Development of an Optimized Real-Time Radio Transient Imager for LWA-SV

Hariharan Krishnan*¶, James Kent†, Jayce Dowell‡, Adam P Bearsdley*, Judd D. Bowman*, Greg B Taylor‡, Nithyanandhan Thyagarajan§‖ and Daniel Jacobs*
*School of Earth & Space Exploration, Arizona State University, Tempe, AZ, USA
†Cavendish Laboratory, University of Cambridge, Cambridge, UK
‡Dept. of Physics & Astronomy, University of New Mexico, Albuquerque, NM, USA
§National Radio Astronomy Observatory, Socorro, NM, USA

***Abstract*—In this paper, we describe our efforts towards the development of a real-time radio imaging correlator for the Long-Wavelength Array station in Sevilleta, New Mexico. We briefly discuss the direct-imaging algorithm and present the architecture of the GPU implementation. We describe the code-level modifications carried out for one of the modules in the algorithm that improves GPU-memory management and highlight the performance improvements achieved through it. We emphasize our ongoing efforts in tuning the overall run-time duration of the correlator which in turn is expected to increase the operating bandwidth in order to address the demands of wide-band capability for radio transient science.**

I. Introduction

Modern radio astronomy is driven towards an upward trend in the size and computational capacity of the conventional FX correlators, primarily by the requirements of sensitivity, field-of-view and resolution. Next-generation radio telescopes (viz. HERA, SKA, PUMA etc.) will require large-N correlators with significantly higher computational and storage requirements. Digital signal processing back-ends are also expected to be capable of handling reasonably large bandwidths on the order of a few hundred MHz. One of the key scientific objectives of modern radio telescopes is to understand the physics of radio transient phenomena and to explore the space weather environment on other sun-like stars that may harbor exoplanets. In the above context, real-time imaging across a wide frequency band at very high temporal resolution is ideal for the detection and localization of Fast Radio Bursts (FRBs), Meteor Radio Afterglows (MRAs), Giant Pulse (GP) emission from pulsars and flares/CMEs from active/solar-type stars.

Modern Graphical Processing Units (GPUs), offering theoretical computational capacities on the order of few Tflop/s, are ideally suited for the growing high-performance requirements of radio astronomy. The current-generation radio telescopes have been successful in deploying the compute intensive task of cross-correlation and integration of multiple data streams from individual antennas of an aperture array on GPUs. Through visibility mapping from cross-correlation of antenna voltage pairs, conventional correlators spatially sample the sky thereby synthesizing the aperture [1]. However, for real-time sky imaging the resulting visibility products must be subsequently gridded and Fourier transformed which compounds the computational intensity.

¶hari.krish@asu.edu
‖Nithyanandan Thyagarajan is a Jansky fellow of the National Radio Astronomy Observatory

II. E-field Parallel Imaging Correlator (EPIC)

In view of the above, we present a generic correlator implementation to reduce the computational requirements for real-time imaging in large-N (N $>$ 1000) arrays with our E-field Parallel Imaging Correlator (EPIC : [2]). The EPIC is based on the Modular Optimal Frequency-Fourier (MOFF : [3]) mathematical formalism for direct Fourier imaging. The fundamental working principle of EPIC is to grid the electric fields from individual antennas and spatially Fourier transform it to the sky image, thus synthesizing the aperture on-the-fly (Fig. 1). In the process, EPIC introduces a significant reduction in computational scaling from $\mathcal{O}(n_a^2)$ to $\mathcal{O}(n_g log_2 n_g)$ (where n_a is the number of antennas and n_g is the number of grid points) [4] in comparison to the traditional correlators.

EPIC is now implemented on a GPU-accelerated architecture and integrated with a python/C++ based high-performance streaming framework, Bifrost [5]. It was successfully deployed and tested on the Long Wavelength Array station [6] located at the Sevilleta National Wildlife Refuge (LWA-SV) in New Mexico, USA [7]. The LWA-SV is a compact array consisting of 256 dual-polarization active dipole antennas arranged in a pseudo-random pattern with an elliptical footprint operating in the frequency range of 10-88 MHz. The real-time digital signal flow into EPIC is described in the block diagram representation in Fig. 1. The GPU implementation of EPIC is composed of multiple kernels/modules that perform various functions, viz. voltage gridding, spatial Fourier transform, cross-correlation in the image plane, making use of some of the standard signal processing blocks of Bifrost [5] as highlighted in Fig. 1.

III. Code Optimization

The instantaneous bandwidth achieved with the initial deployment of EPIC on LWA-SV is limited to $\approx$ 400 kHz per GPU. We seek to optimize the GPU-code of the correlator for better performance in order to increase the bandwidth processable per device in real-time. We decided to begin with the optimization of the *romein* voltage gridding module,

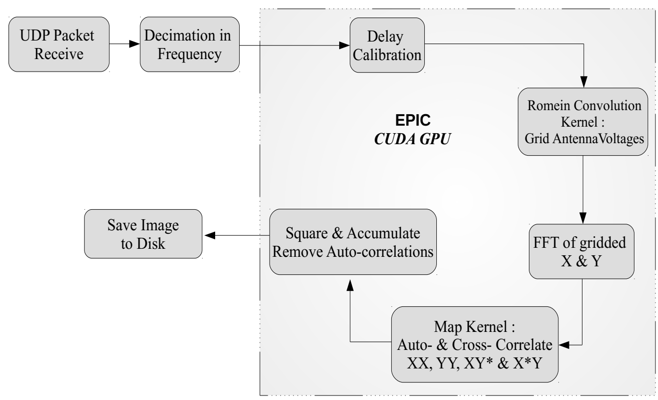

Fig. 1. GPU-implementation of the EPIC architecture in Bifrost for LWA-SV

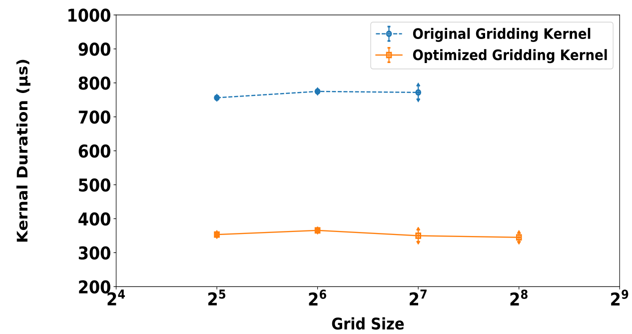

Fig. 2. Comparison of the kernel run-time duration vs Grid-size Dimension: original (blue) and modified (orange) Romein Gridding kernel

which is one of the critical blocks of the EPIC, that is based on a GPU-accelerated convolution algorithm [8] written in C++/CUDA. In the *romein* module, the delay corrected frequency domain signals are convolved with an antenna illumination pattern described as a user-defined convolution kernel and then gridded with a spacing of $< \lambda/2$ on to a 2-D grid [7].

It was observed that the configuration of CUDA threads in the *romein* kernel implementation, and subsequently the memory access pattern, required modifications to improve the kernel performance. Optimizing memory accesses has a huge effect on GPU code efficiency, in the original implementation only one CUDA thread addressed all antennas of the array and handled multiple load/store operations directly from/to the global memory of the GPU. This was modified by re-configuring the CUDA thread-block description to include all antennas as part of one of the thread-block dimension. With this modification one CUDA thread is now able to grid one antenna and the shared memory (which is relatively faster) with increased memory coalescence is now used for load/store operations by the threads. The improved performance can be clearly noticed in Fig. 2 which shows a comparison of the run-time duration for the modified and original versions of the *romein* gridding kernel for different grid-size dimensions. Also, with better memory management the modified gridding module is now able to grid twice the grid-size dimension than earlier.

Discussion

Code-level modifications and efficient GPU-memory management can significantly improve kernel performances in terms of run-time duration. We are currently working to optimize the other modules of the pipeline to achieve similar improvements. This is expected to reduce the overall run-time duration of the correlator that constrains the bandwidth that the correlator can process in real-time. We aim to achieve a reasonable bandwidth of $\approx$ 5-10 MHz per GPU device in order to deploy a commensal transient imager system for the LWA-SV. We also plan to integrate additional modules to the EPIC architecture for the purpose of RFI-mitigation and transient detection to bolster the science capabilities of the commensal system.

Acknowledgment

This work is supported by National Science Foundation awards AST-1710719 and AST-1711164. Construction of the LWA has been supported by the Office of Naval Research under Contract N00014-07-C-0147 and by the AFOSR. Support for operations and continuing development of the LWA1 is provided by the Air Force Research Laboratory (AFRL) and the National Science Foundation under grants AST-1835400 and AGS-1708855. APB is supported by a National Science Foundation Astronomy and Astrophysics Postdoctoral Fellowship under award AST-1701440.

References

[1] Thompson A. R., Moran J. M., Swenson G. W., ”Interferom-etry and Synthesis in Radio Astronomy”, Astronomy and Astrophysics Library, *Cham: Springer International Publishing*, 2017.

[2] Nithyanandan Thyagarajan, Adam P. Beardsley, Judd D. Bowman, and Miguel F. Morales, ”A generic and efficient E-field Parallel Imaging Correlator for next-generation radio telescopes”, *Monthly Notices of the Royal Astronomical Society*, Vol. 467, Issue 1, pp. 715–730, January 2017.

[3] Miguel F. Morales, ”Enabling Next-Generation Dark Energy and Epoch of Reionization Radio Observatories with the MOFF Correlator”, *Publications of the Astronomical Society of the Pacific*, Vol. 123, pp. 1265–1272, November 2011.

[4] Romney J. D., eds: Perley R. A., Schwab F. R. & Bridle S. H., ”Synthesis Imaging: Course notes from an NRAO Summer School”, *National Radio Astronomy Observatory*, New Mexico, 1985.

[5] Miles D. Cranmer, Benjamin R. Barsdell, Danny C. Price, Jayce Dowell, et al., ”Bifrost: A Python/C++ Framework for High-Throughput Stream Processing in Astronomy”, *J. of Astron. Instrum.*, Vol. 06 (4), pp. 1750007, 2017.

[6] G. B. Taylor G. B., Ellingson S. W., Kassim N. E., Craig J., Dowell J., et al., ”First Light For The First Station Of The Long Wavelength Array”, *J. of Astron. Instrum.*, Vol. 01(1), pp. 1250004, 2012.

[7] James Kent, Jayce Dowell, Adam Beardsley, Nithyanandan Thyagarajan, Greg Taylor, and Judd Bowman, ”A real-time, all-sky, high time resolution, direct imager for the long wavelength array”, *Monthly Notices of the Royal Astronomical Society*, Vol. 486, Issue 4, pp. 5052–5060, September 2019.

[8] John W. Romein, ”An Efcient Work-Distribution Strategy for Gridding Radio-Telescope Data on GPUs”, eds.: Bannerjee U. & Gallivan K., *Proceedings of the26th ACM international conference on Supercomputing. ACM*, ACS, New York, pp. 321, 2012.