BRIEF REPORT

Exploring the Cosmic Dark Ages with ALBATROS

Samuel Careau[a,1] **and H. Cynthia Chiang (supervisor)**[a]

[a]Department of Physics, McGill University, Montréal, Quebec H3A 2T8, Canada

This manuscript was compiled on December 5, 2022

ALBATROS is an interferometer designed to record redshifted 21-cm neutral hydrogen lines at <30 MHz, which corresponds to the electromagnetic waves emitted during the cosmic dark ages, a period of the universe that remains unexplored to this day. Each antenna records baseband data which is correlated offline. Due to radio contamination from artificial sources the interferometer needs to be set up in an isolated areas. As such, Marion Island (South Africa) and Expedetion Fjord (Nunavut, Canada) were chosen in order to map both hemispheres.

Cosmology | Interferometry | Cosmic Dark Ages

The cosmic dark ages were an early period of the universe where everything was suspended in a fog of neutral hydrogen. It was preceded by recombination, a period where the universe cooled down enough to allow protons and electrons to interact and combine to create hydrogen and followed by the cosmic dawn, a period characterized by the birth of the first stars, then reionization, a period where stars and other macrostructures began to form in large amounts. The cosmic dark ages are a period characterized by very little activity; however, the hyperfine transitions within the neutral hydrogen atoms resulted in the 21-cm lines which are redshifted to <30 MHz. (1)

Redshift is a fairly simple concept that arises from the expansion of the universe: according to Hubble's law, the speed at which the universe is expanding is given by:

$$v = H_0 d \qquad [1]$$

where v is the speed of expansion, H_0 is the present Hubble parameter and d is the proper distance. (2) Due to the cosmic expansion, the waves travelling through space over long distances inevitably become stretched as well, increasing their wavelength while decreasing their frequency. The redshift parameter z can be found by using the equation:

$$z = \frac{\lambda_{obs} - \lambda_{rest}}{\lambda_{rest}} \iff z + 1 = \frac{\lambda_{obs}}{\lambda_{rest}} \qquad [2]$$

where λ_{obs} is the observed wavelength of an EM wave while λ_{rest} is the rest wavelength. (2) The higher the redshift parameter is, the older a wave is: highly redshifted waves were produced closer to the birth of the universe. For reference, recombination has $z \approx 1000$ and reionization has $z \approx 6$. (3)

Observation of these redshifted 21-cm emission lines may give us insight concerning the dark ages of the universe, an unexplored area of astrophysics and cosmology. While the cosmic dark ages are unexplored to date, Grote Reber used a telescope array to catch glimpses of the cosmos at approximately 2 MHz in the sixties, which is the current record for the lowest frequency measured. (4) One reason for the lack of exploration of this range of frequencies is because the ionosphere can be opaque at these frequencies - factors for opacity include the geographical location, time of year, and time of the day. The goal of the Array of Long Baseline Antennas for Taking Radio Observations from the Sub-antarctic/Seventy-ninth parallel (ALBATROS) is to create a high-fidelity map of the Milky Way at low frequencies. Radio interferometry is suitable for this kind of mapping because it offers high resolution at a relatively low cost. By using long baselines of ∼10 km, the interferometer allows for a high angular resolution as defined by the Rayleigh criterion:

$$\theta_{min} = 1.22 \frac{\lambda}{D} \qquad [3]$$

where θ is the angular resolution in radians, λ is the observed frequency, and D is the diameter of the telescope or interferometer. Using $\lambda = 10\ m$ (f = 30 MHz) and $D = 10\ km$, the theoretical resolution of ALBATROS is:

$$\theta_{min} = 1.22 \frac{10\ m}{10^4\ m} = 1.22\ mrad = 4.19'$$

The nature of the experiment requires the usage of an analog-to-digital. The Nyquist frequency is a direct consequence of the Nyquist–Shannon sampling theorem: in summary, in order to be able to reconstruct a sinusoidal wave that was previously digitized, the sampling rate must be equal (or greater) to twice the critical frequency, or the highest frequency that is being recorded. In other words:

$$f_c \leq f_s/2 \qquad [4]$$

Where f_c is the critical (maximum) frequency with units of Hertz and f_s is the sampling frequency with units of samples/second. f_c is also referred to as the Nyquist frequency. (5) In the case that $f_s > 2f_c$, the signal is over-sampled, with the only drawback that information is in excess and takes up unnecessary storage in the case of digital data.

Due to the remote locations of the arrays and the long baseline between individual stations, the autonomous stations equipped with solar panels and methanol fuel cells are used to collect data over several months. This data will then be retrieved physically and subsequently analyzed in Montreal. During these few months of inaccessibility, we would like to monitor the experiment - as such we want to send daily plots and diagnostics back to Montreal to ensure that everything is running smoothly in the high arctic. Furthermore, we'd like to introduce a flagging system so that we can automatically detect faulty data and send a warning along with diagnostics to Montreal. As such, a Starlink RV unit was modified in order to allow us to send data from these locations to Montreal through satellite internet. This unit is connected to a commercial satellite internet network ran by SpaceX, which allows any antenna from ALBATROS to be connected to the internet.

[1]To whom correspondence should be addressed. E-mail: samuel.careau@mail.mcgill.ca

ALBATROS

ALBATROS is an interferometric experiment that aims to map a high-resolution galactic foreground of the Milky Way below 30 MHz in order to pave the way for future Cosmic Dark Ages exploration. It consists of small, autonomous antennas in remote locations that were chosen for their low RFI (radio frequency interference) contamination and good ionospheric conditions. There are two permanent installations and one temporary installation: the latter is in Uapishka Station, Quebec, Canada which is used for testing improvements before deploying them to the two permanent sites at the McGill Arctic Research Station (MARS), near Expedition Fjord, Nunavut and Marion Island, South Africa, in the southern Indian Ocean (Fig. 1). The coverage of both hemispheres of the globe allow for a complete map of the cosmos above.

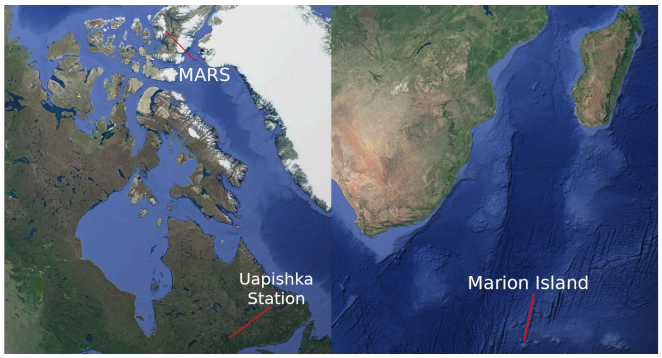

Fig. 1. Map showing the three locations where ALBATROS operates. On the left is North America and on the right is South Africa. These images were taken from Google Earth.

These antennas must operate independently from each other as due to the baseline of the interferometer of 10 km which has the consequence that each antenna requires an independent power source. The antennas are not in an area with any power grid: solar power and methanol fuel cells must be used in conjunction with an array of batteries in order to keep each antenna operational year round, with solar power dominating during the luminous months and the methanol power cells providing power when the luminosity is too low to sustain continuous operation.

The data is recorded on site and then retrieved manually to be correlated numerically. ALBATROS operates on a range from 1.2 MHz to 125 MHz (6) which allows us to detect signals from satellites (notably ORBCOMM satellites). These satellites assist us in refining our synchronization of data when processing data for offline correlation as they are always in the field of view of the antennas. Due to the data being recorded at the same time as these satellites on the same spectrum, correlation of the data in post-processing can be improved significantly.

ALBATROS uses a dipole antennas called Long Wavelength Array (LWA) antennas. They are dual-polarization and the diploes are perpendicular to each other. Mesh ground screens are placed below the antennas in order to reduce the noise from the ground which varies due to its ever-changing water content. Each polarization is fed to a Smart Network ADC Processor (SNAP) board, which is effectively a combination of an ADC and a field programmable gate array (FPGA). The ADC samples the RF signals at 250 Msamp/s over 2048 channels from 0-125 MHz. (6) This sampling rate is chosen according to the Nyquist frequency (eq. 4): according to the equation, the sampling rate should be $2f_c = 2(125\ MHz) = 250\ Msamples/s$. The FPGA also applies a FFT algorithm to incoming data, which turns the incoming RF signal into several amplitude coefficients which are each related to a sinusoidal wave of a specific frequency. The FPGA computes the auto-correlation (polarization with itself) and cross-correlation (both polarizations) and sends the correlated spectra to a Raspberry Pi 4 through the GPIO pins. The Raspberry Pi (RPi) then sends the data to the a USB multiplexer, which selects one from eight hard drives and keeps only that hard drive powered until it is no longer needed or full. This helps save on electricity while also preventing premature usage of the hard drives. When the weather and the time of the year is right, data can be physically retrieved by going to the two permanent installations and retrieving the hard drives. (6)

These electronics are contained within a Faraday cage roughly 100 meters away from the antenna in order to prevent any self-generated RFI and are connected to the antennas through coaxial cables. Balun are connected in series with the coaxial cables: they allow allow the connection between balanced and unbalanced lines without the need for impedance matching, which helps reduce interference from mismatched impedances. (6). A complete block diagram of the hardware used for each autonomous antenna can be found in fig. 2

Hacking the Starlink

The first step of connecting ALBATROS to the internet was solving the power issue. The Starlink RV unit has two parts: the router and the antenna (which also integrates a modem). The router accepts a 120 VAC, 60 Hz input which is then rectified and stepped down to 48 VDC. There are two issues with this stock configuration: firstly, the existing ALBATROS infrastructure is running entirely on 24VDC and converting to DC using an inverter and then stepping up the voltage to 48 VDC is quite inefficient. ALBATROS relies solely on solar power and methanol fuel cells making power conservation a priority. The second issue comes with having a router on this remote system: the network address translation (NAT) could prevent us from contacting individual antennas on the network. The Starlink router is a closed proprietary system which does not allow the users to configure options that are common on consumer routers: some examples are static IP addresses and port forwarding. Static IP addresses are necessary to keep track of which device has which IP address in order to communicate with it remotely. In the case that the router is on Dynamic Host Configuration Protocol (DHCP) then the RPis become essentially anonymous as their internal IP addresses are prone to sudden change.

The first point was addressed by using a power-over-Ethernet (PoE) injector to bypass the need for a router. The PoE injector removes the need for an inverter as PoE only uses DC. Furthermore, there would only be the need to step up the voltage from 24 VDC to 48 VDC which can be easily accomplished by using a DC/DC converter. To implement the PoE injector, we stripped the cable of its connector and measured the potential differences between the eight conductors. Initially we were expecting a number of conductors neighboring 20 wires due to the because the proprietary connector having 20 pins but the cable rubber jacket had indication that it was

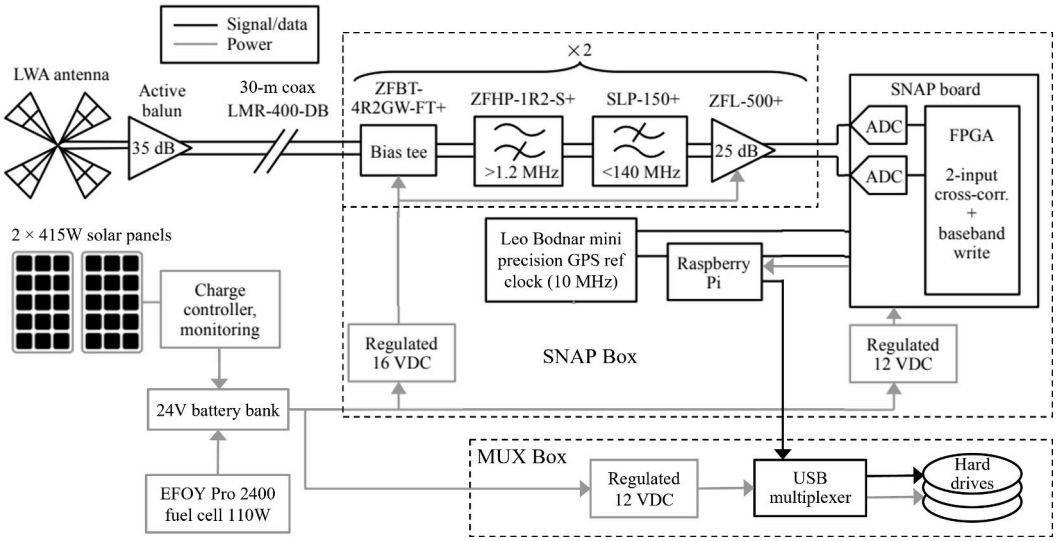

Fig. 2. Block diagram of an autonomous antenna setup. Dashed boxes represent custom-built Faraday cages. MUX stands for multiplexing. This figure was taken from Chiang et al. (6)

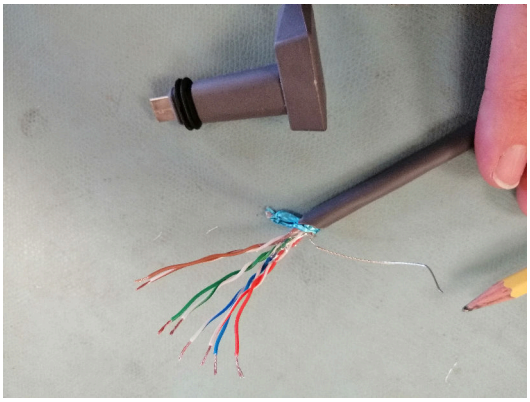

Fig. 3. 8 conductors within the Starlink cable that match the standard colored pairs of Cat5e Ethernet cables. The blue foil is the shielding while the silver wire is the ground.

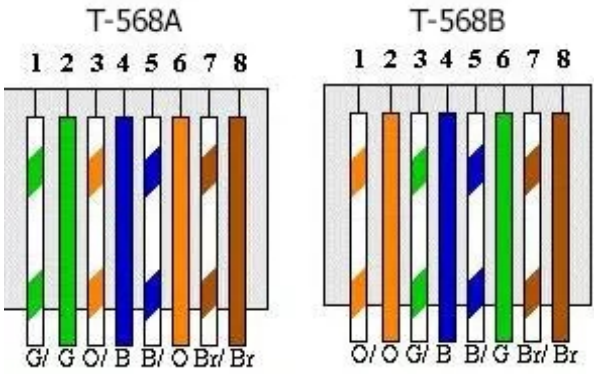

Fig. 4. The T568 standards are used for wiring Ethernet cables. Image source (7)

Cat5e, which would imply that it has eight conductors which turned out to be the case. The cable stripped of its conector can be seen in fig. 3.

The potential difference of each conductor with respect to ground was then measured: +48V on GrW (green-white), Gr (green), O (orange), OW (orange white), -48V on BrW (brown-white), Br (brown), BlW (blue-white), and Bl (blue). By selecting the correct settings on the PoE injector and connecting it to a 48 VDC power source the dish was powered and a connection was established between a computer and the Starlink network. See fig. 5 for a picture of the PoE injector.

A standard RJ45 connector (standard connector for consumer Ethernet cables) was crimped onto those eight conductors and plugged it into the PoE injector. The T-568B standard was used for this connection; a detailed figure can be found in fig. 4.The T-568B standard was chosen because it is more common in telecommunications: the other standard (T-568A) is mostly used in residential settings.

In retrospect, this was a risky endeavour as this kind of modification had not been recorded before. Fortunately we found a schematic produced by Oleg Kutkov (an embedded systems engineer who has done research in Starlink router/satellite disassembly in order to benefit repairs on the Ukrainian frontlines) in fig. 6. This schematic explains summarily how the optional Ethernet adapter works in conjunction with the Starlink router and satellite dish.

Diagnostics & Error Detection

Once the connectivity hardware was ready for testing, I started working on setting up a system to compile diagnostic data from the antennas in order to verify whether the data was collected in the right way and whether the electronics or other antenna hardware could be having issues. Diagnostic data is currently only collected directly on-site; the purpose of this system is to send diagnostic plots daily. Every hour of data collection returns approximately 36 kB in diagnostic data, so it is fairly quick and inexpensive to send it over the Starlink network and retrieve it. Plotting this data makes it easier to detect any anomalies; below is a sample plot of diagnostics collected at the 2021 expedition to Uapishka station in fig. 7

Each subplot has a different purpose. The temperature plot allows the team behind ALBATROS to check that temperatures are withing reasonable operating range. If the tempera-

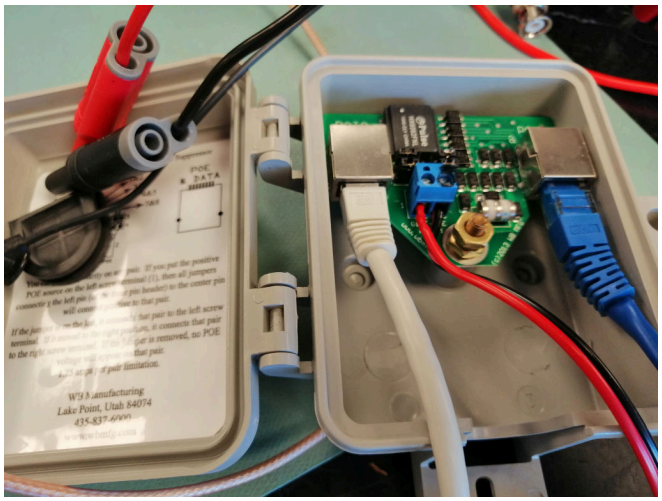

Fig. 5. The PoE injector connected to a 48V source. This setup draws less than 200W at its peak and about 100W with regular use.

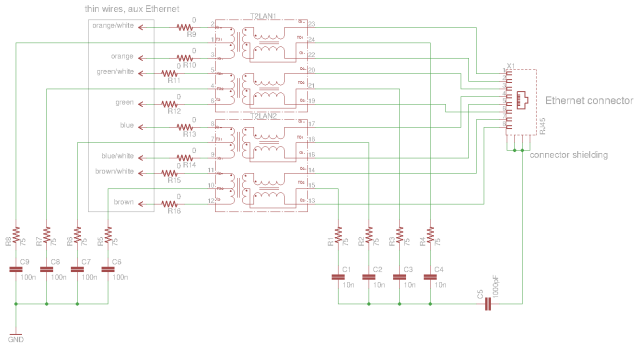

Fig. 6. A schematic produced by Oleg Kutkov while he was dissecting the optional proprietary Ethernet adapter for the Starlink. This schematic helped us confirm that the Starlink worked with PoE and that it only had eight conductors. Image source (8)

tures are too low, it could indicate a fault in the temperature sensor or another issue; if the temperatures are too high, it could indicate a problem with the hardware due to overheating. The FFT overflow counter is straightforward: it counts the number of times there is an overflow when computing the coefficients through the fast Fourier transform algorithm on the FPGA. If it is anything but zero, it indicates that there was a problem with the FFT computations. The time difference between sys_stop and sys_start indicates the difference in seconds between two subsequent data points by comparing the end of the recording of the previous point vs. the start of the recording of the subsequent point. These should always be in the range of a few seconds, anything negative or above a few seconds could indicate some problem with data saving or synchronization. The time differentials is the difference in time between the start of data collection of a point and the start of the start of subsequent point; likewise for the end of data recording ("time_sys_stop"). Most data is situated between 5 and 8 seconds; anything higher than that could indicate problems with disk write if the differential is recurring and anything lower than 2 would be problematic and would impact the graph described previously. The synchronization count describes timeouts on the FPGA and should be zero at all times. Finally, the accumulation count is a counter that goes up with every recording. The plot is differential and it should stay constant at 1 (increment of one between each data point), thus it any other value could indicate an issue various parts of the system.

The second error flagging method comes from analyzing data and spotting outliers. By using reference data that was deemed valid, we can compute a normal distribution for a single channel (or frequency) which gives us a reference frame for which data can be considered good or bad. Due to the data varying greatly depending on the local sidereal time (LST), the way I approached it was to convert local terrestrial time into local sidereal time so that the data taken on different days would be similar. For example, with the galaxy passing overhead every day at the same local sidereal time but not the same local terrestrial time, we should see a pattern repeat itself every sidereal day and a similar pattern repeat itself every solar day, with a difference of about four minutes per day. By accumulating data from each sidereal day binned by hour, we get a good reference frame for what the data should look like for each hour. One such example of a reference plot for approximately one hour of measurements for one polarization on a single channel can be found in fig. 8.

Using this normal distribution, it's easy to automatically detect outliers. Various methods can be used such as the root-mean-square (RMS) in order to define a warning threshold. Once such a an error is detected, it can be passed on to the flagging system and sent to over the Starlink network to be analyzed in Montreal.

I've also worked on introducing a binary flagging system for our scripts. Setting a binary number to 0 and changing bits according to the specific error encountered is easy to carry across different scripts and encodes a lot of information for a single variable. From the right, the first six bits encore the presence of an error or the lack of said error for each of the plots described above. The seventh bit encodes any error within the data itself, i.e. any outliers relative to the normal distribution.

Preliminary Results

After the data is processed from the time domain into the frequency domain, we can correlate the data and refine it into waterfall plots. In figure 9, the plots on the four edges are the waterfall plots, with the horizontal axis representing the frequency (in MHz), the horizontal axis representing the time (in minutes) and the color representing the amplitude (or power) of each sinusoidal wave computed through the FFT.

"Pol" stands for polarization: as such, "pol00" is the auto-correlation of the first polarization with itself, while "pol01" is the cross-correlation between the first and second polarizations of the antenna. The bottom right plot represents the phase difference between both polarizations, with the horizontal and vertical axes being the same as the other waterfall plots and the color being the phase difference in radians. The two middle plots are essentially histograms: each slice is a accumulation of the data obtained for each channel over the course of an hour.

Future Goals

ALBATROS is an ongoing experiment and as such much work remains to be done. We are currently working on one of the issues encountered by separating the Starlink router from its

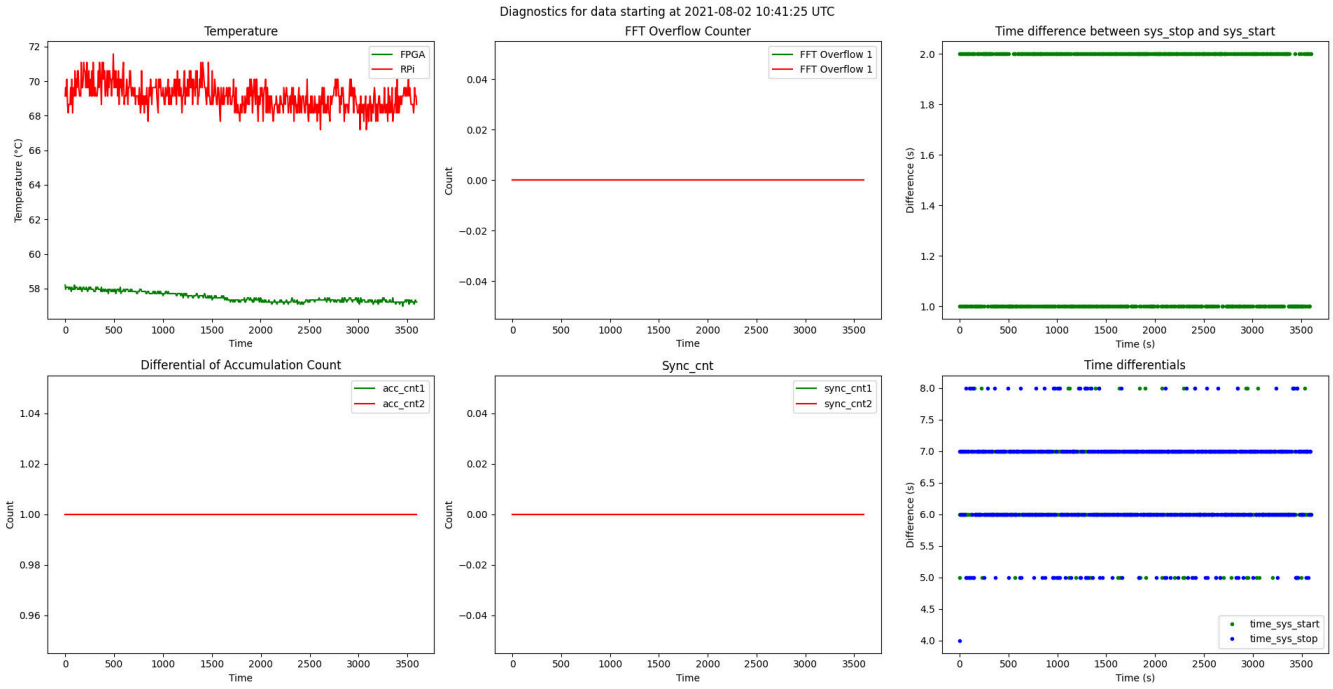

Fig. 7. This plot presents in a visual way the health of the antenna stations. Each subplot represents a different aspect: from the top left going clockwise, the temperature of key systems, the FFT overflow counter, the time differential for start/stop signals, the time differentials between each data point writing to disk, the synchronization count, and the accumulation count.

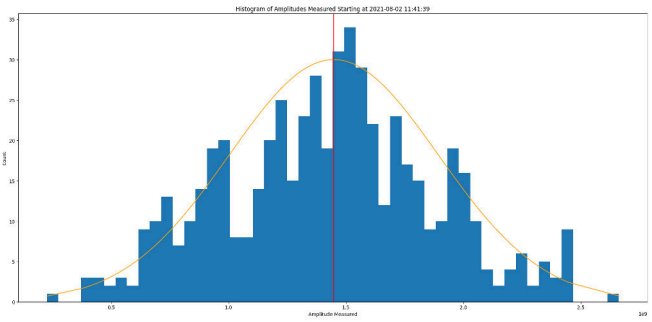

Fig. 8. This is a histogram of the counts of each amplitude measured starting at 2021-08-02 11:41:39 UTC. This histogram has 50 bins. The channel chosen is channel 135, which corresponds to 30.02 MHz in frequency domain. The red line is the median of the distribution while the orange line is a non-normalized normal distribution. Note that this histogram only contains the data from one sidereal day and not the accumulation of several.

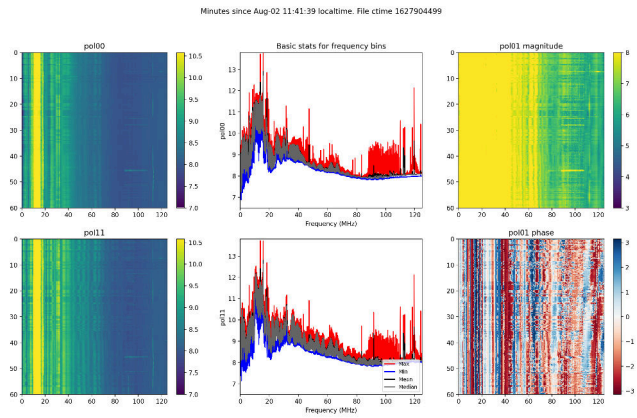

Fig. 9. These are quick plots of the cross-correlated spectra (pol01) and the auto-correlated spectra (pol00 and pol11). The center plots are plots of the frequency vs. the time vs. the amplitude of the FFT that was applied to the input signal. Each frequency channel has a a maximum (red line), a mean (black line), a median (gray line), and a minimum (blue).

satellite dish. Due to the fact that we are bypassing the router, we only get one internal IP address, which lead us to a new problem: how do we aggregate all of the diagnostic data onto a single device to produce plots and send them to Montreal? One possible option would be to use a travel router, which has very lax NAT rules. Another option is to have a central system which will gather data on one RPi and aggregate the compute the diagnostics on-site before sending the final image to Montreal. One of the issues with a centralized system would be reliability - if the RPi dies for any reason, the whole system ceases operation. Alternatives should be considered

Moreover we have to create the script which will collect the diagnostics from each individual RPi and then send them through the Starlink network to Montreal. We will also need this script to connect through secure shell (SSH). We are currently testing with a local machine: so far we have managed to get a server running and successfully connected to it. One other issue we are facing is that we cannot connect directly to the Starlink, but the Starlink can connect to us due to the way that satellite communications work. The external IP address of the Starlink dish can change at any time. The solution we are currently exploring is to have the Starlink establish a SSH connection to Montreal, and then we will connect to the Starlink via SSH within that SSH tunnel, a process known as reverse SSH tunneling. This is effectively our only option to

connect to the ALBATROS antennas reliably.

Furthermore, the scripts for error detection could be improved. The foundation of the program is in place but some features are missing: for example alternative methods of measuring errors could be implemented instead of using the standard deviations or RMS. Additionally, due to the large size of the data, it's nearly impossible to load multiple hours of data at once in a computer's memory, so sampling may need to be done in order to obtain a larger range of results for the normal distribution which should hopefully improve its accuracy in detecting errors.

ACKNOWLEDGMENTS. This journey I have taken part in would have not been possible without the support of several people. I'd like to personally thank H. Cynthia Chiang who has been an incredibly friendly and supportive supervisor over the course of this project. She has introduced me to the world of cosmology while I only knew about the world of astrophysics. I would also like to thank Eamon Egan for his expertise in various domains, but particularly networks and electronics. I would also like to thank Cherie Day and Mohan Agrawal for their help during various stages of this research journey.

1. JR Pritchard, A Loeb, 21 cm cosmology in the 21st century. *Reports on Prog. Phys.* **75**, 086901 (2012).
2. BW Carroll, DA Ostlie, *An introduction to modern astrophysics.* (Cambridge University Press), (2018).
3. IT Iliev, PR Shapiro, A Ferrara, H Martel, On the direct detectability of the cosmic dark ages: 21 centimeter emission from minihalos. *The Astrophys. J.* **572** (2002).
4. G Reber, Cosmic static at 144 meters wavelength. *J. Frankl. Inst.* **285**, 1–12 (1968).
5. JJ Condon, SM Ransom, *Essential radio astronomy.* (Princeton University Press), (2016).
6. HC Chiang, et al., The array of long baseline antennas for taking radio observations from the sub-antarctic. *J. Astron. Instrumentation* **09** (2020).
7. J Ellis, T568a and t568b wiring standards (year?).
8. O Kutkov, Reverse engineering of the starlink ethernet adapter (2022).