Tianlai Data Analysis Center

About Tianlai Data Analysis Center

In order to analyze the Tianlai data requires computer resources in terms of storage and CPU cycles more than any of the collaborators has at hand. On the other hand these resource requirements are rather small when compared to a number of other experiments, such as particle collider experiments. It makes sense to explore the possibility of using Fermilab's computer resources to do the Tianlai analysis. The project requires a large amount of storage (petabytes) which needs to be reduced to terabytes of science data, e.g. 3D HI maps. The reduction techniques are of order N (proportional to the data size), embarassingly parallel, and could be done efficiently on the types of farms of computers which Fermilab uses to analyse collider events.

It is important to keep in mind that Fermilab wants to be seen as "doing science" as opposed to just providing resources for others to do science (whether or not that is actually what is going on). The name "Tianlai Archive Center" therefore gives the wrong connotation. A name like "Tianlai Analysis Center" is better.

The Fermilab proposal must sit well within Fermilab's current computational capabilities in order to avoid encountering a large number of bureaucratic and other hurdles which would surely delay it's implementation beyond what would be useful for the project. That being said the culture here is "to do the job right" and not simply to make a proposal for the maximum amount of resources which are available. Thus one needs to justify that the resources requested would not only be necessary but also that they are sufficient to obtain the science results that are wanted, and with some contingency as well. The idea is that "producing data is expensive and analysis is cheap" and one should therefore strive make maximum use of the data. Along with this goes the idea that one should safeguard the data product. Therefore I think that part of the proposal should be to have an archive of the lowest level (level 0) data somewhere, and possibly at Fermilab. If it existed in China that might be sufficient but if it didn't exist anywhere that might make the project seem too risky to partake in.

Radio Telescope Basics

Number of Correlations

From cross correlating N_{feed} dual polarization feeds the number of visibilities is

$$N^{\times}_{\text{corr}} = \overbrace{2}^{\substack{\text{real \&}\\ \text{imaginary}\\ \text{parts}}} \times \overbrace{3}^{\substack{\text{polarization}\\ \text{pairs}}} \times \frac{\overbrace{N_{\text{feed}}\,(N_{\text{feed}}-1)}^{\substack{\text{feed}\\ \text{pairs}}}}{2} = 3\,N_{\text{feed}}\,(N_{\text{feed}}-1)$$

From auto-correlating N_{feed} dual polarization feeds the number of visibilities is

$$N^{\text{a}}_{\text{corr}} = \left(\overbrace{2}^{\substack{\text{real \&}\\ \text{imaginary}\\ \text{part}}} \times \overbrace{1}^{\substack{\text{only one}\\ \text{polarization}\\ \text{cross pair}}} + \overbrace{1}^{\substack{\text{only}\\ \text{real}\\ \text{part}}} \times \overbrace{2}^{\substack{\text{two}\\ \text{polarization}\\ \text{auto pairs}}} \right) \times N_{\text{feed}} = 4\,N_{\text{feed}}$$

Adding the number of auto- and cross-correlations we find the total number of correlations is

$$N^{\text{feed}}_{\text{corr}} = N^{\text{a}}_{\text{corr}} + N^{\times}_{\text{corr}} = (3\,N_{\text{feed}} + 1)\,N_{\text{feed}}$$

This is before multiplexing the correlations into different channels.

Suppose we multiplex into N_{ch} channels in which case the total number of correlations we might keep track of is

$N_{\text{corr}}^{\text{freq}} = (3\,N_{\text{feed}} + 1)\,N_{\text{feed}}\,N_{\text{ch}}$.

Maximal Data Rate

We average the multiplexed correlations and then store the averaged correlation functions at a frequency f_{sample}. For a transit telescopes the time between samples should be short enough that the Earth has not rotated significantly in that interval. Let D_{EW} be the maximum east-west dimensions of the interferometric array. The beam patterns on the sky have well sampled azimuthal wavenumbers $m \le m_{\text{max}} = \frac{D_{\text{EW}}}{\lambda_{\text{min}}}$. There will be some spillover to larger m but with larger noise. The minimum R.A. angular scale well-probed (half a wavelength) is thus $\pi \frac{\lambda_{\text{min}}}{D_{\text{EW}}}$ so one would generally want $f_{\text{sample}} \gg \frac{1}{\text{day}_{\text{sidereal}}} \frac{D_{\text{EW}}}{\lambda_{\text{min}}}$.

One might want to sample even
more frequently in order to facilitate downstream (i.e. not real-time) elimination or deweighting of RFI (radio frequency interference). The amount of data that is contaminated by short bursts of RFI is smaller the smaller the sample time.

Initial Tianlai Cylinder Telescope

For the initial Tianlai cylinder telescope $N_{\text{feed}} = 96$ and $N_{\text{ch}} = 1024$ so
$N_{\text{corr}}^{\text{feed}} = 27\,744$
$N_{\text{corr}}^{\text{freq}} = 28\,409\,856$.
For initial Tianlai
$D_{\text{EW}} \approx 30$ m
$\lambda_{\text{min}} > 21$ cm
so
$m_{\text{max}} > 142$
and we require
$f_{\text{sample}} \gg 1.6$ mHz.
For the purpose of sky sampling $f_{\text{sample}} \approx 0.1$ Hz would not be unreasonable. For RFR larger number might be desireable.
If we were to sample this at a rate f_{sample} then the rate of averaged correlations generated is (we use sidereal year $\text{yr}_S = 365.256$ days)

$$R_{\text{corr}} = 8.96562 \times 10^{13} / \text{yr}_S \times \frac{f_{\text{sample}}}{0.1\text{ Hz}}.$$

These numbers will be represented with finite precision so if stored as a $\#_{\text{byte}}$ object then the data rate is

$$R_{\text{data}} = 2.7\text{ MB/sec} \times \frac{f_{\text{sample}}}{0.1\text{ Hz}} \times \#_{\text{byte}} = 79.6\text{ TB/yr}_S \times \frac{f_{\text{sample}}}{0.1\text{ Hz}} \times \#_{\text{byte}}.$$