Draft version October 15, 2020
Typeset using LaTeX **twocolumn** style in AASTeX63

arXiv:2010.06748v1 [astro-ph.HE] 14 Oct 2020

An analysis pipeline for CHIME/FRB full-array baseband data

D. Michilli [1,2] K. W. Masui [3,4] R. Mckinven [5,6] D. Cubranic [7] M. Bruneault [1,2] C. Brar [1,2] C. Patel,[6,1,2] P. J. Boyle [1] I. H. Stairs [7] A. Renard [6] K. Bandura [8,9] S. Berger [1,2] D. Breitman [6,10] T. Cassanelli [5,6] M. Dobbs [1,2] V. M. Kaspi [1,2] C. Leung [3,4] J. Mena-Parra [3] Z. Pleunis [1,2] L. Russell [3,4] P. Scholz [6] S. R. Siegel [1,2] S. P. Tendulkar [1,2] and K. Vanderlinde [6,5]

[1] *Department of Physics, McGill University, 3600 rue University, Montréal, QC H3A 2T8, Canada*
[2] *McGill Space Institute, McGill University, 3550 rue University, Montréal, QC H3A 2A7, Canada*
[3] *MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA*
[4] *Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA*
[5] *David A. Dunlap Department of Astronomy & Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S 3H4, Canada*
[6] *Dunlap Institute for Astronomy & Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S 3H4, Canada*
[7] *Dept. of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, B.C. V6T 1Z1 Canada*
[8] *CSEE, West Virginia University, Morgantown, WV 26505, USA*
[9] *Center for Gravitational Waves and Cosmology, West Virginia University, Morgantown, WV 26505, USA*
[10] *Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada*

ABSTRACT

The Canadian Hydrogen Intensity Mapping Experiment (CHIME) has become a leading facility for detecting fast radio bursts (FRBs) through the CHIME/FRB backend. CHIME/FRB searches for fast transients in polarization-summed intensity data streams that have 24-kHz spectral and 1-ms temporal resolution. The intensity beams are pointed to pre-determined locations in the sky. A triggered baseband system records the coherent electric field measured by each antenna in the CHIME array at the time of FRB detections. Here we describe the analysis techniques and automated pipeline developed to process these full-array baseband data recordings. Whereas the real-time FRB detection pipeline has a localization limit of several arcminutes, offline analysis of baseband data yields source localizations with sub-arcminute precision, as characterized by using a sample of pulsars and one repeating FRB with known positions. The baseband pipeline also enables resolving temporal substructure on a micro-second scale and the study of polarization including detections of Faraday rotation.

1. INTRODUCTION

Fast radio bursts (FRBs; Lorimer et al. 2007) are millisecond-duration radio transients whose integrated column density of free electrons (approximately quantified by the dispersion measure, DM) significantly exceeds the maximum value expected by Galactic models (Cordes & Lazio 2002; Yao et al. 2017). The extragalactic nature of the bursts has been confirmed with the localization of a handful of FRBs to host galaxies[1]. The cosmological distance of these host galaxies gives severe constraints on source emission models, requiring radio luminosities far higher than those of any known Galactic source. A comprehensive explanation of the phenomenon is still missing and a variety of models exist[2].

The Canadian Hydrogen Intensity Mapping Experiment (CHIME; Newburgh et al. 2014) radio telescope has been equipped with the CHIME/FRB backend (CHIME/FRB Collaboration 2018) to discover and study FRBs. Thanks to the large instantaneous field of view ($\sim$ 200 square degrees), high sensitivity and almost continual on-sky time, CHIME/FRB has already detected hundreds of FRBs (e.g. Fonseca et al. 2020). CHIME/FRB blindly searches for sources with a real-time pipeline capable of processing the data from the

Corresponding author: Daniele Michilli
danielemichilli@gmail.com

[1] http://frbhosts.org (Heintz et al., in prep.)

[2] https://frbtheorycat.org (Platts et al. 2019)

telescope and automatically selecting potential candidates (CHIME/FRB Collaboration 2018). The data processed by the real-time pipeline has a resolution of ~ 1 ms and no polarization information. However, extracting as much information as possible for every burst is essential since the vast majority of FRB sources are detected just once[3] and only a few have been reported to repeat (Spitler et al. 2016; CHIME/FRB Collaboration et al. 2019a,b; Kumar et al. 2019; Fonseca et al. 2020).

It is interesting to study FRBs at very high resolution for understanding, among other things, the complex and puzzling time-frequency structures of their signal (Hessels et al. 2019), and scintillation of the signal by intervening plasma (Masui et al. 2015). Also, polarization information is key to understanding the emission mechanism and local environment of FRBs (Petroff et al. 2015; Masui et al. 2015). Finally, source localization with the static FFT beams of the real-time search (CHIME/FRB Collaboration 2018; Ng et al. 2017) has uncertainties of roughly the beam width of tens of arcminutes (e.g. CHIME/FRB Collaboration et al. 2019b). However, many of the scientific goals of CHIME/FRB rely on an accurate localization of the detected FRBs. Some examples are the cross-correlation of FRBs with structure in the Universe (Masui & Sigurdson 2015; Rafiei-Ravandi et al. 2019), the study of the luminosity function of FRBs, where a localization is essential to compensate for beam effects when measuring the flux of the detections, and the localization of bright, low-DM FRBs to host galaxies in the local Universe.

To allow detailed studies of FRBs, a triggered baseband recording system capable of storing the electric field measured by each of the 1024 dual-polarization feeds has been developed for CHIME/FRB (CHIME/FRB Collaboration 2018). Here, we report the operations of the baseband system of CHIME/FRB and describe the analysis techniques and software developed to process these data. A summary of the baseband system is provided in §2. Algorithms to form tied-array beams and correct for signal dispersion are described in §3. The process to use these beams to refine source positions is detailed in §4. An automated pipeline that processes the baseband data by using the aforementioned software is outlined in §5. The localization capability of the baseband pipeline is characterized in §6. Finally, conclusions are drawn in §7.

2. CHIME/FRB TRIGGERED BASEBAND RECORDING SYSTEM

[3] http://frbcat.org (Petroff et al. 2016)

The detection pipeline of CHIME/FRB searches for short-duration, dispersed peaks in a stream of total intensity data in each of 1024 formed beams (Ng et al. 2017; CHIME/FRB Collaboration 2018). CHIME/FRB Collaboration (2018) described the plan to create a triggered baseband recording system for the CHIME/FRB experiment. The system is now in place and routinely records snapshots of baseband data around signals of interest. The raw voltages measured by the 1024 dual-polarization feeds of the telescope, which operates between 400 and 800 MHz, are amplified, digitized and processed by a cluster of motherboards called the F-engine (Bandura et al. 2016). For each receiver, the F-engine uses a 4-tap polyphase filter bank to produce a spectrum with 1024 channels (each 390 kHz wide) every 2.56 µs. The baseband data are calculated by rounding this output to 4+4 bit complex numbers. The resulting data rate is 6.5 Tb/s and a memory buffer allows the storage of 35.5 seconds of baseband data at a given time. From the moment a signal arrives at the telescope, the real-time pipeline is able to process the event and trigger a baseband dump in ~ 14 seconds. This leaves a usable data buffer of ~ 20 seconds, which corresponds to a maximum DM of $\sim 1000\,\mathrm{pc\,cm^{-3}}$ at CHIME frequencies. Typically, 100 ms of baseband data are stored around the time of arrival (TOA) of triggering events for each frequency channel. Additional time is usually added around the burst due to the uncertainties in the TOA and DM values reported by the real-time pipeline, which are between 16–128 ms and 1.6–25 $\mathrm{pc\,cm^{-3}}$, respectively. This produces an average of ~ 100 GB of baseband data stored for a detected event.

Currently, all the FRBs detected by the real-time pipeline trigger a baseband dump. There is also the possibility to insert special rules to trigger, for example, known pulsars that are otherwise discarded. If needed, baseband dumps can also be triggered manually for study of specific sources.

3. BEAMFORMING AND DE-DISPERSION

The phase information contained in the baseband data can be used to phase-reference detected waves to any direction within the field of view of the telescope prior to coadding over feeds. This process, referred to as *beamforming*, maximizes the telescope sensitivity to a certain direction and effectively *points* the telescope to that direction. Within the CHIME experiment, a bright source is observed every day to calibrate the gains of each receiver of the telescope (CHIME/FRB Collaboration 2018). Our baseband pipeline uses by default the closest calibration observation in time as a phase reference. A number of beams is then formed in a set

of directions by adding the expected geometrical delays with respect to the geometrical center of the telescope. This operation is performed by using custom developed software. The Python ephemeris library `Skyfield`[4] is used to calculate the geometrical factors for beamforming and calibration. The beamforming routine is the most computationally expensive stage of the baseband pipeline, with one beam formed in roughly one hour with the current configuration. The computational time and memory needed to form the beams limit us to ~ 200 beams formed simultaneously with our current processing machine.

The dispersion of radio signals induced by intervening free electrons along the line of sight impairs the detection of fast transients and it is usually corrected with de-dispersion algorithms. Traditionally, this operation is performed *incoherently* by adding appropriate delays to different frequency channels relative to a reference frequency (e.g. Lorimer & Kramer 2004). However, residual smearing of the signal within single channels will diminish the signal-to-noise ratio (S/N) of the event and smooth out temporal features shorter than the smearing time.[5] With the use of baseband data, it is possible to mathematically remove the effect of dispersion by modelling the intervening medium as a cold tenuous plasma (Hankins 1971). This leads to the following transfer function, whose inverse can be used to rotate the phase of the incoming radio waves and effectively correct for intra-channel smearing,

$$H(\nu+\nu_0) = \exp\left(\frac{2\pi \mathrm{i}\nu^2 k_{\mathrm{DM}}\mathrm{DM}}{\nu_0^2(\nu+\nu_0)}\right), \tag{1}$$

where ν_0 is a reference frequency, ν is the frequency offset from ν_0 and $k_{\mathrm{DM}}^{-1} = 2.41\times10^{-4}\,\mathrm{MHz}^{-2}\,\mathrm{pc}\,\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$ (Manchester & Taylor 1972) is the dispersion constant.

We use baseband data to coherently correct for the dispersion of radio signals, which allows us to resolve the bursts at a high temporal resolution. An example of a comparison between the waterfall plots – the signal intensity displayed as a function of time and frequency – of the same burst obtained with intensity and baseband data, respectively, is displayed in Fig. 1. The burst shown is a CHIME/FRB detection of FRB 20190618A (CHIME/FRB Collaboration, in prep.). The spectro-temporal structures in the burst, unresolved by the intensity data, are clearly visible in the baseband data. This is due to a combination of improvements made possible by the baseband data, namely the bandpass correction by phasing the telescope to the source location,[6] the higher temporal resolution and the ability to coherently correct for the signal dispersion.

4. LOCALIZATION

CHIME/FRB formed beams have a FWHM $\sim$ 15–30 arcmin across the band at zenith. However, due to the interferometric nature of the telescope, the localization of point sources can be improved by mapping the signal intensity around the initial detection. This intensity map can then be fitted with a model of the expected telescope response to obtain a more accurate position. By using this technique, it is theoretically possible to obtain a precision on the localization given by (Masui et al. 2019)

$$\sigma_\theta = \frac{\sqrt{6}}{2\pi}\frac{\lambda}{D}\frac{1}{\mathrm{S/N}}\frac{1}{\cos(\theta)}, \tag{2}$$

where λ is the observing wavelength, D is the maximum separation between the telescope antennae, S/N is the signal-to-noise ratio of the detected event and θ is the zenith angle of the source. For a source detected at zenith by CHIME/FRB, where $D\sim80\,\mathrm{m}$ and considering $\lambda = 600\,\mathrm{MHz}$, this translates to

$$\sigma_\theta \sim \frac{8}{\mathrm{S/N}}\,\mathrm{arcmin}. \tag{3}$$

Therefore, a bright FRB with a S/N = 100 might be localized by CHIME/FRB with a precision of ~ 5 arcsec. However, unmodelled systematic effects will limit the localization accuracy, as discussed in §6.

Our strategy to map the signal intensity is to phase-reference the detected waves to a grid of trial positions around the initial localization and to measure a single value of the S/N value in each of them. This corresponds to forming a grid of overlapping beams in the local reference system of the telescope. The total S/N of the burst measured in each beam is calculated and the resulting array is then fitted with a 2D Gaussian. A Gaussian model with six free parameters (peak position, two widths, the height of the curve, and a minimum value) has been found to fit well the sensitivity response of CHIME formed beams within their FWHM, with typical residuals $\lesssim 5\%$. When performing the fit, we use as the uncertainty on the values of the S/N the noise level measured in single beams. An example of the

[4] https://rhodesmill.org/skyfield

[5] For a source with $\mathrm{DM} = 500\,\mathrm{pc}\,\mathrm{cm}^{-3}$ detected by the real-time pipeline of CHIME/FRB, which has 16,384 channels, the smearing time at the bottom of the band is ~ 0.4 ms, while at the native resolution of baseband data, which have 1024 channels, the smearing time at the bottom of the band is ~ 25 ms.

[6] The telescope primary beam will affect baseband and static FFT beams alike. However, its size is much larger than the size of formed beams.

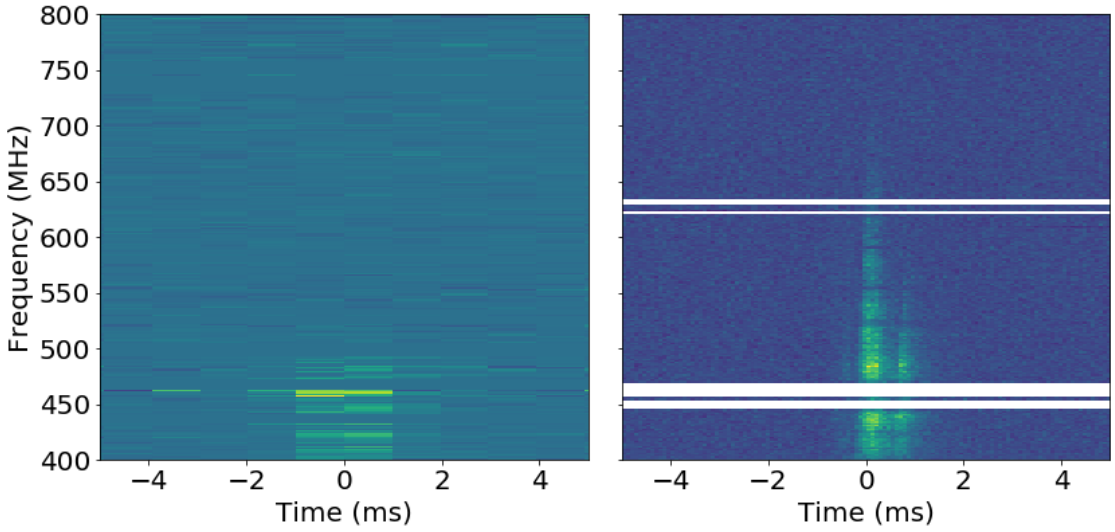

Figure 1. Example of a burst from FRB 20190618A (CHIME/FRB Collaboration, in prep.) detected by CHIME/FRB and processed through the intensity (left) and baseband (right) pipelines. The color scale, independent for the two plots, is proportional to the intensity of the signal. The left plot has a time resolution of ~ 1 ms and has been incoherently dedispersed to a DM of $228.93\,\mathrm{pc\,cm^{-3}}$. The right plot has a time resolution of $81.92\,\mu\mathrm{s}$ (i.e. 32 times the native time resolution) and has been coherently dedispersed to the same DM value. Both panels are plotted with 256 frequency channels. Horizontal white lines in the right plot are frequency channels missing due to the failure of one of the machine's nodes during the baseband dump.

localization of a relatively weak single pulse from FRB 20180916B (CHIME/FRB Collaboration et al. 2019b) is shown in Fig. 2. The coordinate system used in the plot, where the grid of beams is regularly spaced, is a local spherical-polar coordinate system centered at the telescope zenith where X runs from East to West (EW), perpendicular to the telescope meridian, and Y runs from South to North (SN), parallel to the telescope meridian, referenced to 400 MHz. The different frequency channels of single beams are aligned to have the same RA and Dec at the beam center.

5. BASEBAND PROCESSING PIPELINE

A framework with minimal overhead was developed to pipeline the analysis described in the previous sections to produce scientifically useful output from the baseband data recorded by CHIME/FRB without any human intervention. The different steps involved in this pipeline are presented in Fig. 3 and described in detail below. The framework and pipeline are hardware agnostic and designed to be highly scalable and are currently operational on a compute cluster located at the telescope. The baseband pipeline can run automatically on the new events that are identified by the real-time detection pipeline of CHIME/FRB or be manually triggered to process specific events.

5.1. *Pipeline infrastructure*

The automated pipeline consists of three major technologies, Apache Airflow,[7] an open source platform to schedule and monitor workflows; Docker,[8] a platform to deliver software in standard, environment agnostic, packages, called containers; and a custom built coordinator software which manages metadata queuing and orchestration of docker containers through Docker Swarm[9]. When an event is detected by the real-time pipeline and the relevant baseband data are stored on disk, a unique identifier for the event is stored in a *bucket* of the memory of the coordinator. This unique identifier, called the *event ID*, is the key to retrieving and depositing all event information in a database of CHIME/FRB detections. The Airflow workflow, comprising of application programming interface (API) calls to the coordinator in a directed acyclic graph (DAG), scans the baseband bucket periodically (currently, every 15 minutes) and instantiates the analysis pipeline if an event ID is found. Additionally, the coordinator allows the user to manually insert a list of event IDs to be processed into this bucket.

At this point, following the recipe described in the DAG, Airflow launches a series of analysis tasks to be performed on the cluster via the coordinator API. In

[7] https://airflow.apache.org
[8] https://www.docker.com
[9] https://docs.docker.com/get-started/orchestration

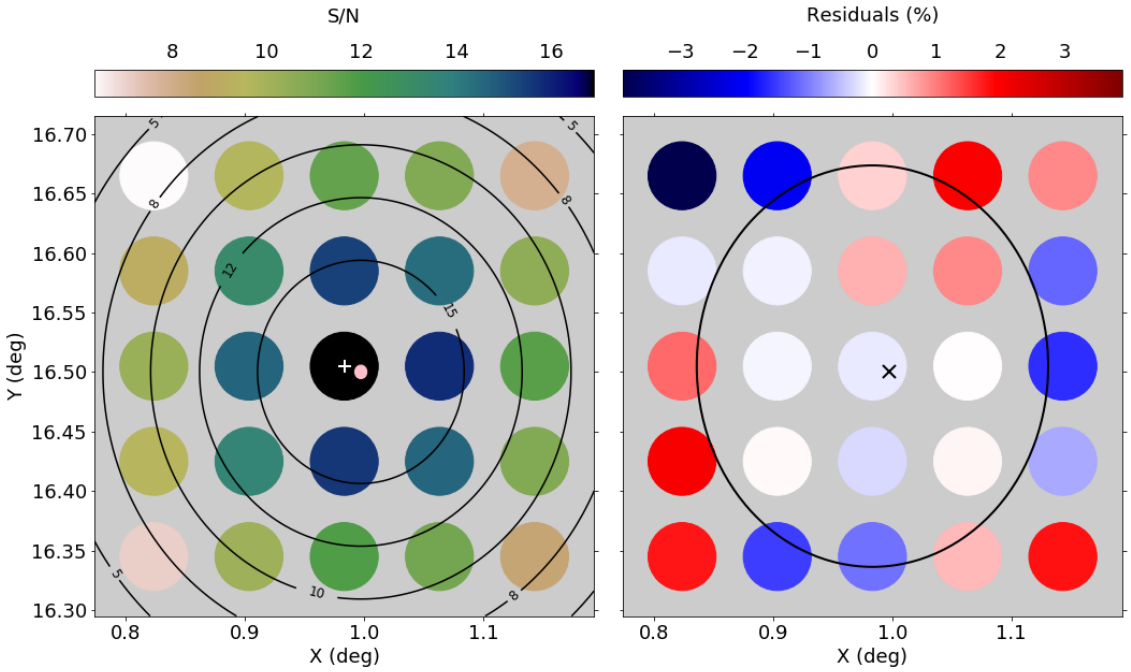

Figure 2. Localization of a relatively weak burst from FRB 20180916B (CHIME/FRB Collaboration et al. 2019b) before applying any post-localization correction, as discussed in §6. The localization is presented in the local coordinates of the telescope, with X going from East to West and Y from South to North, referenced to 400 MHz. Left: Intensity map of the signal measured with a grid of tied-array beams fitted with a Gaussian model of the telescope beam. Colored circles are placed at the position of the beam centers at 400 MHz and their color-scale is proportional to the S/N detected in each beam, as indicated by the color bar at the top. Black contours represent the Gaussian model at different levels, reported as labels on the plot. The central pink ellipse is the 1σ localization region from the fit. The true source position measured by Marcote et al. (2020) is located at the center of the plot and depicted with a white '+' symbol. Right: Percentage residuals of the fit; the color-scale is reported in the color bar at the top. The black cross represents the fit localization, while the black curve is the FWHM of the central beam at 600 MHz.

order to optimize hardware resource utilization while not sacrificing consistency of results, all cluster tasks are mapped to a Docker container. This provides the capability to transform and adapt the analysis pipeline around resource bottlenecks for tasks with significantly different CPU, RAM and I/O requirements. Although the pipeline is designed to be hardware agnostic, in its current configuration it processes the 1024 frequency channels of a baseband dump over 64 containers, each with an allocation of 2 cores and up to 16 GB of RAM (depending on the dump length) running in parallel on the analysis cluster, where each node consists of a dual socket Intel® Xeon® CPU E5-2630 v4 2.20 GHz and 128 GB of RAM.

In each container, two frequency channels are analyzed in parallel at a time until all the channels are processed. A separate dedicated task monitors the status of all the analysis containers and performs any ancillary post-processing routines. This configuration was chosen to reduce the memory pressure on the cluster but could be easily modified to optimize for CPU cycles or I/O bandwidth by either spawning a different number of containers or by modulating the hardware resource budget for each container. Additionally, due to the distributed design of the pipeline framework, multiprocessing designs can be avoided in favor of ease of prototyping, reduced complexity and more accessible single-thread performance.

5.2. *Pipeline tasks*

The baseband analysis pipeline is composed of three stages: refinement, localization and analysis, as shown in Fig. 3, each consisting of multiple tasks, as described in the following sections. Each stage takes approximately one hour to complete on the current analysis cluster, with the time budget being dominated by beamforming operations described in §3.

5.2.1. *Refinement stage*

At the refinement stage, the initial parameters of the signal measured by the real-time pipeline are improved using the baseband data. The uncertainties measured by the real-time pipeline are between 16–128 ms for the burst TOA, between 1.6–25 pc cm^{-3} for its DM and

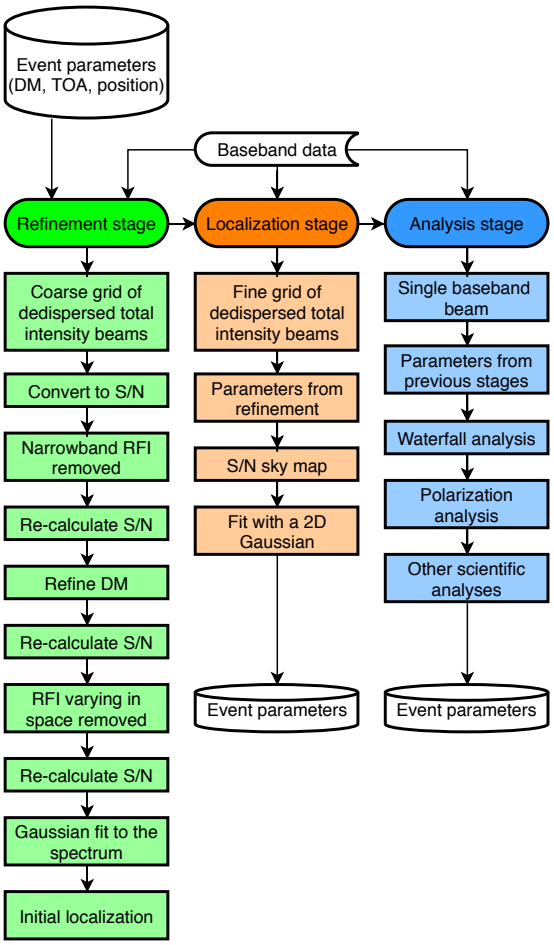

Figure 3. Schematic of the automated pipeline used to process CHIME/FRB baseband data. The colors represent the three different stages of the pipeline, indicated at the top of each branch and highlighted with darker shades and rounded symbols. Single tasks forming the three stages are depicted below each branch and are highlighted with lighter colors and squared symbols. Data and metadata are depicted in white and with different symbols. See text for details.

$\sim$ 15 arcmin for its position. If available, we use as initial localization guess the sky position derived by a parallel offline analysis pipeline called the *intensity localization pipeline*. The intensity localization pipeline refines the source sky position measured by the real-time pipeline by using the frequency-resolved total intensity of the signal detected by multiple static FFT beams. This increases the position accuracy of the source to a few arcminutes (Scholz et al., in prep.).

A set of beams is formed around the initial position and coherently dedispersed at the initial DM estimate. The grid of beams, used to refine the initial localization, is currently configured to produce 3 beams in the NS direction and 10 beams in the EW direction. The larger number in the EW direction is chosen because of the smaller number of independent baselines of the telescope in this direction, which makes the initial EW guess uncertain. The total size of the grid is $\sim 1.1 \times 0.25$ deg^2. It may happen, however, that the initial position is too far away from the true source position, especially for sources detected in far side-lobes of the primary telescope beam. In these rare cases, it is necessary to manually search for a better initial position. We are currently working to further increase the tolerance of the refinement stage to initial positions with large offsets by using an MCMC code (Scholz et al., in prep.).

The total intensity is calculated for each beam as a function of frequency and time. The resulting array is downsampled in time such that the pulse width measured by the real-time pipeline spans three bins. The S/N is calculated for each frequency channel and time bin by normalizing the off-pulse RMS to zero average and unitary standard deviation independently in each frequency channel and beam. In order to find the off-pulse region, an average profile is calculated and any time bin above a S/N of three is removed. This algorithm is applied iteratively in three steps to gradually improve the S/N. In the first step, narrow-band radio-frequency interference (RFI) is excised. This is done by calculating an average S/N for each channel and time bin across all the beams and then using a metric given by the standard deviation divided by the average in each channel. After normalizing this metric to have zero median and unity RMS, and after removing an arbitrary slope to account for intrinsic spectral variations, we remove channels above a specified threshold, currently set to three. This process is iterated upon until no channels above the threshold are found. In the second step, the initial DM value is incoherently refined with a brute-force algorithm that maximizes the S/N with a resolution of 0.1 pc cm^{-3}. Finally, the RFI mask is refined by using the intensity variation of the signal across multiple beams. This is performed by calculating an average value of the RMS in each beam and channel. For each channel, we define a metric given by the fractional difference between the maximum and minimum value of the total intensity across all the beams. In an iterative process, we normalize the metric across all channels and mask values larger than three. This technique is effective in removing RFI for which the intensity changes rapidly in different sky directions. The same RFI mask is applied to all the beams. After each of these three steps, the S/N array is re-calculated with the algorithm described above. In order to increase the S/N of narrow-band bursts, a Gaussian function is fitted to the spectrum of the beam displaying the strongest signal. If the maximum value of the Gaussian within the bandwidth

is larger than 3 times the RMS, channels outside 3 times the standard deviation of the function are removed.

In order to refine the initial source localization, a single S/N value is calculated for each beam of the grid. The position of the beam with the highest S/N value is used as the initial guess position for the subsequent localization stage.

5.2.2. *Localization stage*

As discussed in §4, the source position is further refined by forming a compact grid of largely overlapping beams to map the sky region around the initial guess determined in the previous stage. We tested the optimal size and filling factor of the beam grid to optimize computing power and localization precision. A good compromise in terms of number of beams and localization precision was found to be a grid composed of 5×5 beams covering a square sky region with a side of ~ 0.3 deg, i.e. similar to the FWHM of a single beam, as illustrated in Fig. 2.

The beams are dedispersed to the DM calculated in the refinement stage and the rest of the parameters obtained there are applied. A single pulse profile is obtained for each beam by averaging the frequency channels together. The time bin corresponding to the strongest S/N value across all the beams is selected. The S/N of the profiles in different beams corresponding to the same time bin is used to form a map of the signal intensity at different locations on the sky. A 2D Gaussian function approximating the beam shape is then fitted to this S/N map to obtain the refined source position and its uncertainty, as described in §4.

5.2.3. *Analysis stage*

A single tied-array beam is formed at the source direction calculated in the localization stage. The parameters obtained in the refinement stage are once again used to maximize the S/N. The resulting data product is a complex-valued array resolved in time, frequency and polarization, with the native time and frequency resolutions presented in §2. To facilitate subsequent scientific analyses, this complex array is also converted to power values containing information on the four Stokes parameters.

A series of additional analyses is performed on these arrays to produce scientific results. The analyses include the characterization of the polarization properties of the signal, such as Faraday rotation, polarization fraction and polarization angle. This will be detailed in a forthcoming paper (Mckinven et al., in prep.) The higher temporal resolution allows us to study the spectro-temporal properties of the bursts in detail, theoretically up to nanosecond timescale. This analysis is also important to study various propagation effects on the bursts. For example, potential strong gravitational lensing of the bursts at millisecond timescales is being studied by correlating the phases of beamformed baseband data as a function of time-lag. The baseband data of CHIME/FRB will also be used in the future to correlate the TOA of the signal with outrigger telescopes to perform very-long-baseline interferometry (VLBI) localization of the sources with an expected precision smaller than ~ 50 milli-arcseconds (e.g. Leung et al. 2020).

At the end of the pipeline, the science output is stored in a centralized database of CHIME/FRB detections and diagnostic plots are displayed in a web viewer for inspection.

6. CHARACTERIZATION OF THE LOCALIZATION CAPABILITY

When localizing a source, it is essential to report a reliable uncertainty region. In addition to statistical uncertainties, a number of systematic effects are expected to affect the localization of FRBs detected by CHIME/FRB. Some of these systematics can be modeled and corrected while others remain poorly understood and must be estimated through sources with known positions. The methodology for estimating these systematics is presented below.

Firstly, using interferometric observations of bright sources collected with the cosmology backend of CHIME (Newburgh et al. 2014), we identified a unplanned rotation of the telescope's structure about an axis oriented parallel to the local zenith, about 4 arcminutes clockwise as viewed looking down this axis towards the Earth. This was corrected by updating the assumed feed positions prior to beamfoming. Localization errors of a few arcminutes are also expected from the thermal expansion of the metal structure of the telescope in the SN direction, which uniformly scales the baseline lengths. The thermal expansion coefficient was similarly measured using interferometric data collected over a range of ambient temperatures. We correct single-pulse localization for this effect post-facto by using the measured expansion coefficient and the ambient temperature at the time of observation. A small pointing error, up to a few tenths of an arcminute, is also expected from time-variable ~ 10 ps errors in the distribution of the telescope clock between the digitizers for feeds on the east two cylinders (located in the East Receiver Hut) and the west two cylinders (located in the West Receiver Hut). These clock errors are continuously monitored by injecting a common noise signal into a digitizer located in each hut and examining the phase of the cross-correlation.

Using these monitoring data, the induced localization error is also corrected post-facto.

In addition to the effects we have accounted for, there are several other effects that could degrade our localization accuracy. For example, we have calibrated the telescope orientation around the vertical axis, but have not accounted for any tilt in either the NS or EW directions. Nonuniformities in the telescope structure could also cause offsets in the localization. Also, feed-to-feed variations in the primary beam will introduce a direction-dependent phase that is difficult to correct and will limit localization precision. These beam variations are currently being mapped through a campaign of holographic interferometry (Newburgh et al. 2014; Berger et al. 2016), required for CHIME's cosmology science. These measurements could be used to assess and even correct phase variations prior to beamforming. This remains to be pursued and currently no explicit attempt is made to correct for this effect. The consequence of these and all other unmodelled systematics is characterized by localizing pulses from multiple pulsars and the repeating FRB 20180916B, whose positions were accurately determined by other instruments.

After applying all modelled localization corrections, we use a sample of 56 pulses from 15 sources with a known position to perform an empirical calibration of our localizations and their uncertainty. This sample is presented in Table 1. The sources have been chosen to cover a large fraction of CHIME parameter space in terms of S/N and declination. We plan to further expand this sample in the future to have an even better coverage of the telescope parameter space. We assume that our telescope coordinates X and Y (described above) are the natural coordinates that describe the space in which systematic errors live. The empirical calibration transforms our localizations as follows:

$$\theta_x^i \pm \sigma_x^i \rightarrow \left(\theta_x^i + \delta_x\right) \pm \left(\sqrt{\left(\alpha\sigma_x^i\right)^2 + \Sigma_x^2}\right)$$
$$\theta_y^i \pm \sigma_y^i \rightarrow \left(\theta_y^i + \delta_y\right) \pm \left(\sqrt{\left(\alpha\sigma_y^i\right)^2 + \Sigma_y^2}\right), \quad (4)$$

where the index i runs over the pulses in the sample, and the calibration has 5 free parameters: a scaling of the statistical uncertainty α, systematic offsets in each direction δ_x and δ_y; and a systematic floor in the uncertainty in each direction Σ_x and Σ_y. We adjust the parameters in Eq. (4) based on the following criteria:

1. The weighted mean of the position errors is zero (affects mainly δ_x and δ_y).
2. The reduced χ^2 is unity for low S/N events (affects mainly α).
3. The reduced χ^2 is unity for high S/N events (affects mainly Σ_x and Σ_y).

We empirically define the cutoff between weak and bright events to be S/N = 60 and verify that the fit is not very sensitive to this threshold. The results of this procedure are shown in Fig. 4, and yield $\delta_x = 0.14\pm0.05$ arcmin, $\delta_y = 0.19\pm0.05$ arcmin, $\alpha = 2.1\pm0.1$, $\Sigma_x = 0.28 \pm 0.02$ arcmin and $\Sigma_y = 0.23 \pm 0.02$ arcmin. The value of the reduced chi-square test for the pulsar sample after the correction is $\chi^2_{\rm red} = 0.8$, validating the results of the procedure. No systematic deviation is observed as a function of DM or X and Y coordinates. We apply these corrections to the position of FRBs and other sources localized with the baseband data of CHIME/FRB.

Table 1. Pulses used as a test sample to verify the localization capability of CHIME/FRB. The source name, the number of pulses used for each source, the DM (Manchester et al. 2005) and the reference for the source position used are reported.

Source	Pulses	DM	Reference
		$\mathrm{pc\,cm^{-3}}$	
PSR B0148−06	2	25.7	Deller et al. (2019)
PSR B0329+54	2	26.8	Deller et al. (2019)
PSR B0355+54	8	57.1	Li et al. (2016)
PSR B0531+21	3	56.8	McNamara (1971)
PSR B0643+80	6	33.3	Hobbs et al. (2004)
PSR B1541+09	2	35.0	Chatterjee et al. (2009)
PSR B1642−03	1	35.8	Deller et al. (2019)
PSR B1811+40	5	40.8	Hobbs et al. (2004)
PSR B2210+29	3	41.6	Deller et al. (2019)
PSR B2310+42	6	44.2	Chatterjee et al. (2009)
PSR B2315+21	5	74.5	Deller et al. (2019)
PSR J1652+2651	1	17.3	Lewandowski et al. (2004)
PSR J2139+2242	1	20.9	Sayer et al. (1997)
PSR J2346−0609	1	22.5	Deller et al. (2019)
FRB 20180916B	10	348.8	Marcote et al. (2020)

7. CONCLUSIONS

The automated pipeline to process data stored by the triggered baseband recording system of CHIME/FRB has been presented. Baseband data are stored automatically following the detection of an FRB candidate by the CHIME/FRB real-time pipeline. These data are then processed with a dedicated pipeline run using commercial platforms such as Airflow and Docker to make it easily scalable, portable and making optimal use of machine resources. The pipeline is able to produce sci-

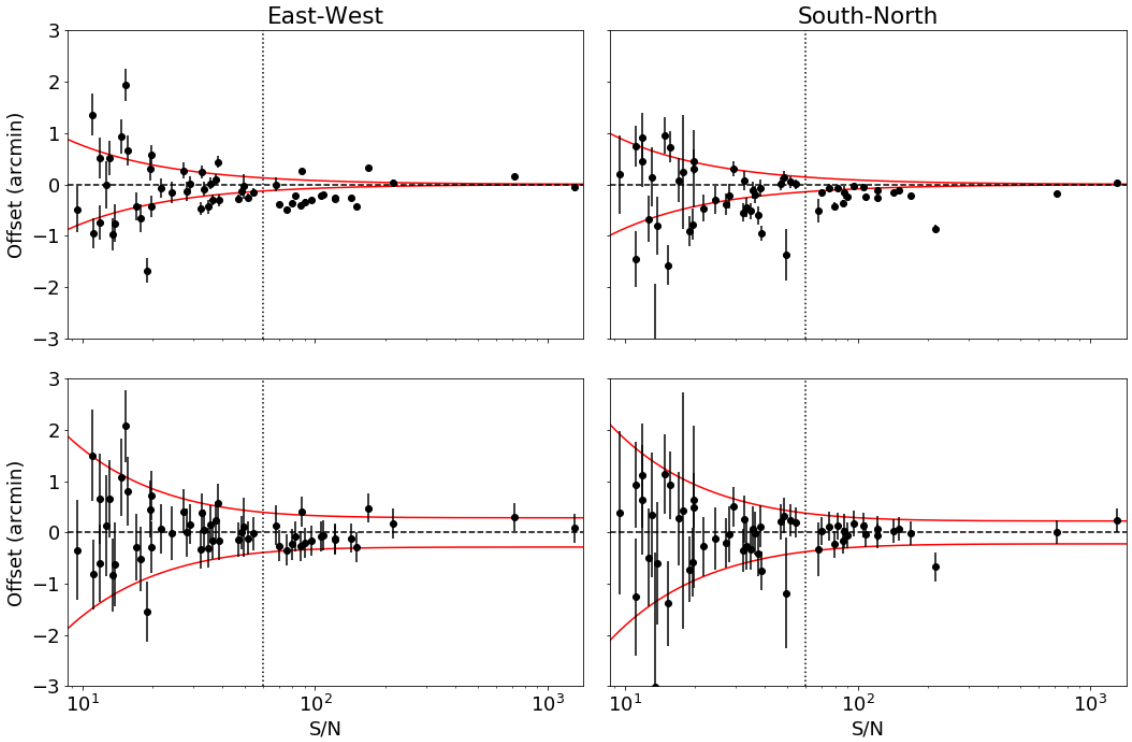

Figure 4. Measured offset of pulses from the sources presented in Table 1 with respect to their expected position as a function of the detected S/N. The offset is presented in the local reference system of the CHIME telescope. Left: offset in the EW direction; right: offset in the SN direction. The top panels report values and uncertainties obtained after including all modelled systematic effects, while the localizations in the bottom panels have been calibrated to have zero weighted average and unitary $\chi^2_{\rm red}$. The horizontal dashed line highlights zero offsets, the vertical dotted line represents an empirical cutoff between weak and bright pulses. Red curves are the theoretical uncertainty limits obtained from Eq. (3) for sources at zenith, before (top) and after (bottom) correcting for unmodelled systematics. Note that some of the pulses are actually detected away from zenith and, therefore, their uncertainties are larger by a factor $\cos(\theta)^{-1}$ (Eq. 2).

ence output without human intervention on a hourly timescale on our machines.

After the initial parameters of the candidate burst are refined, the baseband pipeline localizes its position on the sky. A grid of tied-beam arrays is formed around the initial position and the resulting intensity map is fitted with a 2D Gaussian model. Localization of a sample of known sources enabled us to validate the technique and to characterize localization uncertainties. The best localization precision currently achievable for bright bursts is

$$\begin{aligned} \sigma_x &\sim 17\,\text{arcsec} \\ \sigma_y &\sim 14\,\text{arcsec}, \end{aligned} \tag{5}$$

The results of this localization algorithm are being used for different scientific goals such as correlating FRB positions with large-scale structures in the Universe and localizing low-DM FRBs to host galaxies. We plan to further reduce the impact of systematic effects in the future by improving our modelling and increasing the test sample of known sources used.

Finally, the pipeline produces a beamformed baseband array of data resolved in time and frequency to enable different analyses. These include the study of FRB polarization properties, of spectro-temporal structures of the bursts and of propagation effects such as gravitational lensing. It will also be used in the future to cross-correlate the signal TOA at multiple CHIME/FRB outrigger telescopes under construction to localize sources with tens of milli-arcsecond precision. This will not only enable host-galaxy identification, and hence redshift determinations, but even determine the source location within the host for the majority of CHIME/FRB detected FRBs.

ACKNOWLEDGMENTS

D.M. is a Banting Fellow. R.M. recognizes support from the Queen Elizabeth II Graduate Scholarship and the Lachlan Gilchrist Fellowship. V.M.K. holds the Lorne Trottier Chair in Astrophysics & Cosmology, a Distinguished James McGill Professorship and receives support from an NSERC Discovery Grant and Gerhard Herzberg Award, from an R. Howard Webster Foundation Fellowship from CIFAR, and from the FRQNT CRAQ. C. L. is supported by the U.S. Department of Defense (DoD) through the National Defense Science & Engineering Graduate (NDSEG) Fellowship. M.D. is supported by a Killam fellowship, NSERC, CIfAR, and FQRNT. P.S. is a Dunlap Fellow and an NSERC Postdoctoral Fellow. The Dunlap Institute is funded through an endowment established by the David Dunlap family and the University of Toronto. FRB research at UBC is supported by an NSERC Discovery Grant and by the Canadian Institute for Advanced Research. The CHIME/FRB baseband system is funded in part by a CFI John R. Evans Leaders Fund award to IHS.

REFERENCES

Bandura, K., Cliche, J. F., Dobbs, M. A., et al. 2016, Journal of Astronomical Instrumentation, 5, 1641004, doi: 10.1142/S225117171641004X

Berger, P., Newburgh, L. B., Amiri, M., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9906, Proc. SPIE, 99060D, doi: 10.1117/12.2233782

Chatterjee, S., Brisken, W. F., Vlemmings, W. H. T., et al. 2009, ApJ, 698, 250, doi: 10.1088/0004-637X/698/1/250

CHIME/FRB Collaboration. 2018, ApJ, 863, 48, doi: 10.3847/1538-4357/aad188

CHIME/FRB Collaboration, Amiri, M., Bandura, K., et al. 2019a, Nature, 566, 235, doi: 10.1038/s41586-018-0864-x

CHIME/FRB Collaboration, Andersen, B. C., Bandura, K., et al. 2019b, ApJL, 885, L24, doi: 10.3847/2041-8213/ab4a80

Cordes, J. M., & Lazio, T. J. W. 2002, ArXiv Astrophysics e-prints

Deller, A. T., Goss, W. M., Brisken, W. F., et al. 2019, ApJ, 875, 100, doi: 10.3847/1538-4357/ab11c7

Fonseca, E., Andersen, B. C., Bhardwaj, M., et al. 2020, ApJL, 891, L6, doi: 10.3847/2041-8213/ab7208

Hankins, T. H. 1971, ApJ, 169, 487, doi: 10.1086/151164

Hessels, J. W. T., Spitler, L. G., Seymour, A. D., et al. 2019, ApJL, 876, L23, doi: 10.3847/2041-8213/ab13ae

Hobbs, G., Lyne, A. G., Kramer, M., Martin, C. E., & Jordan, C. 2004, MNRAS, 353, 1311, doi: 10.1111/j.1365-2966.2004.08157.x

Kumar, P., Shannon, R. M., Osłowski, S., et al. 2019, ApJL, 887, L30, doi: 10.3847/2041-8213/ab5b08

Leung, C., Mena-Parra, J., Masui, K., et al. 2020, arXiv e-prints, arXiv:2008.11738. https://arxiv.org/abs/2008.11738

Lewandowski, W., Wolszczan, A., Feiler, G., Konacki, M., & Sołtysiński, T. 2004, ApJ, 600, 905, doi: 10.1086/379923

Li, L., Wang, N., Yuan, J. P., et al. 2016, MNRAS, 460, 4011, doi: 10.1093/mnras/stw1262

Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J., & Crawford, F. 2007, Science, 318, 777, doi: 10.1126/science.1147532

Lorimer, D. R., & Kramer, M. 2004, Handbook of Pulsar Astronomy, Vol. 4 (Cambridge University Press)

Manchester, R. N., Hobbs, G. B., Teoh, A., & Hobbs, M. 2005, AJ, 129, 1993, doi: 10.1086/428488

Manchester, R. N., & Taylor, J. H. 1972, Astrophys. Lett., 10, 67

Marcote, B., Nimmo, K., Hessels, J. W. T., et al. 2020, Nature, 577, 190, doi: 10.1038/s41586-019-1866-z

Masui, K., Lin, H.-H., Sievers, J., et al. 2015, Nature, 528, 523, doi: 10.1038/nature15769

Masui, K. W., Shaw, J. R., Ng, C., et al. 2019, ApJ, 879, 16, doi: 10.3847/1538-4357/ab229e

Masui, K. W., & Sigurdson, K. 2015, PhRvL, 115, 121301, doi: 10.1103/PhysRevLett.115.121301

McNamara, B. J. 1971, PASP, 83, 491, doi: 10.1086/129160

Newburgh, L. B., Addison, G. E., Amiri, M., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9145, Proc. SPIE, 91454V, doi: 10.1117/12.2056962

Ng, C., Vanderlinde, K., Paradise, A., et al. 2017, in 2017 XXXIInd General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS), 1–4

Petroff, E., Bailes, M., Barr, E. D., et al. 2015, MNRAS, 447, 246, doi: 10.1093/mnras/stu2419

Petroff, E., Barr, E. D., Jameson, A., et al. 2016, PASA, 33, e045, doi: 10.1017/pasa.2016.35

Platts, E., Weltman, A., Walters, A., et al. 2019, PhR, 821, 1, doi: 10.1016/j.physrep.2019.06.003

Rafiei-Ravandi, M., Smith, K. M., & Masui, K. W. 2019, arXiv e-prints, arXiv:1912.09520. https://arxiv.org/abs/1912.09520

Sayer, R. W., Nice, D. J., & Taylor, J. H. 1997, ApJ, 474, 426, doi: 10.1086/303446

Spitler, L. G., Scholz, P., Hessels, J. W. T., et al. 2016, Nature, 531, 202, doi: 10.1038/nature17168

Yao, J. M., Manchester, R. N., & Wang, N. 2017, ApJ, 835, 29, doi: 10.3847/1538-4357/835/1/29