Fermilab LDRD Proposal

Project Title: Tianlai Data Analysis Center

Principal Investigator: Albert Stebbins
Lead Division/Sector/Section: PPD/Astrophysics
Co-Investigators (w/institutions):
John Marriner (Fermilab Emeritus)
Xuelei Chen (National Astronomical Observatories of China)
Reza Ansari (LAL, Université de Paris XI, Orsay)
Peter Timbie (University of Wisconsin, Madison)

Proposed FY and Total Budgets: (summary of budget page (in dollars))

	SWF	SWF OH	M&S	M&S OH	Contingency	Total
FY15						
FY16						
FY17						
Total						

SWF: Salary, Wages, Fringe   SWF OH: overhead on SWF
M&S: Material and Supplies M&S OH: overhead on M&S
Contingency (estimate of additional funds that might be required with justification)

Initiative: 2015 Broad Scope

Project Description

This project will fund the Computing Division support that is required for the analysis of data from the Tianlai[1] pathfinder 21cm intensity mapping redshift survey [1,2]. The Tianlai survey is a Chinese funded pilot project to demonstrate the feasibility of using a wide field of view, radio interferometer to map the density of neutral hydrogen in the universe. The resulting maps can be analyzed to extract cosmological parameters, such as the Dark Energy equation of state, by searching for the distinctive pattern of density fluctuations that are known as baryon acoustic oscillations, which were first measured by SDSS. The wide field of view, the large number of data channels, and the wide bandwidth present new, unsolved analysis challenges and the goal of this LDRD project is to invent or adapt the novel analysis techniques which will be required to extract science results from the data produced by the Tianlai survey.

There is a unique window of opportunity for Fermilab to be involved in solving the most important developmental issues for 21cm redshift surveys, with hardware just now being deployed on the ground. The opportunities will be different and probably less ground breaking in the coming years.

[1] Tianlai can be translated from Mandarin as "Heavenly Sound"

Significance

The discovery of accelerated cosmological expansion (Dark Energy) has prompted a flurry of experimental activity, as it is a robust low energy manifestation of physics beyond the standard model, and can be measured precisely by a variety of experimental techniques. Fermilab scientists have a strong interest in the study of Dark Energy and have contributed to the effort to study it through past, present and future large optical surveys (SDSS, DES, DESI, LSST) and through a broad range of theoretical efforts. Of the various techniques, the observation of Baryon Acoustic Oscillations (BAO) is most relevant for this proposal. Optical surveys have been extremely successful in mapping the inhomogeneities in the universe by measuring the distribution of luminous galaxies in redshift space (angle and position). These maps can be used to observe the characteristic BAO signal and to extract cosmological parameters, including the Dark Energy equation of state. As we map larger and more distant volumes of the universe, using optical techniques to measure precise redshifts becomes more challenging. However there are alternatives that may be more cost effective.

Galaxy redshifts can be measured at radio frequencies from the hyperfine 21cm emission of atomic neutral hydrogen (HI). However resolving individual galaxies at cosmological distances requires extremely large radio telescopes, *e.g.* the Square Kilometer Array. A radically different technique, *intensity mapping*, was proposed by collaborator Peterson [3] where one uses maps of 21cm emission where individual galaxies are not resolved. The 21cm line is unique in cosmology in that it is the dominant astronomical line emission for all positive redshifts, *i.e.* for all cosmological emission. So to a good approximation the wavelength of a spectral feature can be converted to a Doppler redshift without having to first identify the atomic transition. Intensity mapping makes cosmological redshifts surveys feasible with 100m class radio telescopes which only have angular resolution of a few arc-minutes. The overall promise of the intensity mapping technique was studied at Fermilab [5]. The direct determination of redshift using 21cm data can be compared to the optical technique, which requires the identification of a suitable subset of target galaxies (photometry), then obtaining an optical spectrum, and finally identifying some unique combination of emission and absorption lines that allow an unambiguous determination of the redshift for that galaxy (spectroscopy). SDSS produced both photometry and spectroscopy while deeper photometric surveys, *e.g.* DES and LSST, have left the spectroscopy to other projects *e.g.* DESI. Another reason to develop 21cm mapmaking techniques is for its future potential. 21cm emission and absorption occur even before galaxies form, *i.e.* during the cosmic "dark ages". The technique developed for this project can be extended to mapping inhomogeneities in the majority of the cosmological volume, which can only be seen during their dark ages.

There are a number of reasons why using the 21cm line in large scale cosmological redshift surveys is only now being studied in a pilot project:

1. It was not recognized that making "intensity maps" with telescopes that cannot resolve individual distant galaxies would be useful.
2. Advances in semi-conductor electronics have made the large, multi-element arrays necessary for these measurements increasingly feasible. Primary among the advances are a) the dramatic decrease in the noise temperature of room temperature preamplifiers, a factor of

~10 below the physical temperature, obviating the need for cryogenics, b) the exponential growth in capabilities for the necessary online data processing, c) the decrease in the electrical power required to operate the electronics.

3. Foreground emission in these bands is orders of magnitude larger than the 21cm emission and the possibility of removing it was not fully explored,
4. Many experienced radio astronomers harbor significant doubts whether technical issues including the ability to eliminate radio interference (RFI) and to calibrate multi-element antenna arrays can be overcome

The intensity mapping concept (1) has been widely accepted as valid primarily through simulations and the Tianlai array will establish the feasibility of the apparatus and data acquisition systems. This LDRD project is intended to address the analysis challenges described above (3 and 4). The PI has spent significant effort on theoretical approaches to the foreground removal problem and would like to test them. A successful demonstration would open the door to a most promising future for 21cm large scale structure (LSS) surveys.

There are a few other ongoing 21 cm intensity mapping pilot projects to which the goals of this LDRD proposal are relevant, but Tianlai is chosen for analysis because the PI has access to the data due to his long association with the project. Tianlai refers to two interferometric arrays of feed antennas that will constitute the 21cm intensity mapping telescope: the main array consists of 3 cylinder telescopes and next to this is an array of 16 six meter dishes. Installation is nearing completion in western China and should be fully operational by the end of the summer 2015. Initially the cylinders will have only a small fraction of the receivers they could accommodate, but if this pilot project is successful they will be outfitted with a full complement of receivers, which will greatly increase the sensitivity and eventually the number of cylinders will also be expanded [1]. While intensity mapping should work best with dedicated arrays like Tianlai the most successful 21cm intensity mapping result to date used the single dish non-interferometric Green Bank Telescope (GBT) [4]. The Tianlai collaboration includes two prominent members of the GBT collaboration: Peter Timbie (co-I) and Jeff Peterson (collaborator). Tianlai should outperform GBT as it has similar collecting area but has ~100 times more receivers and hence will be able to map the sky 100 times faster. The Tianlai site has considerably less RFI as well. The pilot project will map a volume of ~50 Gpc^3 ($0.77<z<1.03$ and 50% of the sky), which is comparable to that surveyed by the DES, with better spectral resolution, albeit coarser angular resolution and larger map noise.

The proposed data reduction and analysis is an R&D project because the multi-element feed arrays, the wide field of view, the cylindrical geometry of the main reflectors and the significant foreground radiation all present significant challenges in the analysis. The response to the challenges are not straightforward extensions of existing techniques, but requires developing, testing and analyzing different algorithms that have never been implemented (in a data reduction pipeline) nor tested on real data and inevitably will need to be refined. The goal of the LDRD project is to develop the “know-how” to analyze the data. The techniques developed are expected to be invaluable for future 21cm surveys such as the full-scale Tianlai array [1,2] or other similar projects, such as a natural follow-on survey in the Southern hemisphere. The unique skills gained by any RAs who choose to get involved will position them to be leaders in the nascent field of 21cm cosmology. The 21cm survey itself, covering a volume comparable to DES but at a higher redshift, will be an important contribution to cosmology providing new constraints on the dark energy equation of state. Tianlai is a small collaboration consisting of several scientists in China, a few in France, and even fewer in the US. A Fermilab Tianlai analysis center would play to the Fermilab strength in computational science and lead to a highly visible role in producing the cosmological science results.

Research Plan

a. Objective

The goal of this LDRD pilot project is to demonstrate that 21cm intensity mapping can be used to map large-scale structure (LSS) and place constraints on cosmological parameters. To do this we will develop data analysis techniques to go from radio interferometer data to 3D maps of HI LSS from which foregrounds have been cleaned. From these maps we will compute the power spectrum of HI clustering and from the power spectrum place constraints on cosmological parameters. This 2nd step will use standard algorithms and will not be discussed here.

b. Concepts to be demonstrated

To achieve the objective one must subtract radio foregrounds, which are predominantly synchrotron emission, and are $\sim 10^5$ larger than the 21cm emission. The two can be distinguished because the former have smooth spectra and the latter do not. The 21cm features we are most interested in are the small scale BAO `wiggles'. Unfortunately radio telescopes are diffraction-limited and the beam patterns depend on frequency, which mixes angular and frequency dependence. Since the foregrounds are not smooth in angle it is a nontrivial task to identify and subtract the smooth frequency foregrounds with sufficient accuracy to reveal the 21cm emission. It is easier if one goes to very small radial scales, but to get the most cosmological information out of the data one must remove the foregrounds over the largest range of scales possible. Developing and demonstrating the efficacy of methods that can optimally remove foregrounds is a prime goal of this project and is necessary to achieve the main objective.

To achieve the foreground subtraction goal as well as to make accurate 3D maps one must have a very accurate model of the electric field patterns and complex gain of every feed antenna. These can change with time and one will need to be able to measure these quantities from the interferometer data itself. Developing and demonstrating the efficacy of methods to model and calibrate large interferometric arrays is also necessary to achieve the main objective.

c. Methods

One can outline our methods by stepping through the data chain: stages of processing information starting with the telescope in China and ending with 3D maps of the structures in the universe. The data chain for the Tianlai cylinder interferometer starts with voltages from 92 dual polarization telescope feed antennas that are digitized and transformed into 1024 frequency channels via a Fast Fourier Transform (FFT) spanning 700-800 MHz. Each channel produces a time stream of complex amplitudes that is cross-correlated with every other channel for each frequency and each time interval. The cross-correlations are known as visibilities. They will be averaged over ~1 second intervals to reduce the volume of data and recorded on to tapes or disks which are shipped from the telescope to the National Astronomical Observatories of China in Beijing. The time ordered data (TOD) are written to tape and then shipped to Fermilab where they are archived. At Fermilab the TOD are scanned for RFI, which is deweighted and flagged. The remaining signal is used to self-calibrate the feeds and the visibilities are adjusted by the calibration. This internal self-calibration is only a relative calibration. The calibrated signal is averaged in time bins of ~1 minute. This refined TOD are broken into lengths of 1 sidereal day and combined into an “average sidereal day’s data” (ASD). The ASD is about 10 TB in size and is a rudimentary image of the raw data. The data for any given day is dominated by thermal noise in the amplifiers, which is uncorrelated between amplifiers, and therefore the visibility signal averages toward zero as more TOD is accumulated and averaged. As the data accumulates the noise contribution will decrease, and will ultimately be dominated by foregrounds, with the 21cm emission being a minor component which must be extracted. The next step in the analysis is to convert the visibilities into a map of the sky. This step is complicated by the need to carefully calibrate the absolute frequency and spatial response of each feed antenna. Finally, the foregrounds must be separated from the cosmic 21cm signal of interest to us. In the following paragraphs, we discuss key aspects of this procedure.

The data rate for the cylinder telescope can be computed as follows: there are 96 dual polarization feeds or 2×96 data streams. With 1024 channels the number of correlations consists of $1024\times(2\times96)^2$ =38M complex numbers. The on-site correlators generate these numbers at a rate of 100kHz but this is averaged down to ~1Hz sampling before they get to Fermilab. A 38M/s data rate is not large but these data need to shipped, read in, and processed continuously. We have not fixed the byte count nor the sample rate, and we do not know the uptime, but a rough guess is 1 petabyte of data per year. This LDRD funds two years of effort, which is sufficient to validate the methods.

The step of producing the ASD from the TOD is expected to be computationally simple, but involves a large enough data set that the assistance of a computer professional is highly desirable in order to make efficient use of the Fermilab computing facilities. With the proper tools in place, data processing can also be initiated by the PI and/or collaborators from the US, China, or France. The TOD arriving at Fermilab over time is large (~1 petabyte) and reducing from TOD to ASD involves processing all of the accumulated data. While the computational job by itself is significant there are other challenges. Although random thermal noise in different amplifiers does not contribute, on average, to the visibilities, other instrumental effects do. Even though Tianlai is located in an extremely radio quiet site, there will be some, human generated radio interference (RFI) that can contribute to the visibilities. RFI that is large compared to the thermal noise is easily eliminated, but persistent RFI that is hidden in the noise is more problematic. Natural sources, such as the sun, will be present in the data as well as transmissions from artificial satellites. There are also a number of possibilities for correlated noise generated in the observing apparatus itself. Much of the potential problem is eliminated by careful hardware design, but the feed antennas necessarily radiate as well as receive so that the amplifier thermal noise is broadcast to nearby feeds. None of these spurious signals will have the same sidereal period as the cosmic signal, and that difference is a key to eliminating some of the more subtle problems. As the ASD noise is reduced through additional observation, the subtle problems will become more apparent. For this reason, we expect that the TOD to ASD step will be repeated a number of times as we perfect our algorithms. The algorithms for this step, as well as the other steps, will be developed by scientific effort, most of which will be provided by participants in the Tianlai collaboration and coordinated by the PI of this proposal.

Calibration will rely primarily on phenomena that occur naturally in the data stream. The data stream contains all the visibilities (all pairs of feed antennas); the number of visibilities greatly exceeds the number of feed antennas. Since the response of the array can be characterized to a good approximation by the spatial and frequency response of the individual feeds, the redundancies inherent in the visibilities provide a good method for relative calibration. Absolute calibration should be done over the entire field of view of the telescope and can be done differently in different parts of the field. Earth rotation scans the fields through all right ascensions but not declination. Bright radio sources can be used as calibrated sources at declinations where they exist. Pulsars are not as bright but have unique time signatures. Artificial satellites are also good point sources that scan an angular path across the array. Co-I Peter Timbie and his students have developed EM simulations of the array and will take a leading role in understanding the response of the deployed antennas.

Several algorithms have been put forward for filtering out foreground contamination [7-9] some from within collaboration. All these will be tested on the Tianlai data. The residual after low pass filtering should be an image of the 21cm emission in redshift space, *i.e*. a 3D map of the neutral hydrogen distribution. The map will be filtered, meaning some of the 21cm distribution is lost during foreground removal, so this map must be accompanied by the filter. The map will also be accompanied by a noise model. The 3D map will be noisy even after a year of data, but the power spectrum should have high signal-to-noise. This power spectrum can be fit to cosmological parameters. Since 21cm emission is unpolarized, an important null test of all of these methods is performed by checking if the foreground subtracted maps have removed all significant linear and circular polarizations,.

The high-pass filtered, deconvolved map that is subtracted to obtain the 21cm map will be a high signal-to-noise map of Galactic and extra-Galactic polarized foreground emission with exquisite frequency resolution in the 700-800MHz band over half the sky. Currently there is no comparable dataset available, so this will be an important contribution to radio astronomy! The Galactic contribution will be dominated by synchrotron emission and the slope and curvature of the spectrum in each resolution element will give an indication of the energy distribution of high-energy electron cosmic rays along each line of sight. The polarization of the radiation will tell us about Galactic magnetic fields. Models of the Galactic cosmic ray distribution and magnetic fields that can be inferred from these maps are useful for Galactic science as well as dark matter indirect detection.

d. Timetable and Deliverables

The first set of deliverables, which we expect to be available early in the first year includes

1. Data visualization software and metrics for data quality.
2. The Average Sidereal Day (ASD) data i.e. an average over days of all the visibilities as a function of the rotation of the Earth.
3. A noise model of the apparatus.
4. Characterization of the spurious signals that contribute to the ASD.

The ASD will accumulate data over the course of the 2 years of the LDRD. A critical aspect of the proposal is that the ASD be made available to the Tianlai collaborators and that they be provided with access to the Fermilab computing farms to test and run their algorithms. Some collaborators may choose to transfer some or all of the ASD to other computing facilities and we intend to allow for that possibility.

A second set of deliverables that will also be required to do the science and can be expected by the end of the LDRD includes

1. A model of the response function for the full array, including frequency and spatial response of each feed and any significant cross-couplings between feeds; this model will include the response to specific high intensity point sources, to bright radio pulsars, and to satellites.
2. A study of the time variability of the calibration over long and short time baselines.
3. Intensity and polarization map of total (mostly foreground) emission as a function of angle and wavelength.
4. 3D map of filtered intensity of cosmic 21cm radiation as a function of angle and redshift with noise model.
5. Inferred redshift space power spectrum of 21cm emission, with errors,
6. Inferred cosmological parameters derived from the power spectrum.

Prepublication access to the data will be at the discretion of the PI in consultation with the rest of the Tianlai collaboration. We would like to make the final maps publicly available, if they are of sufficient quality, but we are not requesting funding for this here.

e. Summary

In summary, we propose to process data from the prototype Tianlai array using the Fermilab computing resources, which are well adapted to this analysis. The LDRD funding supports the R&D that will be required to develop the analysis tools to produce science results from these data. Support is required for hardware (tapes and a tape drive), but the primary need is for the support of a computer professional for the more technical aspects of processing this large data set. It should be noted that we also expect and require vigorous engagement of the Tianlai collaboration to develop the algorithms and software that will be required to overcome the potential problems with the analysis.

The management of this LDRD project is independent of the Tianlai project. Although successful completion of the LDRD project requires analysis of data from Tianlai Fermilab has no contractual obligation to provide data analysis services nor is subject to any MOU's or performance goals imposed by the Tianlai collaboration.

Future Funding

Key components of the deliverables will be made public in scientific journals:

1. A description of algorithms developed to extract 21cm maps from interferometric visibilities as well as metrics for how well they perform. Some of the code will be made public for inspection or use by other experiments.
2. The extracted redshift space power spectrum of 21cm emission in the volume surveyed as well as cosmological parameters including the inferred scale dependent bias of neutral hydrogen.
3. The extracted radio foreground emission from Galactic and Extra-Galactic radio emission.

We would also like to make the final maps publicly available, if they are of sufficient quality, but we are not requesting funding for this here.

The beneficiaries of these deliverables will be 1) Fermilab and the DOE in terms of achieving their mission of advancing fundamental physics, 2) the larger scientific community, especially those pursuing cosmological 21cm observations, and this might include Fermilab and/or the DOE.

We expect to achieve the deliverables within the scope of this LDRD so no future funding is expected to be required for this specific project. However, the project is part of a suite of international efforts to develop 21cm cosmology. Future funding and expansion of the Tianlai array will, in the immediate future, be primarily funded by China, with some contribution from France and the US team members. Fermilab could choose to continue to play a major role in data analysis from a full Tianlai as part of the Cosmic Program. This would be a larger data stream but the methods would have been worked out. The PI expects to continue to devote part of his theoretical efforts toward support of 21cm cosmology. Other projects such as CHIME (the Canadian Hydrogen Intensity Mapping Experiment) may also benefit from deliverables of this LDRD. CHIME has been and likely will continue to be funded from Canadian sources. There are a number of possibilities for an intensity mapping experiment in the southern hemisphere including an expression of interest by the SKA (Square Kilometer Array) but it may be that a dedicated smaller telescope array would be more cost effective. SKA is also funded by non-US sources. Successful completion of the LDRD project will position Fermilab to be a data analysis center for any of these projects although there are no specific plans for that possibility. There is also an opportunity for the US and the DOE to get involved in a southern hemisphere effort, especially if a non-SKA option is pursued.

References

A description of Tianlai telescope and estimates of its performance are given in

[1] Xu, Wang, Chen *Forecasts on the Dark Energy and Primordial Non-Gaussianity Observations with the Tianlai Cylinder Array* Astrophys J **798** 40 (2015)

Comparative estimates of the performance of proposed 21cm surveys are given in

[2] Bull, Ferreira, Patal, Santos *Late-time Cosmology with 21 cm Intensity Mapping Experiments* Astrophys J **803** 21 (2015)

The HI intensity mapping technique was first proposed in

[3] Chang, Peterson, McDonald *Baryon Acoustic Oscillation Intensity Mapping of Dark Energy* Phys Rev Lett **100** 9 (2008)

and was shown to work on the Green Bank Telescope

[4] Chang, Pen, Bandura, Peterson *An intensity map of 21-cm emission at redshift z~0.8* Nature **466** 463 (2010)

Fermilab scientists have shown an interest in 21cm intensity mapping as illustrated by this early design study and by a 12 of 39 author participation in a contribution to Astro2010

[5] Seo, Dodelson, Marriner, McGinnis, Stebbins, Stoughton, Vallinotto *A Ground-Based 21 cm Baryon Acoustic Oscillation Survey* Astrophys J **721** 164 (2010)

[6] Peterson, Aleksan, Ansari, Bandura, Bond, Bunton, Carlson, Chang, DeJongh, Dobbs, Dodelson, Darhmaoui, Gnedin, Halpern, Hogan, Le Goff, Liu, Legrouri, Loeb, Loudiyi, Magneville, Marriner, McGinnis, McWilliams, Moniez, Palanque-Delabruille, Pasquinilli, Pen, Rich, Scarpine, Seo, Sigurdson, Seljak, Stebbins, Steffen, Stoughton, Timbie, Vallinotto, Teche *21 cm Intensity Mapping* Astro2010: The Astronomy and Astrophysics Decadal Survey (2010)

PI and co-I have helped develop methods for removal of foregrounds in 21cm surveys

[7] Cho, Lazarian, Timbie *A Technique for Foreground Subtraction in Redshifted 21 cm Observations* Astrophys J **749** 164 (2012)

[8] Shaw, Sigurdson, Pen, Stebbins, Sitwell *All-Sky Interferometry with Spherical Harmonic Transit Telescopes* Astrophys J **781** 57 (2014)

[9] Shaw, Sigurdson, Sitwell, Stebbins, Pen *Coaxing Cosmic 21cm Fluctuations from the Polarized Sky using m-mode Analysis* Phys Rev D **81** 083514 (2015)

Qualifications

The PI has worked in the field of cosmology as a theorist with emphasis on how to relate data to theory for the past 30 years. He has been involved in several cosmological survey projects such as SDSS, SNAP (not funded), and DES. He developed novel methods to extract the existence of primordial waves using B-mode polarization of the Cosmic Microwave Background. He has been involved with an international group of scientist developing 21cm intensity mapping since 2009. This group evolved into both the Tianlai project and a similar Canadian effort: CHIME. He has worked on theoretical development of some the methods that will be tested by this LDRD. A brief CV is also attached.

Resource Availability and Recent LDRD Funding

(not included in 6 page limit):

a. Discuss scientific or technical obligations of the investigators that may limit the available time for working on the LDRD project (e.g., other funded research, participation in scientific committees, etc.); use units of FTE's to estimate the time

This project involves only one active Fermilab investigator, the PI. The PI currently heads Fermilab's Theoretical Astrophysics Group although it is expected this responsibility will be given to another member of the group in the next year. The PIs past theoretical efforts to develop methods for 21cm cosmology is part of Fermilab's theory program and will not be funded by the LDRD but will contribute to the success of the LDRD. The LDRD proposes to fund 0.05FTE of the PIs time for the purpose of management, oversight, and troubleshooting the project. The other 0.95FTE will continue to be funded by the theory program. Most of the PIs theoretical work is not related to this project.

b. List other LDRD commitments of the investigators; include both current (funded projects) and pending (new proposal) commitments; use units of FTE's to estimate the time

There are no other LDRD commitments of investigators.

c. Summarize accomplishments of funded LDRD projects for the last five years (include project title, investigators, and year of project)

The investigators have no prior LDRD projects.

Budget Narrative

A. We plan to ship the data on LTO-6 tapes, each of which holds 2.4TB of data and can be purchased bar coded for inventory purposes. The read speed is 160MB/sec which is only several times the maximal instantaneous data rate so the tape drives would be fairly heavily used. The tapes will be purchased by the LDRD project and delivered to China, where they will be filled with data and returned to Fermilab. The tapes would be the property of Fermilab, and Fermilab scientists would have unrestricted access to the data. A petabyte of data would require ~400 tapes and the cost is ~$40 per tape so the total is ~$16k for the tapes. The LDRD would also purchase a tape drive (~$2k) for the PI who could look at samples of the data "off-line" in his office to monitor and debug problems with the incoming data.

B. We are not requesting computer professionals to develop scientific data reduction software but we do need help with software for data inventory and management as well as scheduling the data reduction runs.

C. This LDRD will pay for costs associated with data storage and computer usage. With the proper tools in place data processing can also be initiated by the PI and/or collaborators from the US, China, or France.

D. The PI who will manage the project will charge 5% of his time to this LDRD.