

Annual Report for Blue Waters Allocation

- **Project Information**

- Title: Image Uncertainty Quantification and Radio Astronomical Imaging.
- PI: Athol J. Kemball, University of Illinois at Urbana-Champaign.
- Names and affiliations of students or collaborators: Michael Katolik (Graduate Student; UIUC); Di Wen (Graduate Student; UIUC).
- Contact: akemball@illinois.edu

- **Executive summary (150 words)**

This research project has as its primary goal the exploration and development of new algorithms and data analysis techniques required to enable science in the challenging transition underway at present to the so-called Great Survey Era in observational astronomy. This era is characterized by highly data- and compute-intensive technological trends and associated analysis challenges that collectively require extreme-scale computational solutions. In the current period of reporting, work has continued on several carefully selected problems in this area including: i) new approaches to advanced calibration and imaging in contemporary radio interferometry; ii) pixel-level image fidelity assessment for ill-posed inverse image formation; and, iii) new approaches to power spectra estimation in large optical surveys and associated supporting simulations. Several graduate students advised by the PI are currently involved in these projects.

- **Description of research activities and results**

- **Key Challenges:** Observational astronomy is undergoing a rapid and disruptive transformation over the past decade into a sharply increasingly compute- and data-intensive discipline. This technical transformation would not be possible without geometric advances in foundational electrical and computer engineering technologies, specifically processor chip and storage density, hardware speed, and software tools and frameworks that permit their efficient use in advanced high-performance computing (HPC) systems. These technical advances allow new grand challenge problems in astrophysics to be approached for the first time. Synergistically, in observational astronomy, the technological trends benefiting computational engineering are also vastly improving the sensor density (and associated data acquisition rates) of astronomical instrumentation used in current and future telescopes. These two trends converge strongly in the current, so-called, Great Survey Era. Telescopes under construction at present, including the Large Synoptic Survey Telescope (LSST) and the Square Kilometer Array (SKA) are being constructed to explore the Universe over very large volumes of solid angle, redshift, wavelength, and time, in order to measure the signature effects of key unknown constituent influences in the evolution of the Universe: dark energy, and dark matter. The Universe, sampled by these instruments over vast volumes, can detect the influence of these unknown matter-energy constituents acting over cosmological times, thus yielding critical insights into their possible physical

origins. The nature of dark energy is arguably the greatest unsolved problem in Physics today. The Great Survey Era demands high-sensitivity wide-field imaging and opens other windows of scientific exploration in addition to the foundational cosmological questions concerning dark energy and dark matter.

This BWP research project has as its primary goal the investigation of targeted grand challenge problems that need to be solved in the Great Survey Era, specifically by employing new large-scale computational approaches that were not previously feasible. In the period of this report, these sub-problems have included:

- a) Novel algorithms for next-generation radio-interferometric instrumental calibration and imaging.
 - b) Uncertainty quantification for ill-posed inverse imaging in astronomical interferometry.
 - c) Statistical estimators of dark matter properties in large N-body simulations as preparation for future large optical surveys.
- **Why it Matters:** The primary value of this work is that Blue Waters allows new approaches to be explored to key algorithms and grand challenge problems early in the Great Survey Era. These facilities and instruments represent significant capital investments in the federal basic science portfolio. This research is relevant to a current facility, the Atacama Large Millimeter Array (ALMA), which was constructed for \$1.4B, and will be important also for the future Square Kilometer Array (SKA) (several billion dollars), and the Large Synoptic Survey Telescope (LSST) (\$1B). Enabling science from these telescopes, including LSST, is critical to ensure that the full return on these capital investments is realized by the community. This science, for the reasons described above, is deeply computationally and data-intensive. Traditional algorithmic approaches need revision for efficiency and the broadest possible community access.
- **Why Blue Waters:** The algorithms evaluated and developed here are not feasible on other HPC systems. They explicitly use new approaches that are substantially more computationally expensive than traditional solutions adopted thus far in the field, and are being developed in this research project specifically with future data rates and computational demands in mind. For example, radio interferometers produce data at a rate broadly proportional to N_{ant}^2 , where N_{ant} is the number of antennas or stations. Current interferometers typically have between 25 and 50 antennas; the SKA will deploy thousands in order to achieve the required collecting area and sensitivity with a commensurate increase in data rate. In addition, we are exploring computational statistical approaches in several cases, providing an additional dimensional multiplier, here over ensemble or realization. In

observational optical astronomy data rates are growing in proportion to the number of pixels in CCD cameras and their associated electronic data acquisition readout speeds.

In terms of specific code performance on Blue Waters, the current research project spans a very heterogeneous code base given the nature of the problem domain of astronomical image formation. This domain utilizes several $O(10^6)$ SLOC) community codes that are typically developed over decades. This software engineering pattern arises because of the complexity of the data schema, domain-specific data management requirements, and custom imaging and calibration heuristics and techniques. At the lowest layers of the software stack these codes do re-use optimized standard numerical libraries, for example for Levenberg-Marquardt optimization, large-scale linear algebra, and other common elements of standard numerical analysis. However, the community codes, considered as a whole, have a net computational profile that is not highly optimized for extreme-scale efficiency. They suffer from undue serialization, inefficient (small-scale) I/O, over-reliance on shared-library dynamic linking, client-server architectures, and poor load-balancing. In the worst cases, even thread safety cannot be assumed. To mitigate, we use a component abstraction behind a domain-specific generic interface to allow re-use of multiple community code elements, built within a common build system, and one in which we can reduce these inefficiencies on a targeted basis. As an aside, we note that this is an important issue in modern HPC computing in general given the increasing heterogeneity of applications from large communities drawn into data- and compute-intensive science without extensive prior experience. Once an algorithm is proved in concept, we move to an optimized implementation using standard parallel programming patterns and technologies, such as MPI. Based on our development in the current research project, we can efficiently use Blue Waters for the full scope of our research applications.

- **Accomplishments:** Our summary accomplishments on Blue Waters in the current reporting period include:
 - **Novel interferometric calibration algorithms:** In the current reporting period this work has continued to focus on novel metaheuristic and stochastic approaches to large-dimension parameter estimation methods to characterize the instrumental and propagation terms in the integral equation that forms the data model in astronomical interferometry. Specifically, this work includes a unique approach to regularization which is needed for this ill-posed inverse problem. In previous reports, we have demonstrated strong results that show the feasibility and convergence of this novel method when using large swarm-particle samples; an approach possible on the scale of Blue Waters. For clarity, we include Figure 1 and Figure 2 from our previous report. The code for these tests was stabilized and we entered production runs early in this reporting period, using a

substantial fraction of our allocation within the first several months. At that point we paused our use of Blue Waters on this project, saving resources for planned runs later in the year, as we moved from simulation studies to real data from contemporary interferometers. Work on a research paper, theoretical development of certain relations concerning our algorithm, and a high service level for the PI has led to a lower than expected utilization beyond that date. We did not unfortunately reach the point where it was appropriate to complete the final runs, although that point is now imminent. This work is in collaboration with a Ph.D. student (M. Katolik) advised by the PI.

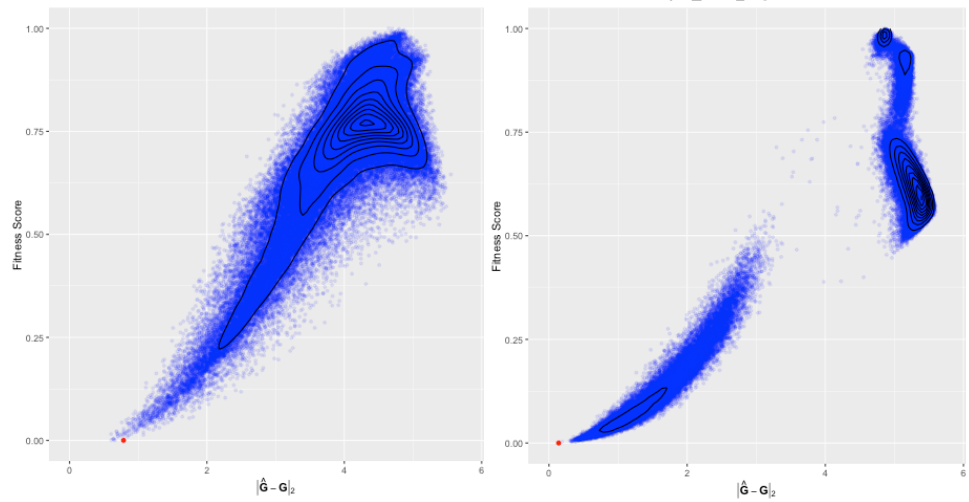


Figure 1. Distribution of particles in a metaheuristic optimization study concerning a novel interferometric calibration method. The x-axis is the metric distance from the true solution, so truth is at zero metric distance. The y-axis shows the fitness objective function value for each particle. Contours are drawn as Gaussian kernel density estimates. The figure at left shows the initial random distribution of 100,000 particles and the figure at right the distribution after ten iterations of the metaheuristic algorithm update. The red dot in each figure depicts the lowest objective fitness function value at a given iteration.

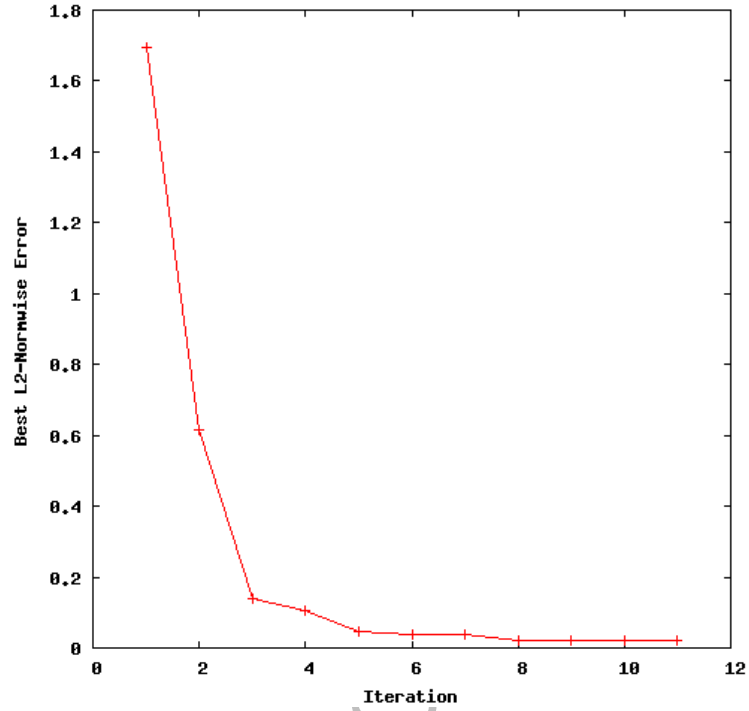


Figure 2. L2 norm of best swarm particle as a function of iteration, for the simulation study shown in Figure 1.

- **Pixel-level imaging fidelity assessment:** During this period of reporting, we have continued work on simulations of large-scale image fidelity as a problem in uncertainty quantification using the integral equation that forms the data model in interferometric image formation. This is an unsolved problem. In this period we have generated new simulation datasets, spanning 10-250 antennas, with a range of specific error distributions of an instrumental and propagation nature. This is a challenging theoretical problem, when considered with classical techniques such as Fisher Information theory to determine estimator performance analytically. The computational work, uniquely possible on Blue Waters, is invaluable in challenging several common (but possibly flawed) assumptions regarding noise covariance and statistical distribution in this domain.
- **Dark-matter substructure:** In last year's report we described Blue Waters team work analyzing ALMA data to infer dark matter substructure in a lensing galaxy. In collaboration with a student (D. Wen) this reporting period has included new work on using counts-in-cells techniques to estimate cosmological power spectra. These techniques are especially valuable in understanding dark matter clustering and distribution, and powerful tools needed for the LSST era. The student has developed an

initial MPI implementation and is testing this method on large simulation data cubes.

- **List of publications and presentations associated with this work**

- a) Yashar D. Hezaveh, Neal Dalal, Daniel P. Marrone, Yao-Yuan Mao, Warren Morningstar, Di Wen, Roger D. Blandford, John E. Carlstrom, Christopher D. Fassnacht, Gilbert P. Holder, Athol Kemball, Philip J. Marshall, Norman Murray, Laurence Perreault Levasseur, Joaquin D. Vieira, Risa H. Wechsler **2016**, *Detection of lensing substructure using ALMA observations of the dusty galaxy SDP.81*, *Astrophysical Journal*, 823, 37H.
- b) Wen, Di, 2017, presentation: “Spatial Distribution of Dark Matter Substructure”, conference: “Cosmology, Gravitational Waves and Particles”, NTU, Singapore, 6-10 February 2017.

- **Plan for next year**

Allocation utilization over the full current reporting period (~50%) failed to reach the target of full utilization as in the previous year. The reasons concerning programmatic research factors are described above, but utilization in the second half of this reporting period was also adversely affected by urgent service duties of the PI within the LSST project. The PI has taken measures to mitigate this risk in the coming year.

For the next year, I would like to sincerely request 200,000 XE NH and 40,000 XK NH in order to continue this important research program. This estimate is based on the known computational complexity of the current stable research code, and the scope of the parameter studies planned. In addition, the PI will advise an additional Ph.D. student (to reach a total of five) starting in 2017. This student is intended to work on a Blue Waters project.

I would also like to request 500 TB of nearline project storage to allow a new research initiative involving re-analysis of large archival datasets from the national optical astronomy observatory using new algorithmic approaches. This is especially important to our institutional strategic interests. To allow efficient staging and processing of the nearline data I would also like to request 50 TB of Lustre project disk space.

The estimated utilization schedule for the requested allocation is uniform: Q1: 25%, Q2: 25%, Q3: 25%, Q4: 25%).