The FAST Extragalactic HI Survey
Data Reduction Pipeline
– Current Progress and Status

Bo Zhang¹, Jian Xiao²

¹FAST Science Group, NAOC
²Division of Intelligence & Computing, Tianjin University

November 22, 2018
Contents

- Current status of FAST extragalactic HI observations
- Basic design of FAST extragalactic HI survey data reduction pipeline
- Convolutional neural network-based automatic RFI flagging code
- CPU-GPU hybrid convolution-based gridding algorithm
- Summary

Supported by the Joint Research Fund in Astronomy (The Key Technologies for High-Performance Data Pipeline of FAST, U1731125, 2018.01-2020.12)
Contents

- Current status of FAST extragalactic HI observations
- Basic design of FAST extragalactic HI survey data reduction pipeline
- Convolutional neural network-based automatic RFI flagging code
- CPU-GPU hybrid convolution-based gridding algorithm
- Summary
Current Status of FAST Telescope

- Commissioning of the telescope in progress
- Observing modes: drift scan; tracking; on-the-fly; basket-weaving
- Backends: spectral line (GPU-based) + pulsar (ROACH) + off-the-shelf spectrum analyzer (+ SETI/FRB backend + CRANE)
- 19-beam receiver with band coverage of 1.05-1.45 GHz saw first light on May 25th, 2018
- L-band sensitivity $A/T > 2000\, m^2/K$ achieved
FAST Extragalactic HI Observations

- Extragalactic emission/absorption lines observed
- Test calibration: consistent with Arecibo's results
- System temperature/gain measurements still in progress; improvements expected in near future

- ALFALFA: $W_{50} = 112 \pm 7$ km/s, flux = $0.89 \pm 0.09$ Jy km/s
- FAST: $W_{50} \sim 100$ km/s, flux $\sim 1$ Jy km/s
FAST HI Observations: M31 Drift Scan

September 12th M31 drift scan (Beam 1)
Contents

- Current status of FAST extragalactic HI observations
- Basic design of FAST extragalactic HI survey data reduction pipeline
- Convolutional neural network-based automatic RFI flagging code
- CPU-GPU hybrid convolution-based gridding algorithm
- Summary
Basic Steps of HI Data Reduction

- **Level 0**: data format conversion
- **Level I**: flux calibration, bandpass/baseline calibration, RFI flagging
- **Level II**: data cube preparation, signal extraction
- **Level III**: cataloging, writing research papers, data archiving & VO interfacing ...
While for Existing Single-Dish Data Reduction Codes...

- **ALFALFA's LOVEDATA code (IDL):**
  - Designed for ALFALFA survey only; not a universal tool
  - Manual intervention required for key steps

- **Gildas/CLASS (fortran90):**
  - General-purpose software widely used by single-dish telescopes
  - Manual intervention required

- **AIPS++, CASA (C++, Python):**
  - Designed as universal radio astronomy data reduction software; widely used by interferometers
  - Partial automatic data reduction made possible with reconfigurations of xml files; MPI supported
  - Source code too complicated to modify

- **Livedata/Gridzilla(C++) & ASAP (AIPS-based):**
  - Functions & performance not in satisfactory
Our Main Challenge: Data Volume

- Data from FAST spectral line backend: SDFITS, 2 GB each
  - Wide+Narrow band: 1000-1500 MHz with 65536 channels for wide band; 31.25 MHz bandwidth with 65536 channels for narrow band
  - Full resolution: 1000-1500 MHz, 1024K channels

- Data rate:
  - Wide+Narrow band: 2 (N+W band) × 19 beams × 4 Stokes × 65536 channels floating point data/s + header; ~ 2.2 GB/min
  - Full resolution: 19 beams × 4 Stokes × 1048576 channels floating point data/s + header; ~ 17.6 GB/min

- Survey Duration: ~ 220 days for single pass, 440 days for two-pass scan → PB-sized data generated; > 10 billion voxels in extragalactic HI data cube!

- Our goal: automatic data reduction with high performance computing & minimum manual interventions
Progress on Levels 0 & I

- Level 0 in progress: SDFITS → HDF5 for massive parallel computing
- Level I calibrations: various tests with observational data in progress; more investigations on instrumental properties needed to get a better result

Before

After
Contents

- Current status of FAST extragalactic HI observations
- Basic design of FAST extragalactic HI survey data reduction pipeline
- Convolutional neural network-based automatic RFI flagging code
- CPU-GPU hybrid convolution-based gridding algorithm
- Summary
What are RFIs?

- **Radio Frequency Interference**
- **Origin of major RFIs in 1.0 – 1.5 GHz band:** navigation/broadcasting/telecommunication satellites & aviation distance measuring equipment; unavoidable!

**Radio interference**
- Electromagnetic interference, also called radio frequency interference (RFI) when in the radio frequency spectrum, is a disturbance generated by an external source that affects an electrical circuit by electromagnetic induction, electrostatic coupling, or conduction.
RFI Flagging – Traditional Ways

- Manual inspections
- Surface fitting and smoothing:
  - Assuming astronomical signal as smooth surface; spikes removed
  - Not suitable for spectral line observations
- Combinational threshold:
  - Detects series of samples with higher values than pre-defined thresholds
  - Widely used with moderate computing requirements
  - Thresholds needed to be chosen carefully
- Singular value decomposition
  - Performing singular value decomposition of original data matrix; highest singular values flagged as RFIs
  - Suitable for detections of periodic RFIs only
- Major problems: accuracy, robustness, speed...
Machine Learning & RFI Flagging

- Application of traditional machine learning algorithms:
  - Characterizing Interference in Radio Astronomy Observations through Active and Unsupervised Learning (Doran 2013)
  - Machine learning approach to radio frequency interference (RFI) classification in radio astronomy (Wolfaardt 2016)

- Application of deep learning:
  - Radio frequency interference mitigation using deep convolutional neural networks (Akeret 2016)

24 hours of observed TOD from the Bleien Observatory

Original data  SumThreshold  CNN

Akeret (2016)
So...How to Utilize Machine Learning to Deal with RFIs?

- Maintain connections between RFI characteristics: U-Net – a modified convolutional neural networks (CNN)
- Generate training samples; avoiding random interferences: GAN - generative adversarial network
- Performance optimization
Our Modified U-Net Model

- Complex RFIs: more layers
- Narrow Band RFIs: small convolution kernels
- Wide band RFIs: multi-layer convolution
- Concentrated RFIs: Max-pooling
- Random RFIs: Batch-Normalization
Zhicheng Yang, Jian Xiao, Bo Zhang et al. An improved U-Net for Automatic Recognition of Radio Frequency Interference, ADASS 2018
Preliminary Results with Simulated Data

- Precision: the proportion of genuine RFIs against samples identified as RFIs
- Recall: the proportion of RFIs correctly identified among all RFI samples
- F1 Score: average of precision & recall's reciprocals

- More training with observational data required
- Further optimization needed
Contents

- Current status of FAST extragalactic HI observations
- Basic design of FAST extragalactic HI survey data reduction pipeline
- Convolutional neural network-based automatic RFI flagging code
- CPU-GPU hybrid convolution-based gridding algorithm
- Summary
Why Gridding?

- Generate uniformly spaced data cubes from resampling of more randomly distributed observations
- Basic equation:

$$R_{i,j}(\alpha_i, \delta_i) = \frac{1}{W_{i,j}} \sum_n R_n(\alpha_n, \delta_n) w(\alpha_i, \delta_i; \alpha_n, \delta_n)$$

Fabello et al. (2011)
Two Processes: Scatter vs Gather

**Scatter**
- No search, but has race condition
- Sample point (input)
- Grid point (output)

**Gather**
- No writing race, but needs search
- Grid point (input)
- Sample point (output)
**Ring-Based Searching Strategy**

- Quickly find all the potential points contributing to a certain grid point
- HEALPix hierarchically subdivides the sphere surface into $12N^2$ pixels (cell)
- All cell centers are located at $4N-1$ of constant latitudes (rings)

Qi Luo, Jian Xiao, Ce Yu et al. HyGrid: A CPU-GPU Hybrid Convolution-based Gridding Algorithm in Radio Astronomy, ICA3PP, 2018
1. groups and sorts input points by HEALPix ID (cells in different colors) on CPU  
2. Load the compact array into GPU’s memory

searches the neighboring input points
HyGrid Results with Simulated Data

Qi Luo, Jian Xiao, Ce Yu et al. HyGrid: A CPU-GPU Hybrid Convolution-based Gridding Algorithm in Radio Astronomy, ICA3PP, 2018
HyGrid Performance Tests

Acceleration with different input size
HyGrid vs CyGrid

CPU ordering vs GPU gather
Contents

- Current status of FAST extragalactic HI observations
- Basic design of FAST extragalactic HI survey data reduction pipeline
- Convolutional neural network-based automatic RFI flagging code
- CPU-GPU hybrid convolution-based gridding algorithm
- Summary
Summary

- Various types of extragalactic HI signals already detected with FAST in its commissioning phase
- Data conversion & calibration codes for FAST extragalactic HI sky survey under development
- Improved U-Net-based automatic RFI flagging algorithm proved effective; training with observational data needed
- CPU-GPU hybrid convolution-based gridding algorithm with better performance developed
- More test with real observational data required
Thank You!