Fermilab LDRD Proposal

Project Title:

Principal Investigator:
Lead Division/Sector/Section:
Co-Investigators (w/institutions): (if applicable)

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 (150-200 words): Summarize in 150-200 words the scientific/ technical objectives of the proposal, methods that will be used, and expected deliverables and their expected impact. This description should be understandable to a technically literate lay reader.

It is proposed to reduce analyze the correlator data from the Tianlai 21cm intensity mapping redshift survey producing 3 dimensional maps of the 21cm emission from ~50 Gpc^3 volume at a redshift near unity. The power spectrum of inhomogeneities of neutral hydrogen in this volume will be computed and constraints put on cosmological parameters in the same way as galaxy redshift survey are used. This is a pilot project and the main point of this LDRD effort is to develop and refine techniques which can be used on future larger surveys which will have reduced noise, cover a larger redshift range, and have better angular resolution. Tianlai is a collaboration of Chinese, French, and US scientists working with a Chinese funded dedicated cylinder radio interferometer which is nearing completion in western China. The PI is part of this collaboration.

Significance (~1-2 pages): Describe the scientific/technical problem that the proposal addresses, explain why this problem is significant, and introduce your novel approach for addressing this problem. Include a critical comparison of your proposed approach with the latest published work and explain how your project would advance the state of the art and influence its field of research. Begin with the "big picture" and funnel the reader to the significance of the specific problem addressed in the proposal.

By mapping the distribution of galaxies in the universe one can learn about the early universe where cosmic inhomogeneities were produced as well as about the content of the universe today, including gross properties such as the equation of state of the mysterious dark energy which is causing universal acceleration and detailed features such as the tiny masses of neutrinos. These were/are goals of past, present and future large Fermilab projects: the Sloan Digital Sky Survey, the Dark Energy Survey, the Large Synoptic Survey Telescope and the Dark Energy Spectroscopic Instrument. All of these surveys use optical light to detect galaxies and measure their distance. As we explore larger and more distant volumes of the universe optical techniques to do the mapping become increasingly more expensive. However there are alternatives.

Neutral hydrogen (HI) in the universe is producing copious numbers of radio photons via the hyperfine spin flip transition which produces a narrow line at a wavelength of 21cm. The 21cm line is unique in cosmology in that it is the dominant astronomical line emission over the broad range of frequencies corresponding to cosmological redshifts. So to a good approximation the frequency of a feature can be converted to a Doppler redshift or blueshift without having to first identify the atomic transition. Making a map of the redshift and angle distribution of this line would give us a map of the spatial distribution of HI in the universe. HI is just as good a tracer of the large scale structure (LSS) of the universe as optically bright galaxies which are used in more traditional redshift surveys. Any of these LSS maps can be used to study dark energy by tracking the angular and redshift scale of baryon acoustic oscillations (BAOs). An advantage of the 21cm technique is that it is very easy to determine very accurate redshifts which is the most difficult part for optical redshift surveys. Another reason to pursue 21cm LSS mapmaking is it's future potential. HI 21cm emission and absorption occurs even before galaxies form, i.e. during the "dark ages". In principle this technique can be extended to study the LSS in the majority of the cosmological volume which we can only see during their dark ages.

Three main reasons why 21cm is not currently a prominent redshift survey technique is:
1. it wasn't appreciated that making "intensity maps" with telescopes that cannot resolve individual distant galaxies would be useful,
2. foreground emission in these bands is orders of magnitude larger than the 21m emission and the possibility of removing them was not fully explored,

3. it is only the recent availability of inexpensive fast digital electronics make the hardware costs fairly reasonable: in the $10M range for a significant survey (we are not asking for fund hardware here).

The feasibility of the 21cm intensity mapping and foreground removal has been established "in theory" and practical demonstration of these techniques are now in progress and what is being proposed here. A successful demonstration would open the door to a most promising future for 21cm LSS. The PI has spent significant effort on theoretical approaches to the foreground removal problem and would like to test them.

There are a few ongoing 21 cm intensity mapping pilot projects and the one which the PI is involved with is Tianlai[1]. Tianlai refers to two interferometric arrays which will be dedicated to 21cm intensity mapping: the main one 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 have only a small fraction of the receivers they could accommodate, and if this pilot project is successful they will be outfitted with a full complement of receivers which will greatly increase the sensitivity. 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). The Tianlai collaboration includes two prominent members of the GBT team: Peter Timbie (co-I) and Jeff Peterson (collaborator). One should do much better with the Tianlai. 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 Dark Energy Survey (although with poorer angular resolution and larger map noise than DES).

While what is proposed is data reduction and analysis this is an R&D project because the algorithms used have never really been implemented (in a data reduction pipeline) or tested on real data and inevitably will need to be refined. The main product of the R&D effort will not be technological but rather intellectual know how. The techniques developed will be invaluable for future 21cm surveys such as a extensions of the Tianlai pilot project and other similar projects, such as a natural follow-on survey in the Southern hemisphere. The unique skills gained by any RAs which get involved will position to be leaders in the nascent field of 21cm cosmology. The 21cm LSS 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. If funded a Fermilab Tianlai analysis center would play a very visible role in producing the cosmological science results which would be the most visible contribution to the particle astrophysics community.

---

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

Research Plan (~3-4 pages): Provide a brief overview of your research plan and your specific objectives or aims. For each objective/aim, provide a section with the following:

    a. State the objective/aim

    b. Describe the scientific hypotheses to be tested or technical concepts to be demonstrated to achieve the objective/aim.

    c. Discuss the methods, materials, facilities, protocols to be employed, and techniques for analyzing data and validating results as appropriate.

    d. Describe the expected results and impact (e.g., fundamental breakthroughs, enabling technologies)

    e. Provide a deliverable(s) for year one (within the first year of funding), year two (if seeking a multi-year project), and at the completion of the project.

It is proposed that much of data reduction and analysis for the Tianlai 21cm redshift survey be done at Fermilab producing 3 dimensional maps of the 21cm emission from ~50 Gpc^3 volume at a redshift around unity.   The power spectrum of 21cm inhomogeneities in this volume will be computed and constraints put on cosmological parameters.   The focus will be on learning how to do this chain of operations well: from real radio interferometer data to cosmological parameters.   Going from LSS maps to parameters is already a well developed art so the main effort will be on going from radio data to LSS maps.

Let us review the data chain for the Tianlai cylinder interferometer. The voltages from 92 dual polarization telescope feeds are mixed down from 700-800MHz and digitally recorded. This time stream are FFT'd into 1024 frequency channels. Each channelized time stream is then correlated with the channelized time stream from every other feed. These correlations or "visibilities" are averaged over 1 second intervals and eventually written to tapes or disks which are shipped to the National Astronomical Observatories of China in Beijing. Copies of this data are written to tapes and shipped to Fermilab. The 1Hz time ordered data (TOD) is written to tapes and then shipped to Fermilab where they are "archived". At Fermilab the TOD is searched for RFI (radio frequency interference) which if flagged, self calibrated, and weighted and then averaged over a time interval (to be determined but at least a minute). This refined TOD is broken into 1 sidereal day lengths and co-added into an "average sidereal day's data" (ASD) as well as a noise model for the ASD. The ASD is about 10TByte in size and is a rudimentary image of the raw data. However as the data accumulates the noise contribution will decrease and will eventually be dominated by foregrounds with the 21cm emission being a minor component which must be extracted.

Going from the ASD to a 3D map is the trickiest part of the data chain. One must deconvolve the interferometric beam pattern in such a way as to "line up" synthesized beam patterns in all the different frequency channels. The foregrounds which have a very smooth spectrum in the frequency direction (but not the spatial direction) can then be subtracted by low pass filtering in the channel direction. This lining up / low pass filtering can be done in different ways. One is the the signal-to-foreground Karhunen-Loeve method developed with the PI, there is another method develop by Co-I Ansari and a similar purification technique developed by the PI. We will try them all. The residual after low pass filtering should be a filtered image of the 21cm emission in redshift space, *i.e.* a real 3D map of the neutral hydrogen distribution. This will be accompanied by a noise model. The 3D map will be noisy even after a years data, but the power spectrum should have high signal-to-noise. This power spectrum can be fit to cosmological parameters. The data being shipped to Fermilab is about a petabyte and the final 3D map about 10 terabytes. Final as well as intermediate stage of the data should be made available to the collaboration, including members from China and France. We have no aversion to sharing the final 3D map with the scientific community but as this is a significant expense this is not part of the what is asked to be funded here.

The high passed filtered, deconvolved map which 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 (21 cm emission is unpolarized). Currently there are no comparable data set 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 will tell us about Galactic magnetic fields. Models of the Galactic cosmic ray distribution and magnetic fields can be inferred from these maps may be useful for Galactic science. It is relevant input to such programs as GALPROP used for indirect detection of dark matter. Another byproduct of the survey will be a catalog of radio transients. Although these will be "caught" in Beijing, not Fermilab, they will available to collaboration.

The challenges to be overcome, beyond the shear volume of the data, include development and refinement of the data analysis methods, which will involve scientist time looking at the data at various stages of reduction. Data visualization software and metrics for data quality must be developed. Another big uncertainty is modeling of the telescope beam patterns. This will mostly done by other members of the collaboration (Peter Timbie) but the ASD data must be used to validate these models and this is part of the analysis chain. The telescope itself may change with time in a variety of ways which can change the beam patterns and this must be monitored.

Now let us say something about methods. Tianlai is a transit telescope meaning no moving parts. It stares while the Earth rotates. The astronomical signal we are looking for does not change significantly with time so that each time the Earth rotates (1 sidereal day) the TOD should repeat. The noise from the telescope and the sky does not repeat but should average down as we co-add different days. There are other astronomical signals which do vary with time, but in fairly predictable way, namely the Sun, Moon, planets and pulsars. The radio sky also has unpredictable time variations from compact radio sources, and both pulsars and the Sun fluctuate in brightness. A larger source of unpredictable variability comes from manmade RFI. By searching for and flagging such unpredictable variability one can deweight parts of the data which are contaminated by these fluctuations when making the ASD. Moving bright radio source such as the Sun, Moon, and Jupiter can be subtracted, but uncertainties in the beam patterns will lead to uncertain residuals in the subtractions. So one should deweight data likely to be contaminated by those residuals. The same goes for non-moving bright radio sources such as Cas A, Sag A, Cen A. Regions around these sources must be deweighted when computing a power spectrum because of large residuals. The entire weighting scheme has yet to be developed.

The Tianlai cylinder array is designed with very many redundant baselines, which means the contribution of the sky signal (including the sky noise) to the correlations between redundant pairs of feeds should, ideally, be identical. One can use this to calibrate the and instrumental noise and relative complex gain of each of the feeds. This can also be used to discover malfunctioning components. How best to do this and how well it works is something which will be learned during the course of this project.

The structure of the raw data arriving at Fermilab is very simple: for each pair of the 96 feeds there are complex correlations at each of the 1024 frequency channels. These will be stored in FITS (Flexible Image Transport System) files which is a standard in astronomy and can hold metadata which would identify the time interval over which the data was averaged, identify the feeds and flag information about RFI, radio transients, position of moon and planets, and operational condition of the telescope. There will also be ancillary data files which inventory the FITS files and give a time history of the telescope operations and configuration (which can in principle change). Data I/O and visualization of these FITS files can be done with standard tools. The FITS format will be used throughout the data chain. The final 3D maps will be stored in HEALPix format which is a standard for CMBR analysis. The main difference between our maps and the CMBR maps is that we will have 1024 channels whereas CMBR data has only a few bands.

The main cost that is funded by the LDRD is computer professional help in handling the data as it arrives and running the processing jobs. The TOD data arriving at Fermilab over time is large (~1 petabyte) and reducing from TOD to ASD involves processing/ reducing all of the accumulated data. This is a significant task. The TOD to ASD step

will likely be done a number of times as we perfect our algorithms. We are not requesting computer professionals to develop scientific data reduction software but we may need help with software for data inventory and management and well as scheduling the data reduction runs. This LDRD will also 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. The map making step will also be done by different members of the collaboration and we expect there to be different algorithms developed by different members of the Tianlai collaboration. These members should have free access to the ASD at Fermilab and be able to run their algorithms on the farms here.

In this paragraph we break down the data reduction to it's simplest form. The data shipped to Fermilab is the time ordered correlations TOD, while the near final data products are the 3D maps of 21cm intensity and foreground intensity and polarization. The latter depend linearly on the former so one can think of the entire procedure as a large matrix multiplication. The receivers are assumed linear implying that different frequencies do not mix so each frequency channel is independent and this large matrix is block diagonal with 1024 channel blocks. Since the TOD is periodic with period 1 sidereal day the averaged sidereal days data can be Fourier transformed into m-modes (m is like the azimuthal index of a spherical harmonic) and so long as the total weight is uniform across the day (although not every day gets the same weight) then the different m components do not mix and the channel block are also block diagonal, where the number of m blocks is at most the sampling rate times the sidereal day (~1435 for 1 minute sampling). So we see that the matrix is extremely sparse. A large part of the intellectual effort will be to optimally determine the non-zero elements of the matrix as well as the daily weights. The number of non-zero matrix elements is still extremely large and it is likely they will be computed "on-the-fly" with algorithms that are being developed.

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 consist of $1024 \times (2 \times 96)^2$ =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 by todays standards but this needs to shipped and read in and processed; and could be continuous for months and years which could add up to no petabytes per year. We have not fixed the byte count nor the sample rate and we do not know the uptime, but a good guess is 1 petabyte per year. This LDRD funds two years efforts which is sufficient to validate the methods.

We plan to ship the data on on LTO-6 WORM (write once read many) tapes which holds 2.4TB of data (uncompressed) and can be purchased bar coded for inventory purposes. The (uncompressed) 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 plan is to purchase the tapes with the LDRD and have them delivered to China where they will be

filled with data and sent back the Fermilab. Fermilab would retain possession of the tapes and data. A petabyte of data would require ~400 tapes and the cost is ~$40 per tape so this 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 in case there are issues with the incoming data.

Future Funding (~1/2 page):

    a. Describe how your results will be disseminated; include likely journals in which your work would be published and conferences, workshops, and planning activities at which the work would be presented.

    b. List the probable future funding sources (sponsors), including DOE programs, other agencies, state agencies, or private sector investment.

    c. For each probable funding source:

        i. Explain how the sponsor would benefit from the pursuit of this work.

        ii. Discuss probable contacts with sponsors and plans for responding to current and planned proposal calls.

        iii. Estimate the likelihood that the sponsor would provide future funding including anticipated range of such funding.

References (not included in 6 page limit): Cite a concise set of relevant literature that supports the scientific/technical significance of your project and the innovativeness of your proposed methodologies.

Qualifications (optional, not included in 6 page limit): The C.V. of the PI may be attached and/or a brief statement (~1/2 page) discussing the qualifications of the PI to carry out the proposed research may be described.

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

    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 Table (not included in 6 page limit, separate document): The last page of the proposal consists of a completed budget table. In the budget table, include a cost breakdown for each objective/ aim discussed in the Research Plan. There is no specific budget limitation, but keep in mind that the LDRD funding has limited resources. If subcontracting work, it should be clear in the proposal what work is being subcontracted and justified why that work is not able to performed within Fermilab.