Tianlai Data Analysis Center @ Fermilab

Albert Stebbins
Xinjiang China

2015 Tianlai 21cm Workshop
September 8, 2015

Tianlai Data: Signal

- Fundamental Observational Quantity Measured: **Stokes Parameters**

  $I_\nu[\alpha,\delta]$, $Q_\nu[\alpha,\delta]$, $U_\nu[\alpha,\delta]$, $V_\nu[\alpha,\delta]$ (or $\mathbf{T}_\nu[\alpha,\delta]$ collectively)

- Discretely sampled: ν into 1024 frequency channels: ν_a
- Discretely sampled into visibilities: (α,δ) = (RA,dec)

$$C_a^{i,j}[\mathrm{LST}] = \iint d\alpha\, d\delta \int d\nu\, w_a[\nu]\, \mathbf{B}_\nu^{i,j}[\alpha-\mathrm{LST},\delta] \cdot \mathbf{T}_\nu[\alpha,\delta]$$

- (α,δ) = (RA,dec) LST - local sidereal time
- Visibilities averaged into time samples of length Δt: $C_{a,\mu}^{i,j} = \langle C_a^{i,j}[t] \rangle_t$
- i,j label the voltage streams - 2×32×3 initially
- *a* labels the channel
- μ labels the time interval
- What is an appropriate value for **sample time**, Δt?
- N.B. in this approximation
- $C_{a,\mu}^{i,j}$ is **linear** in $\mathbf{T}_\nu[\alpha,\delta]$ and
- $C_{a,\mu}^{i,j}$ is **periodic** each sidereal day

Tianlai Data: Noise

- Non-linear / non-periodic **contamination** of $C_{a,\mu}^{i,j}$
- the data stream is digitized which introduces non-linear round-off error
- additive noise: thermal / non-thermal noise in the instrument
- multiplicative noise: variations in gain / phase / other instrumental variation
- ground pickup (which can be time variable)
- moving objects (creatures, clouds, sun, moon, planets, satellites)
- radio transients, e.g. pulsars

- **Time Ordered Data** (**TOD**)
- $C_{a,\mu}^{i,j}$ including both signal and noise

Data Reduction

- TOD produced at the Tianlai site and then shipped (on tape) to be analyzed along with any metadata.
- besides a description of the TOD, what other metadata should be included?

- TOD “scanned” for contamination
- Identified contamination can be either “corrected” for “flagged”, e.g.
- TOD could be corrected for multiplicative noise (e.g. phase variations)
- characteristics of RFI can be flagged and used to de-weight that part of the data
- noise level of each feed could be noted and used to construct a noise model

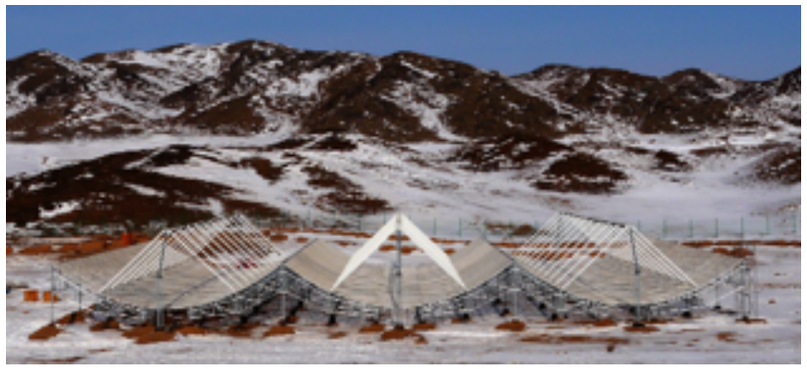
- **Average Sidereal Day** (**ASD**):
- Split TOD into sidereal days and average days
- used RFI flags and and noise notes to determine optimal weights for average.
- Non-periodic signal (residual from ASD) can be used to define contamination model, .e.g. a model of the planetary contamination.
- correct ASD for contamination model

Map Making / Parameter Estimation

- Given beam patterns, $\mathbf{B}_\nu^{j,j}[\alpha\text{-LST},\delta]$, deconvolve ASD to reconstruct $\mathbb{T}_\nu[\alpha,\delta]$, an estimate of $\mathbf{T}_\nu[\alpha,\delta]$.
  - How do we know the beam patterns, $\mathbf{B}_\nu^{j,j}[\alpha\text{-LST},\delta]$?
  - the discrete data stream only samples a finite subspace of the infinite dimensional Hilbert space which describes $\mathbf{T}_\nu[\alpha,\delta]$ ∴ no unique reconstruction algorithm even in the absence of noise.
  - If we do want to make a map $\mathbb{T}_\nu[\alpha,\delta]$, then which algorithm do we choose?
  - Decompose map into 21cm and foreground contribution

$$\mathbb{T}_\nu[\alpha,\delta] = \mathbb{T}_\nu[\alpha,\delta]_{21cm} + \mathbb{T}_\nu[\alpha,\delta]_{foreground} \pm \text{noise} \pm \text{systematics}$$

  - How do we do this decomposition?
- From $\mathbb{T}_\nu[\alpha,\delta]_{21cm}$ determine P[k,z] and cosmological parameters.



1.6 petabyte / year

TOD

$(3\ N_{feed}+1)\ N_{feed}\ N_{ch}\ \#_{bytes}\ /\ \Delta t$

$N_{feed}=96\ N_{ch}=1024\ \#_{bytes}=2\quad \Delta t=1sec$

This requires significant resources and some planning.

4 terabyte

ASD

$\div 365$

1 terabyte

Maps

$\#_{stokes}\ N_{ch}\ \#_{bytes}\ \Omega_{survey}\ /\ \delta\theta^2$

$\delta\theta=1'\ \Omega_{survey}=25{,}000 deg^2\ N_{ch}=1024\ \#_{bytes}=4=\#_{stokes}$

LDRD Funding Pending

- **L**aboratory **D**irected **R**esearch and **D**evelopment
- internal source of funding not requiring DOE review or appropriation
- selection of funded proposals will be announced this weekQ
- proposal document available upon request
- PI: Stebbins co-I's: Marriner, Chen, Ansari, Timbie

- $265k requested over two years (incl. overhead)
- majority of cost is for data storage, tapes, and shipping
- computing is "free" on Fermigrid (part of Open Science Grid DOE/NSF)
- ***does not pay for***
- analysis software development
- labor to run jobs (can be done remotely)
- extracting science from reduced data

Why Only 2 Years?

- LDRD funds only “research and development” (see ldrd.fnal.gov)
- LDRD does not fund science projects
- Tianlai pathfinder is an R&D project, and upgrades are not.
- What we learn from pathfinder will determine how we want to process pathfinder+ making it difficult to write a specific proposal for pathfinder+.
- majority of cost is for data storage, tapes, and shipping

Get Organized!

- Need to determine soon:
- TOD data format, metadata, and sample rate.
- media for transport of data to Fermilab (grant could pay for media).
- Fermilab would transcribe media onto it's own system which will be stored.
- media can be recycled or not (cost savings)
- Need software ASAP!
- self calibration, RFI identification and mitigation, TOD → ASD
- One can develop "at home" and then test at Fermilab.
- Unlicensed software easily shared between nodes using Open Science Grid (Linux based).
- Data processing should be planned with the collaboration
- I think we should not discourage parallel data processing algorithms within limits.

Who has access?

- Anyone the project management designates!
- Technically the PI (me) needs to authorized each individual user.
- Access is given to people with electronic credentials.
- *Do not "share" access credentials!*
- Data redistribution
  - the TOD will accumulate to a large amount of data so it might be difficult to export (say to France or back to NAOC) but with some planning resources it might be doable.
  - the smaller ASD dataset is might routinely be shared between the Tianlai institutions - although access would be need be approved by project management.
  - the maps would also be shared and might (if of sufficient quality) be made public. The requires resources and is not part of the LDRD funding.

What If Not Funded?

- Identification of alternate data analysis/reduction sites within collaboration should be explored
- This might be useful in any case especially looking forward to pathfinder+ and full Tianlai. Fermilab R&D funding does not guarantee participation in these larger projects (although it probably helps).
- While petabyte data is too big to do at Fermilab without funding - smaller data could be done more informally.
- decreasing the sampling rate to 0.1Hz instead of 1Hz might be sufficient - we would still learn quite a bit about intensity interferometry.
- ~100TByte storage is probably the minimum the maps would also be shared and might (if of sufficient quality) be made public. The requires resources and is not part of the LDRD funding.
- Need software ASAP!


I WANT YOU