern derivation can be found in Thompson, Moran & Swenson (2017), where the Fourier relationship between the sky coordinates and the interferometer coordinate system is described

$$I(l, m, w) = \iint V(u, v, w) \times \exp\left[2\pi i \left(ul + vm + w\left(\sqrt{1 - l^2 - m^2} - 1\right)\right)\right] \mathrm{d}u\, \mathrm{d}v. \quad (1)$$

Mathematically this can be described by an outer product between a vector representing a single frequency channel of Fourier-transformed voltages from all antennas, and its conjugate transpose. Thus given N antennas outputting N electric field patterns in a channel, we derive a resultant $N \times N$ visibility matrix. Because of Hermitian symmetry, only the upper or lower triangle is retained for efficiency in FX correlators. This relation is as follows:

$$V_{12}(u, v, w) = E_1(x_1, y_1, z_1) \otimes E_2(x_2, y_2, z_2)^*. \quad (2)$$

Here $E(x, y, z)$ represents the electric field measured by an antenna at some location in an orthonormal coordinate system, with $V(u, v, w)$ representing the resultant visibility matrix. The (u, v, w) coordinate system represents the vector separation (baseline vector) between the different antennas.

2.1 MOFF

The multiplication-convolution theorem from Fourier transform theory allows us to rearrange equation (1) to form the MOFF algorithm (Morales 2011) of

$$I(l, m, w) = \left\langle \left| \iint E(x, y, z) \exp\left[2\pi i \left(xl + ym + z\left(\sqrt{1 - l^2 - m^2} - 1\right)\right)\right] \mathrm{d}x\, \mathrm{d}y \right|^2 \right\rangle. \quad (3)$$

It is important to note that $E(x, y, z)$ constitutes the electric field in the Fourier domain convolved with the antenna illumination pattern. It is not a point function, but an electric field distributed across some physical extent in the measurement plane. Taking this into account equation (3) becomes

$$I(l, m, w) = \left\langle \left| \iint \left[W(x, y, z) * E'(x, y, z)\right] \times \exp\left[2\pi i \left(xl + ym + w\left(\sqrt{1 - l^2 - m^2} - 1\right)\right)\right] \mathrm{d}x\, \mathrm{d}y \right|^2 \right\rangle, \quad (4)$$

where $W(x, y, z)$ defines a 'gridding' function which constitutes a convolution in antenna space, E' represents the point measurement of the electric field within the measurement plane, and $*$ the convolution operator. In addition to the antenna illumination pattern,

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

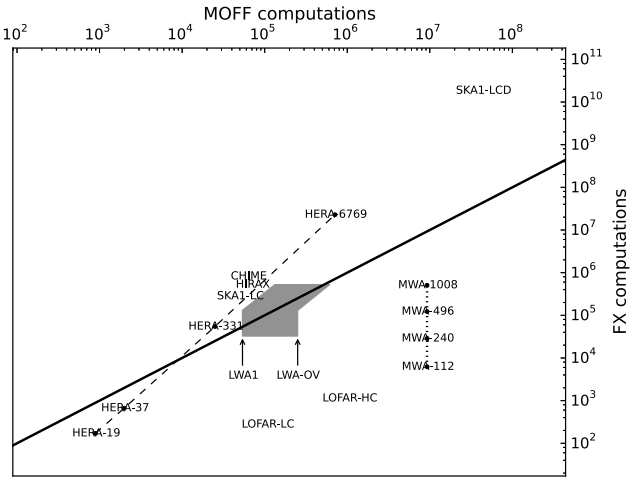

Figure 1. (Reproduced from Thyagarajan et al. 2017) A selection of current and planned instruments, with the solid black line delineating the boundary in efficiency between EPIC and FX correlators. Instruments below this line are more efficient with a standard FX correlator. Above the line it is more efficient to use EPIC.

$W(x, y, z)$ can optionally incorporate any wide-field effects resulting from non-coplanarities in the array, as well as ionospheric effects (Morales 2011). In our implementation, we assume a coplanar array. Correcting for non-coplanarities will be dealt with in a forthcoming paper. With this in mind, equation (4) becomes

$$I(l, m) = \left\langle \left| \iint \left[W(x, y) * E'(x, y) \right] \times \exp\left[2\pi i \left(xl + ym \right) \right] \mathrm{d}x\, \mathrm{d}y \right|^2 \right\rangle . \quad (5)$$

Thus equation (5) is a gridding of an electric field pattern directly for each antenna, followed by a spatial Fourier transform to produce the image. This transform is followed by squaring the image, or cross-multiplying between polarizations, and accumulating images over time. This produces what are commonly called 'dirty images', which is the true sky brightness distribution convolved with the PSF of the instrument (Taylor, Carilli & Perley 1999).

The EPIC architecture uses the MOFF algorithm as the basis for imaging. The computational cost of the EPIC architecture scales theoretically as $\mathcal{O}(n_g \log n_g)$, compared to $\mathcal{O}(n_a^2)$ for the classical FX correlator. For highly dense arrays, depending on array geometry, a MOFF-based correlator, such as EPIC, may be more efficient than an FX correlator (Thyagarajan et al. 2017). The limiting factor for the EPIC architecture is the Fourier transform size of the grid, whereas that for an FX correlator is the number of antenna pairs. A comparison of instruments and their most suited correlator type is shown in Fig. 1 (reproduced from Thyagarajan et al. 2017).

An additional characteristic of the EPIC architecture is the typically lower data rates in saving images at high time cadence (Thyagarajan et al. 2017). Images from the EPIC architecture are already calibrated and science ready compared to visibilities from an FX architecture which typically require additional processing offline to form science-ready images. This means that the sky can be imaged at a higher time resolution than is possible using an FX correlator. The ramifications for EPIC, as will be seen, is that all-sky imaging at sub-millisecond sampling periods is feasible, potentially yielding new insights into a wide range of time-domain phenomena at radio frequencies.

3 IMPLEMENTATION

3.1 Bifrost

The GPU-accelerated implementation of the EPIC architecture is done using the Bifrost framework (Cranmer et al. 2017). Bifrost is a highly abstracted library for building high performance streaming systems. The back end framework is built using C++ which calls high speed CUDA libraries and bespoke kernels implemented by the Bifrost authors. For ease of use an abstracted PYTHON front end is provided.

Bifrost is based around the concept of blocks, where each block performs some operation on the data, and then outputs it into a high-speed ring buffer in memory, which facilitates moving data between blocks. The output ring buffer from one block becomes the input ring buffer for the next block. Each block loads a 'gulp' of data from the ring, and processes it before placing a gulp into the output ring. The block processes data until the input ring are emptied or the pipeline is shutdown.

Many standard signal processing techniques are implemented into the Bifrost back end with GPU capability where appropriate. These include operations such as finite impulse response filters, fast Fourier transforms (FFTs), and various matrix algebra operations.

The PYTHON front end also includes a high-performance mapping function which takes a string of C++/CUDA and uses a Just-In-Time (JIT) compiler to generate and execute valid CUDA code on-the-fly using the Bifrost back end. This provides significant flexibility in doing small mathematical operations without the need to write multiple custom blocks and implement them directly into the Bifrost framework.

The majority of the EPIC implementation in Bifrost was done using the standard signal processing blocks as well as the Bifrost map function, with a notable exception of high-speed convolution and gridding. For this operation a custom kernel was created based on a high-speed gridder developed by Romein (2012).

3.2 Pipeline

The real-time streaming processor implementation comprises several Bifrost blocks[1] as a PYTHON program, with all significant compute and memory operations done seamlessly through Bifrost's high-performance C++/CUDA back end. An overview of this pipeline and the relationship between the various blocks is shown in Fig. 2.

Channelized raw data are received via UDP in a 4 + 4-bit complex integer format in the `UDP Receive` block. This complex integer representation serves to reduce the bandwidth required by the local ethernet connection. After the data have been captured, the channelized data are first decimated in frequency to obtain a bandwidth that can be processed without packet loss, and then transposed to move the data ordering from [*Time*, *Channel*, *Antenna*, *Polarization*], to [*Time*, *Channel*, *Polarization*, *Antenna*]. This is important as it facilitates contiguous loads in the gridding convolution step, discussed below.

[1] The source code for EPIC as well as the Bifrost-based pipeline implementation for the LWA is available at https://github.com/nithyanandan/EPIC.

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

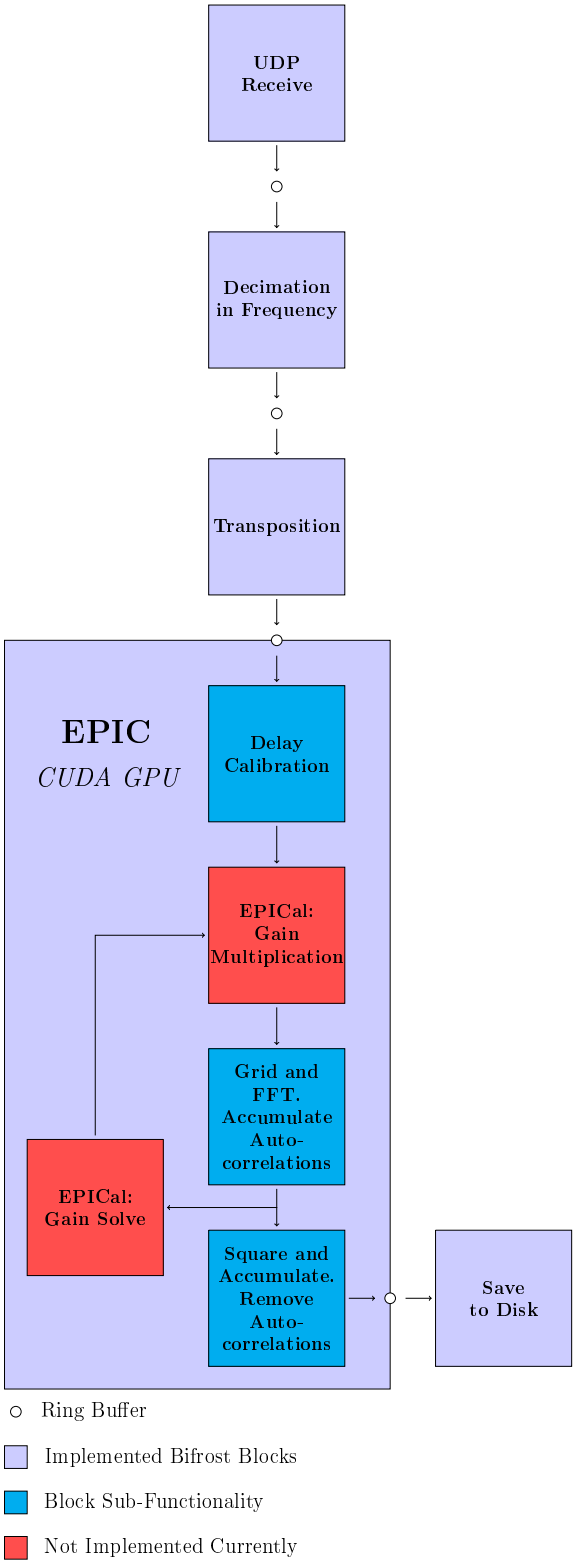

Figure 2. Block diagram of the Bifrost-based implementation of the EPIC architecture at LWA-SV. The blocks are named by their function and the arrows indicate the direction of data flow. The large EPIC block corresponds to a single operational block in the Bifrost pipeline, with its major sub-functionality displayed. Where the calibration steps sit are also included, despite not yet being implemented. The EPIC block maps closely with the architecture discussed in Thyagarajan et al. (2017).

After the `Transposition` block, the complex integer data are unpacked and promoted to a standard 64-bit complex floating-point number (32-bit real, 32-bit complex) and compensation for the signal path delays are applied using a JIT compiled Bifrost `map` function. The delay calibrated data are then convolved with the antenna illumination pattern, which is a user-defined convolution kernel. This is then gridded on to a 2D grid with a spacing of $< \frac{\lambda}{2}$, where λ is the wavelength of the channel, to ensure we are sampling all of the sky modes by sampling at the Nyquist wavelength.

This convolution and gridding operation is done using the Romein Convolution algorithm (Romein 2012) in the `Grid and FFT` block in Fig. 2, designed specifically for high-speed visibility gridding where locality is poor and memory bandwidth is high. The gridding convolution algorithm is described in more detail below.

Once the data have been gridded for a single time-step, the gridded data are inverse Fourier transformed to produce a complex-valued image on the sky. These images are then cross-multiplied in the `Square and Accumulate Image` block to form the polarized images, which are then accumulated to a user-defined time interval depending on the science use case. After accumulation to the threshold time, the image is written to disc in a binary format and converted to a FITS image in a post-processing step. This ensures the real-time processing is not held up by high-cost image manipulation operations.

Optionally autocorrelation removal can also be done to remove the zero-spacing power inherent in EPIC. Together with this, the imprint of the image of the gridding illumination kernels can be removed after the fact in a post-processing step as they are pre-generated and thus known previously.

3.3 Romein convolution algorithm

The Romein convolutional algorithm (Romein 2012) proved to be a critical step in the implementation of EPIC. Previous EPIC reference codes have attempted to use a direct convolution mapping using matrix multiplication, as described by the operator formalism in Thyagarajan et al. (2017).

Unfortunately, on a GPU this results in unacceptably high memory bandwidth which causes this step to bottleneck the code significantly. The Romein convolution was used instead as it is designed to reduce the GPU memory bandwidth significantly by only doing explicit memory store operations when necessary. The algorithm is designed to preferentially accumulate any grid updates into a high-speed local register on the GPU core.

The Romein convolution algorithm additionally allows multiple convolution kernels to be combined together and applied simultaneously. This not only allows convolution of the electric field with the illumination pattern, i.e. A-projection (Bhatnagar et al. 2008; Morales & Matejek 2009), but additionally provides scope for including wide-field and antenna effects, such as W-Projection (Cornwell, Golap & Bhatnagar 2008). The implementation of non-coplanarity correction to ensure wide-field fidelity will be a focus of future work.

The Romein convolution algorithm, written in C++/CUDA, was implemented by modifying the Bifrost back end and to add the necessary functionality. The additional Bifrost module is intended to be a generic, type-agnostic convolution kernel. This module is then called in the pipeline script from the Bifrost library using PYTHON's `ctypes` interface.

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

Figure 3. Aerial view of the LWA station at the Sevilleta National Wildlife Refuge. Most antenna elements are in a dense configuration towards the right of the image. A test antenna, is visible at the bottom. The signal processing hardware is contained within a modified, radio frequency shielded shipping container, visible in the left of the image.

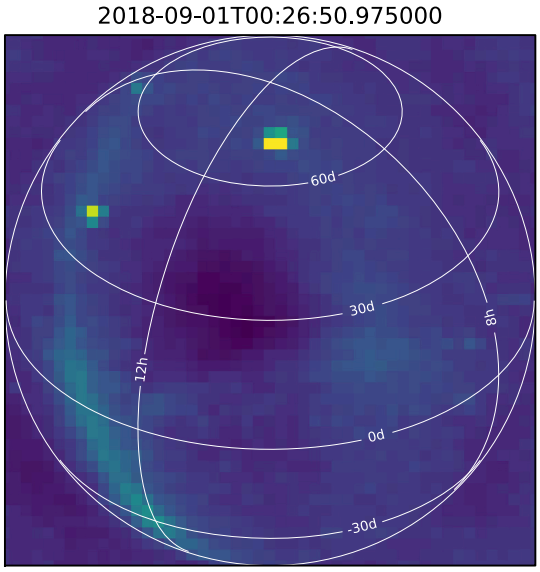

Figure 4. All-sky pseudo-Stokes-I image showing a meteor reflection detection during an observation on the LWA-SV site (upper centre). The plasma left by the meteor impacting the atmosphere reflects the signal from a 55.25 MHz TV transmitter located beyond the horizon. Lines of constant right ascension and declination in J2000 are marked in white. Cygnus A is the bright point in the upper left of the image. A .mp4 video file of this event is available, and has been submitted alongside this manuscript. The video is of images outputted sequentially at a 50 ms cadence.

4 DEPLOYMENT AND FIRST LIGHT

4.1 Long wavelength array

The LWA is a low-frequency radio interferometer observing between frequencies of 10–88 MHz, with two operational stations, one located at the Karl G. Jansky Very Large Array site and the other at the Seviletta National Wildlife Refuge (see Fig. 3), both in the state of New Mexico, USA (Henning et al. 2010; Taylor et al. 2012; Ellingson et al. 2013). Its high-density configuration makes it an excellent candidate for deployment of EPIC.

The LWA-SV array consists of 256 dipole antennas arranged in a dense pseudo-random arrangement inside a 110 m by 100 m elliptical aperture that is elongated north–south. An additional antenna is located approximately 300 m west of the core of the array, acting as an outrigger to help with calibration and to improve the angular resolution of the telescope. This outrigger was explicitly excluded during our implementation to ensure a high density, keeping the resultant image FFT size as small as possible.

The analogue signal from each dipole is initially low pass filtered and amplified at the front end before being transmitted over coaxial cable to the electronics shelter. Inside the shelter the analogue signal is further filtered and then digitized using ROACH2 boards. The boards use the CASPER ADC16x156-8 digitizer boards to sample the dipole signals at 204.8 MHz. The digitized signals are then Fourier transformed into 4096 25 kHz channels with a time resolution of 40 μs. At this point the frequency domain data, between 10 and 88 MHz, are requantized into 4 + 4-bit complex integer data, packetized, and routed over a 10/40 GbE network to a cluster of seven general purpose machines. Each machine is equipped with two Intel Xeon E5-2640 v3 processors, 64 GB of RAM, a Mellanox ConnectX-3 40 GbE network interface card, and two NVIDIA GTX 980 (Maxwell) GPUs.

4.2 Deployment

The initial deployment took place on the LWA-SV site during the week of the 2018 August 27–31. The EPIC architecture was deployed on a single cluster node, receiving a sixth of LWA-SV's total bandwidth. Operation of EPIC was achieved with no modifications to the LWA system or hardware apart from swapping the FX correlator software pipeline for EPIC. The LWA's public software library was used to perform delay calibration to account for different antenna cable lengths and to provide the array geometry (Dowell et al. 2012). A simple square top-hat function with 3-m extent was used as the illumination pattern for the dipole antennas. No additional calibration was performed. The observations reported here were run at an image accumulation time of 50 ms in order to allow observations of short-duration transient phenomena in the radio sky. Four channels of 25 kHz were processed with a combined bandwidth of 100 kHz centred at a frequency of 55.25 MHz. The stability of the system was tested with a 24 h operation under the EPIC correlator. Images were generated at the raw 40 μs time cadence of the LWA-SV and then accumulated to obtain the final cadence of 50 ms. A $\lambda/2$ grid spacing was used, resulting in approximately 64^2 image pixels.

4.3 Detection of meteor transient as proof of concept

EPIC images the whole sky as visible to the LWA-SV station. During our initial observations, multiple small transients were identified. The majority of them are radio frequency interference (RFI), which most often shows up on the horizon, indicating a terrestrial origin. Occasionally RFI can appear overhead, reflected off of airplanes or satellites. These signals are generally narrow bandwidth and highly polarized, making them easy to recognize.

After ruling out RFI events, some physical transients were noted, the brightest of which in our observing window was a meteor striking the Earth's atmosphere, a still frame pseudo-Stokes-I image of which is shown in Fig. 4, with a corresponding light curve shown in Fig. 5. A pseudo-Stokes image is one that is formed from straightforward linear combinations of the coherency vectors from the linear polarization parameters, but is acknowledged to not

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

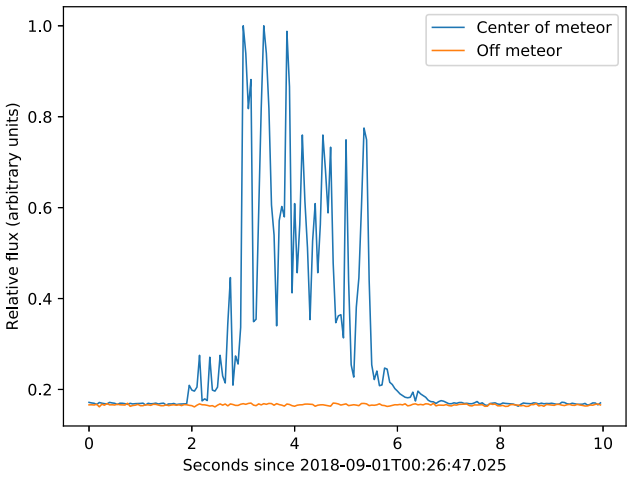

Figure 5. Light curve of the brightest pixel around the transient during the meteor passage noted in Fig. 4, with a comparison to the radio background. The time resolution is 50 ms. The reflection light curve shows considerable structure due to changes in the plasma tail as it expands and is distorted by atmospheric winds. The light curve is consistent with some of the examples in Helmboldt et al. (2014).

exactly represent the true Stokes vectors due to cross-coupling and polarization leakage effects. The meteor striking the atmosphere generates a plasma, which acts as a reflector for an over the horizon analogue TV transmitter at 55.25 MHz, illuminating the meteor plasma's path. This is almost identical to the methodology of studying meteor events through the use of radar (Prentice, Lovell & Banwell 1947). Studies of reflections such as these provide information about the speed of the neutral wind in the mesosphere and lower thermosphere through the observed Doppler shift of the reflection (Helmboldt et al. 2014). The total number of meteor reflections can also be used to inform estimates of the terrestrial accretion rate (see Kortenkamp & Dermott 1998, and references therein). Such events have been observed by the LWA previously, as well as self-emission from meteor trails (Obenberger et al. 2014) and lightning (Obenberger et al. 2018). This demonstrates the potential of an EPIC system for image-based all-sky transient detection and monitoring.

5 BENCHMARKS

During the first light deployment at the LWA-SV site, the performance was measured and characterized. The performance is a consequence of both the deployment system and hardware, as well as EPIC's execution method in comparison to an FX correlator.

Overall, in the first iteration, up to 800 kHz of bandwidth is processed per GPU card on the LWA-SV correlator, when running with only a single instrumental polarization, which is useful for maximizing bandwidth for faint transients and facilitates averaging over the band. With the LWA-SV system's current hardware layout, this corresponds to 9.6 MHz of single polarization bandwidth when EPIC is run on both GPUs of all six data capture servers. When running with both X and Y polarizations, which allows the formation of Stokes images, half the overall bandwidth is available: up to 400 kHz per card or 4.8 MHz for the entire system. We explore the factors contributing the per-GPU bandwidth below and discuss ideas for improvement.

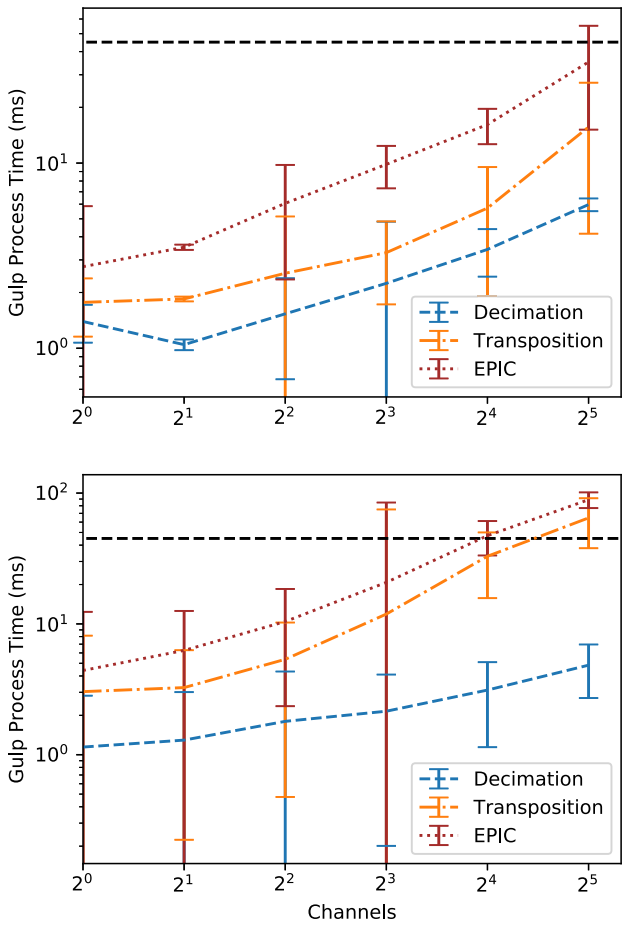

Figure 6. Processing time as a function of the number of channels being processed for (top) single polarization and (bottom) dual polarizations, with EPIC running on LWA-SV with a time gulp of 50 ms and grid size of 32 pixels on a side. At 32 (2^5) channels is when we began experiencing packet loss on dual polarizations, on the incoming data stream carrying electric field data, which marks when the system is no longer able to keep up with the input data rate. The black dashed horizontal line denotes the 90 per cent processing time for the gulp size.

5.1 Maximum throughput

To characterize the overall throughput of the system, we monitored the UDP streams being broadcasted by the ROACH2 boards running the front end Fourier transforms and channelization. If the system is keeping up with the input data, then there will be no packet loss. If compute requirements increase on the node, for example by increasing the number of channels per card or changing the frequency tuning such that a larger grid/FFT size is needed, then packet loss will occur as the pipeline struggles to keep up with the incoming data stream.

There are additional overheads in the system, such as running a normal Linux operating system in the background that can cause the occasional reductions in processing performance. To ensure that the pipeline does not drop packets due to such variations, we found empirically that a time 'gulp', i.e. the amount of time represented by a single chunk of data, such that the data can be processed in ≈90 per cent of the observed time is useful, providing a 10 per cent margin for system processing variations. For example, if ingesting 50 ms worth of data in a single gulp from the ring buffer, to ensure

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

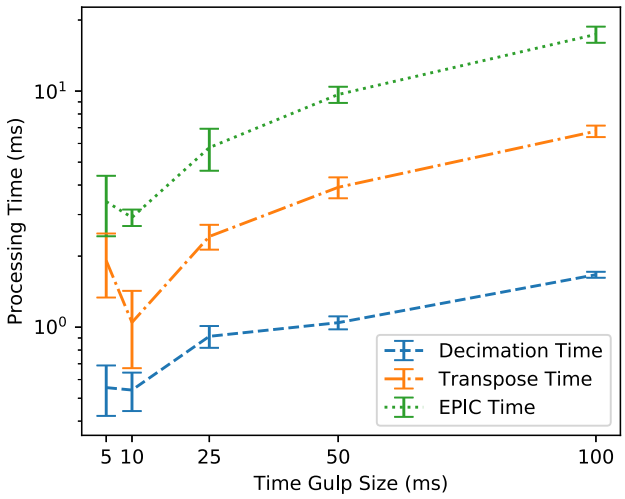

Figure 7. An exploration of how the system scales as a function of the time gulp sizes for 100 kHz of bandwidth and dual polarization. Each vertical bar is sub-divided to show the time used by each block in the pipeline. The legend corresponds to the blocks in Fig. 2, with 'Image and Accumulation' corresponding to the pipeline element on the CUDA GPU. These data were derived from at least 600 trials of each time gulp size.

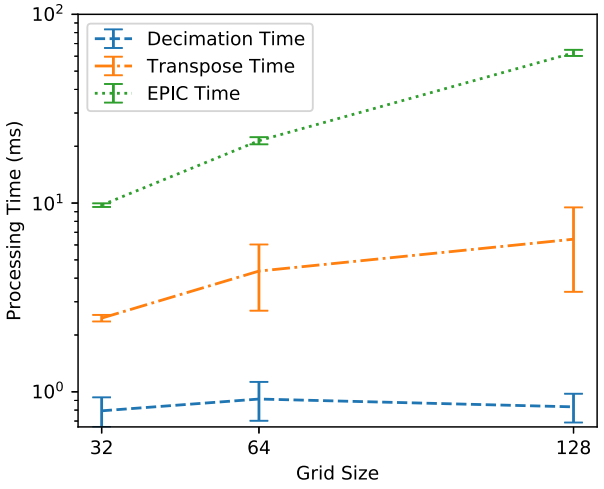

Figure 8. An exploration of how the system scales as a function of the grid size for a time gulp of 25 ms, 100 kHz of bandwidth, and dual polarization. The grid size is the size of one dimension of our squared grid. These data were derived from at least 600 trials of each grid size. 'EPIC Time' in this instance is the GPU element specified in Fig. 2. A grid size of 128 is more than can be processed in real time by the system, as it causes packet loss on the input UDP stream, but it is plotted here to show scaling.

the system can keep up, the GPU should process it in 45 ms to keep the system running smoothly.

The results of our initial tests on the system are shown in Fig. 6 where the gulp processing time and UDP packet loss fraction are shown as a function of the number of frequency channels processed. As computational resources are exceeded by increasing the number of channels, the pipeline backs up, and packet loss increases to indicate that system capacity has been exceeded.

With a grid size of 64^2 and the time gulp size set to 50 ms, we are capable of running up to 16 channels (400 kHz) with dual polarizations before packet loss increases to indicate the pipeline stalling. Single polarization mode runs over twice as fast as the dual polarization mode. In Fig. 7, the scaling of the system as a function of time gulp size is shown when processing 100 kHz of bandwidth and dual polarization. The scaling with gulp size is roughly linear, with the GPU coping well at a variety of representative time gulp sizes between 5 ms and 0.1 s. Similarly, Fig. 8 shows the scaling of the pipeline with the grid size for a gulp size of 25 ms with the same bandwidth and polarization set-up as in Fig. 7. We see that the processing times for grid sizes of 32 and 64 pixels on a side are roughly comparable, indicating that the EPIC processing time may not be dominated by the Fourier transform at these grid sizes. This can also been seen in Table 1 where the processing time per gulp is explored for a representative pipeline run. The scaling of the processing time between 64 and 128 pixels on side is around 2.5 times whereas theory would predict an increase of 4.7. The reason for this is potentially the underlying cuFFT library being more efficient for larger FFT sizes compared to smaller ones (Kent & Nikolic 2016). We note that larger sizes for both the time gulp and grid size are unable to be tested because of the lack of sufficient memory on the GTX 980 GPUs available at LWA-SV.

5.2 GPU performance

Here we assess the overall performance and suitability of EPIC for a GPU programming model. This can be explored using a roofline model, a common visualization in high-performance computing to analyse the execution properties of a particular algorithm (Demmel, Williams & Yelick 2009).

Table 1. Representative approximate breakdown of processing time by block as a fraction of the time gulp. This was done for a grid size of 64^2, 2048 40 μs time samples and eight channels.

Block	Processing Time (per cent of gulp time)
Decimation	2
Transposition	4
EPIC	90
EPIC – FFT	35
EPIC – Gridding	20
EPIC – Cross-Multiply	40
EPIC – Data Transformation	5
Save	4

The roofline comparison between the GPU computed elements of an FX correlator and EPIC is shown in Fig. 9. The example here is computed using a representative roofline for an NVIDIA GTX 980 GPU, used in this implementation, and for the elements of the pipeline that execute on the GPU. A GTX 2080 roofline is also provided, as a potential upgrade for the LWA correlator. The FX correlator is clearly in the compute bound regime, whereas EPIC is memory bound. This means increasing the memory bandwidth available rather than compute power will be more beneficial for EPIC, in contrast to an FX correlator which is predominantly compute bound.

Upgrading the LWA-SV correlator to use GTX 2080's, which have over double the memory bandwidth and compute performance, should yield a significant performance increase in bandwidth that can be processed. Assuming an increase in performance of at least two times with the new cards, notwithstanding additional optimization, over 15 MHz of LWA bandwidth should be able to be processed.

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

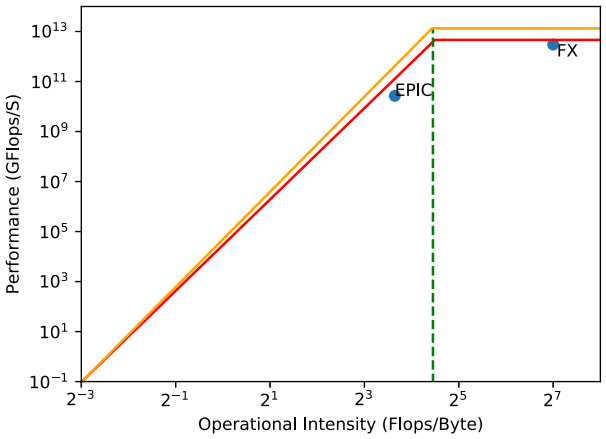

Figure 9. A roofline model comparing an FX correlator outer product with the EPIC pipeline on a NVIDIA GTX 980 GPU. The red line represents the maximum number of operations per second for a GTX 980 GPU, and the orange line for a GTX 2080, a potential upgrade card for the LWA. The rooflines for both cards were worked out using the memory bandwidth and peak compute performance for single-precision floating point (FP32) operations. The memory bandwidth of the GTX 980 and GTX 2080 are 226 and 616 GB s^{-1}, with peak FP32 performance of 4.6 and 13.3 TeraFlops, respectively.

We note two important caveats that such a direct comparison is not entirely appropriate. First, the EPIC architecture provides end-to-end real-time imaging (from raw antenna voltages to science-ready calibrated images), whereas an FX correlator predominantly consists of a single mathematical operation, namely, outer product of the raw voltages and thus does not calibration or imaging, which incur additional costs. Secondly, if fast time-domain studies (time-scales $\lesssim 1$ ms) are to be performed with an FX-based correlator, the cost of gridding and imaging will be much higher since they have to be performed at such fast cadences and have not been included in these estimates.

6 CONCLUSION

The first version of a working EPIC direct imager, through direct implementation of the MOFF algorithm, has been developed fully, and its implementation and operation demonstrated at the LWA-SV site. Observations of transients from reflections of terrestrial transmissions off passing meteors on time-scales of $\sim$2 s at a cadence of 50 ms are reported. These serve to verify EPIC as a science-capable interferometric imaging capability.

The Bifrost framework aided in implementing EPIC. The C++/CUDA back end abstracts away complicated constructs, such as the ring buffers, which form the communication backbone between processing steps. The major advantage is the native CUDA support, facilitating access to the power of the GPGPU paradigm. Extending the Bifrost framework, such as adding extra GPU-enabled processing blocks, was straightforward.

This successful deployment and working demonstration of the principles behind EPIC and the MOFF formalism mark a paradigm shift in correlator technology. The impact is especially acute for high-density arrays such as SKA1-Low and the completed HERA configuration. It can also offer the capability, with its next iteration of development as a self-triggering transient survey instrument for arrays such as the Low Band Observatory (ngLOBO; Taylor et al. 2017), and the LWA Swarm Telescope (Dowell & Taylor 2018). Higher frequency instruments such as the MWA can benefit from the unique capabilities of EPIC for exploring FRB phenomena, with its ability to image the entire celestial hemisphere simultaneously at high time cadence. Additional potential scientific uses range from ionospheric disturbance mapping to direct observations of compact astrophysical sources.

While the first version of the EPIC system is now operational, significant improvements are planned. Future work will include:

(i) EPICal – EPIC requires calibration in real time. A solution has been demonstrated (Beardsley et al. 2017) but not implemented yet in this deployment. This will be an important feature addition since direct imaging approaches do not allow for post-acquisition methods to improve the image calibration.

(ii) Wide-field effects – Effects of non-coplanarity and wide-field effects may be significant at low frequencies. Thus, it may be necessary to deal with non-coplanarity and wide-field effects in EPIC. EPIC's antenna-based gridding convolution naturally allows for the non-coplanar effects to be fully incorporated and corrected for (Cornwell et al. 2008; Morales 2011). A forthcoming paper will elucidate the principles and practicalities behind doing this on EPIC-based imagers.

(iii) Optimization – EPIC has unique computational challenges associated with it, which will benefit from broad optimization of key kernels to remove bottlenecks during the convolutional gridding and the FFT stages.

(iv) Real-time transient detection – Our transient detection in this manuscript was done using offline analysis of the data. A prototype transient detector has been implemented, however it is in the early stages. An online automated transient detector, with effective RFI filtering, will provide another strong science capability to the EPIC architecture.

The addition of aforementioned features will further increase EPIC's scientific repertoire, through correction of antenna based terms in the imaging process, as well as precision calibration in real time, yielding precision astronomical observations across the whole sky, with high resolution. These are planned to be implemented in the next iteration of development of EPIC on the LWA. We plan to continue observing in EPIC 'mode' at LWA-SV for longer periods of time at a high time resolution, to facilitate blind, source-agnostic surveys where transient phenomena might appear.

ACKNOWLEDGEMENTS

This work is supported by National Science Foundation awards AST-1710719 and AST-1711164. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X GPU used for prototyping and testing the EPIC pipeline. Construction of the LWA has been supported by the Office of Naval Research under Contract N00014-07-C-0147 and by the AFOSR. Support for operations and continuing development of the LWA1 is provided by the Air Force Research Laboratory and the National Science Foundation under grants AST-1835400 and AGS-1708855. APB is supported by an National Science Foundation Astronomy and Astrophysics Postdoctoral Fellowship under award AST-1701440. JK is funded by the Engineering and Physical Sciences Research Council, UK. We also acknowledge Emma Maton for her help in proofreading the final manuscript.

REFERENCES

Amiri M. et al., 2019, Nature, 566, 230
Beardsley A. P., Thyagarajan N., Bowman J. D., Morales M. F., 2017, MNRAS, 470, 4720

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019

Bhatnagar S., Cornwell T. J., Golap K., Uson J. M., 2008, A&A, 487, 419
Born M., Wolf E., 1999, Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light, 7th edn. Cambridge Univ. Press, Cambridge
Bunton J. D., 2011, IEEE Trans. Antennas Propag., 59, 2041
Caleb M. et al., 2017, MNRAS, 468, 3746
Cornwell T. J., Golap K., Bhatnagar S., 2008, IEEE J. Sel. Topics Signal Process., 2, 647
Cranmer M. D. et al., 2017, J. Astron. Instrum., 6, 1750007
Daishido T. et al., 1991, in Cornwell T. J., Perley R. A., eds, ASP Conf. Ser. Vol. 19, IAU Colloq. 131: Radio Interferometry: Theory, Techniques, and Applications, Astron. Soc. Pac., San Francisco, p. 86
DeBoer D. R. et al., 2017, PASP, 129, 045001
Williams S., Waterman A., Patterson D., 2009, Communications of the ACM, 52, 65
Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, Proc. IEEE, 97, 1482
Dowell J., Taylor G. B., 2018, J. Astron. Instrum., 07, 1850006
Dowell J., Wood D., Stovall K., Ray P. S., Clarke T., Taylor G., 2012, J. Astron. Instrum. , 1, 1250006
Ellingson S. W. et al., 2013, IEEE Trans. Antennas Propag., 61, 2540
Foster G., Hickish J., Magro A., Price D., Zarb Adami K., 2014, MNRAS, 439, 3180
Helmboldt J. F., Ellingson S. W., Hartman J. M., Lazio T. J. W., Taylor G. B., Wilson T. L., Wolfe C. N., 2014, Radio Sci., 49, 157
Henning P. A. et al., 2010, Proc. Sci., The First Station of the Long Wavelength Array, SISSA, Trieste. PoS(ISKAF2010)024 preprint (arXiv:1009.0666)
Kent J., Nikolic B., 2016, Quantifying Power Efficiency of FFTs on Nvidia GPUs, http://ska-sdp.org/sites/default/ files/attachments/powerefficiencysdp_part_1_-_signed. pdf
Kortenkamp S. J., Dermott S. F., 1998, Icarus, 135, 469
Morales M. F., 2011, PASP, 123, 1265
Morales M. F., Matejek M., 2009, MNRAS, 400, 1814
Obenberger K. S. et al., 2014, ApJ, 788, L26
Obenberger K., Bennett J., Malins J., Parris R., Pedersen T., Taylor G., 2018, Geophys. Res. Lett., submitted
Prentice J. P. M., Lovell A. C. B., Banwell C. J., 1947, MNRAS, 107, 155
Romein J. W., 2012, in Bannerjee U. , Gallivan K., eds, Proceedings of the 26th ACM international conference on Supercomputing. ACM, ACS, New York, p. 321
Romney J. D., 1985, in Perley R. A., Schwab F. R., Bridle S. H., eds, Synthesis Imaging: Course notes from an NRAO Summer School, National Radio Astronomy Observatorym , New Mexico, p.
Taylor G. B., Carilli C. L., Perley R. A. 1999, ASP Conf. Ser. Vol. 180, Synthesis Imaging in Radio Astronomy II. Astron. Soc. Pac., San Francisco, p. 7
Taylor G. B. et al., 2012, J. Astron. Instrum., 1, 1250004
Taylor G. et al., 2017, preprint (arXiv:1708.00090)
The CHIME/FRB Collaboration, 2018, ApJ, 863, 48
Thompson A. R., Moran J. M., Swenson G. W., 2017, Interferometry and Synthesis in Radio Astronomy. Springer International Publishing, Cham
Thyagarajan N., Beardsley A. P., Bowman J. D., Morales M. F., 2017, MNRAS, 467, 715
Zernike F., 1938, Physica, 5, 785

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

D4qOZYW.mp4

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article-abstract/486/4/5052/5484888 by Arizona State University West user on 11 June 2019