Big Data Challenges for Large Radio Arrays

Dayton L. Jones, Kiri Wagstaff, David R. Thompson, Larry D'Addario, Robert Navarro, Chris Mattmann, Walid Majid, Joseph Lazio, Robert Preston, and Umaa Rebbapragada
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109
818-354-7774, Dayton.Jones@jpl.nasa.gov

Abstract—**Future large radio astronomy arrays, particularly the Square Kilometre Array (SKA), will be able to generate data at rates far higher than can be analyzed or stored affordably with current practices. This is, by definition, a "big data" problem, and requires an end-to-end solution if future radio arrays are to reach their full scientific potential. Similar data processing, transport, storage, and management challenges face next-generation facilities in many other fields.**

The Jet Propulsion Laboratory is developing technologies to address big data issues, with an emphasis in three areas: 1) Lower-power digital processing architectures to make high-volume data generation operationally affordable, 2) Date-adaptive machine learning algorithms for real-time analysis (or "data triage") of large data volumes, and 3) Scalable data archive systems that allow efficient data mining and remote user code to run locally where the data are stored. [1]

TABLE OF CONTENTS

1. INTRODUCTION

Future large radio astronomy arrays, particularly the Square Kilometre Array (SKA), will be able to generate data at rates far higher than can be analyzed or stored affordably with current practices. This is, by definition, a "big data" problem, and requires an end-to-end solution if future radio arrays are to reach their full scientific potential. Similar data processing, transport, storage, and management challenges face next-generation facilities in many other scientific fields as well as a large number of data-intensive industries (financial, biotech/medical, telecommunications, etc.).

The Jet Propulsion Laboratory (JPL) is developing technologies to address big data issues, with an emphasis in three main areas:

1. Lower-power digital processing architectures to make high-volume data generation operationally affordable.

2. Date-adaptive machine learning algorithms for real-time analysis (or "data triage") of large data volumes.

3. Scalable data archive systems that allow efficient data mining and remote user code to run locally where the data are stored.

Power consumption and cooling of systems like cross-correlators for large arrays can be prohibitively expensive. An optimized ASIC architecture can provide order-of-magnitude reductions in power usage compared with traditional correlator design approaches [1-3].

Real-time data-adaptive software innovations at JPL have focused on the detection of fast (< 1 second) transient signals in high-rate data streams. These include known signals such as pulsars as well as signals that deviate from a standard dispersed-pulse profile. Here is a prime example of a situation where it is impractical to store the high time and spectral resolution data from many antennas for later analysis. A fast radio transient detection system using data-adaptive algorithms has been deployed on the Very Long Baseline Array, a facility of the National Radio Astronomy Observatory [4-6]. This technology has also been adapted to radio frequency interference excision and could be used for real-time anomaly detection in array monitoring data.

Work in the data archiving and data mining area is based on previous JPL investments in a data archive framework for Earth science missions (PCS/OODT) [7]. This framework is being used for an ever-increasing number of non-astronomy applications, and is currently being adapted for use by radio observatories.

Recent progress in each of these areas, along with possible paths for further development of the relevant technologies, will be described in the following sections. Dealing with big data issues in an integrated end-to-end manner will be an essential aspect of the design of many large observational systems in the future. We see future application of these technologies in both ground-based and space-based systems, for both astronomy and non-astronomy uses.

[1] 978-1-4577-0557-1/12/$26.00 ©2012 IEEE
[2] IEEEAC paper #1047, Version 2, Updated Nov 22, 2011

2. Data Generation

Most big data problems begin, almost by definition, with the generation of data at a rate too high to be handled by the data transport infrastructure or the real-time data processing systems available, or data volumes too large to be stored for traditional off-line analysis. This situation occurs with increasing frequency in science, where sensor technology allows ever-increasing numbers of pixels, spectroscopic resolution, and time sampling. A prime example in radio astronomy is the SKA, which will produce raw data from its many antennas at a combined rate of order PB/s. Similar examples abound in high energy physics, many types of surveillance operations, and some data-intensive industries.

One aspect of high-rate data generation that is becoming increasingly important is the high power consumption (and associated cooling requirements) of large sensor networks. In the case of radio astronomy interferometer arrays, digital cross-correlation of data from many antennas is often one of the significant power uses. JPL has developed an improved ASIC architecture (see Figure 1) that minimizes power used for data movement and memory. This approach can reduce the power consumption of a large correlator by more than an order of magnitude compared to current architectures.

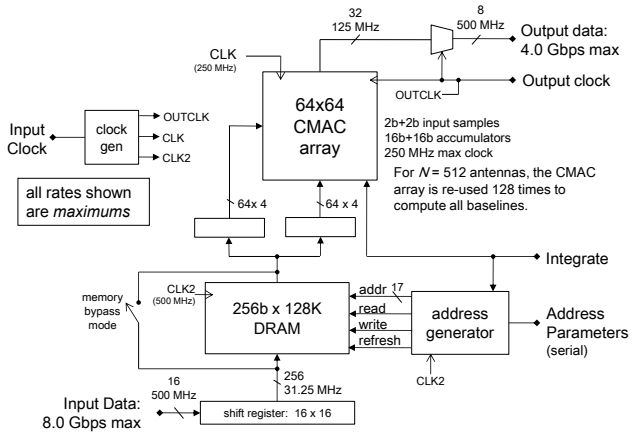

Figure 1– Conceptual block diagram of a demonstration correlator ASIC. High speed I/O uses differential signaling, with 16 parallel input bits and 8 parallel output bits. Rates shown are achievable in the IBM 32nm process [3].

A related effort involves fast transient radio signal detection at the Australian SKA Pathfinder array (ASKAP). Here the challenge is processing large quantities of data from beam-forming electronics in real time, prior to cross-correlation. The basic approach in shown in Figure 2. This development is based on FPGAs, which implement algorithms to correct for dispersion caused during signal propagation through the ionized interstellar medium, combine signals from multiple antenna, detect transients, and signal buffer memories in the beamformers to store the raw data during a transient event.

The time resolution of this system is about 1 ms, and the trigger to the data buffers must occur within about 1 second.

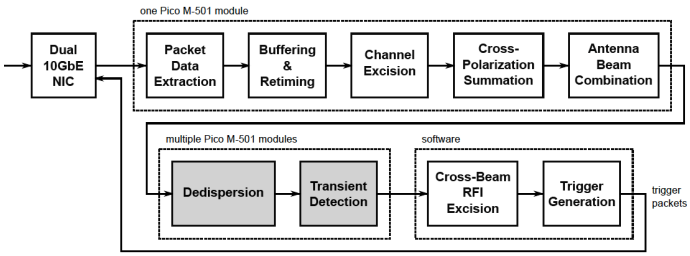

Figure 2 – Block diagram of multi-antenna, multi-beam radio transient detection for the Australian SKA Pathfinder ASKAP. Data from beamformers enter at left; trigger causes buffered input data to be saved.

This is an example of "data triage." Only a small fraction of data prior to cross-correlation (and time averaging) can be stored for later analysis. It is essential that decisions about what small fraction of raw data should be saved are made rapidly and correctly. Once data are time-averaged, there is no way to go back and extract information on fast transient signals that may have been present.

3. Data Triage

The concept of data triage can be generalized in many ways, even within the rather narrow niche of fast transient radio source detection. Figure 3 shows a fast transient detection system based on more powerful and flexible approaches to data triage. This system has recently become operational on the Very Long Baseline Array (VLBA), a continent-wide array of 10 radio antennas operated by the National Radio Astronomy Observatory [4]. It uses machine learning (data-adaptive) algorithms developed at JPL instead of hard-wired detection schemes [5].

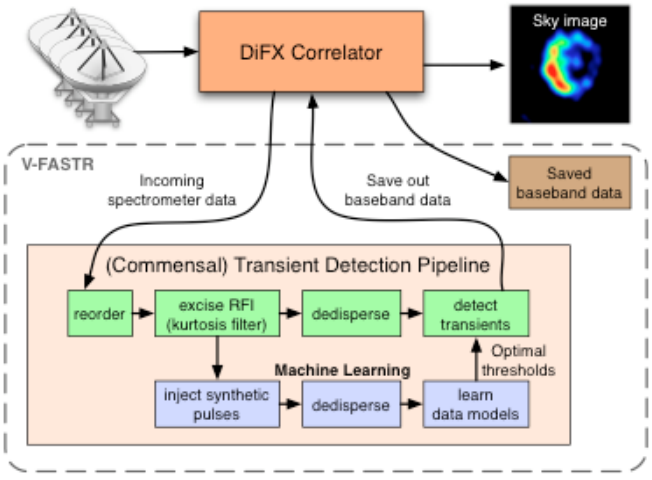

Figure 3– Block diagram of fast transient detection system currently installed on the VLBA [5]. Intelligent software (machine learning, light blue boxes) makes this system highly efficient and adaptive to interference.

The advantage of machine learning algorithms is that they can improve the effectiveness of data triage. In the radio astronomy case, an important part of the decision process is deciding if a given transient signal is from an astronomical source or terrestrial radio frequency interference (RFI). Machine learning allows the system to continuously improve its knowledge of what RFI looks like.

As an example, figure 4 shows simultaneous digitized signal voltages from nine VLBA antennas. The faint peaks marked with arrows are periodic pulses from a known radio pulsar. Everything else is thermal noise from the receivers, RFI, or system gain variations.

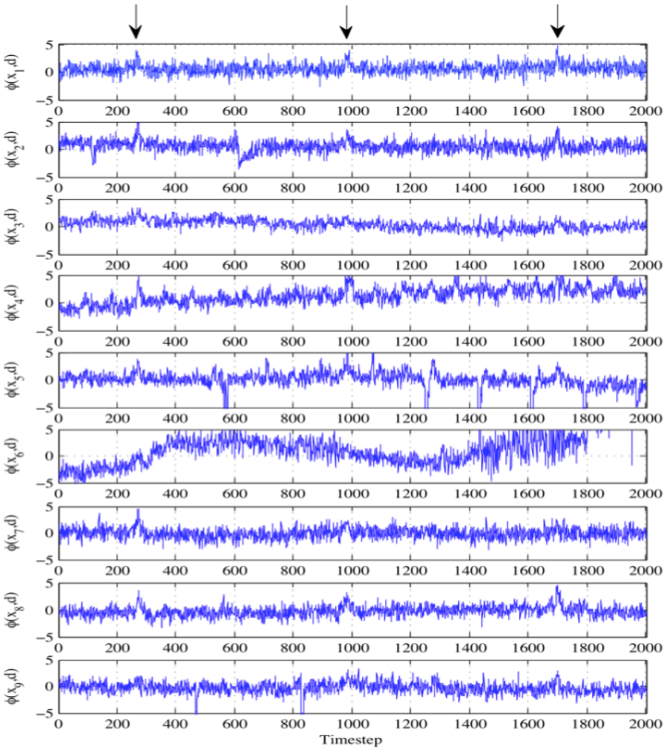

Figure 4 – Time series from nine VLBA antennas [6]. The horizontal axis covers about 4 seconds of time. The vertical axis is proportional to received signal amplitude from each of the antennas.

The ability to distinguish "interesting" signal peaks from the rest of the data is critical. A data-adaptive algorithm can be trained with examples of random noise, RFI, instrumental errors, and other known types of error and will reject signals that share those properties. In addition, the algorithm can take advantage of the large geographic separation between VLBA antennas to separate local RFI from distant sources.

Figure 5 shows how the comparison of signals from more than one antenna can help discriminate against RFI using decision boundaries as defined in [6]. In this example, a quadratic function selects the interesting events, and rejects RFI, more effectively than other decision functions.

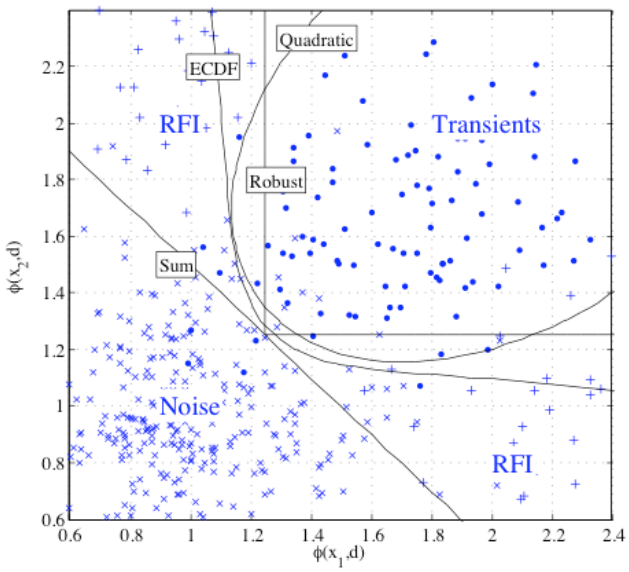

Figure 5 – Multi-station transient detection methods (labeled curves; see [6]) and their ability to separate noise, RFI, and true transients. The axes show signal strength from each of two widely separated antennas.

Figure 6 is a more quantitative summary of the performance of the various functions used in Figure 5.

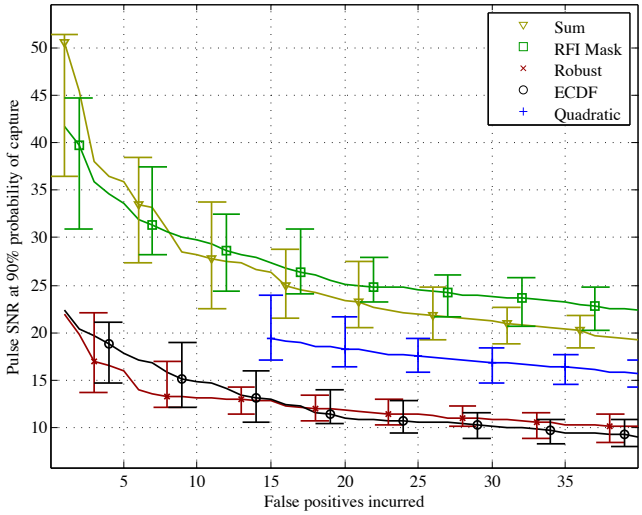

Figure 6 – Multi-station SNR detection limit for various false positive rates. Lower is better. Green and yellow lines show standard incoherent sum algorithms. Red and black show V-FASTR approach used at VLBA [6].

Figure 6 illustrates the sensitivity of various single-pulse detection approaches applied to a dataset of several hundred individual pulses from a radio pulsar (PSR 0329+54). More lenient detection thresholds achieve a higher detection rate for true pulses at the expense of more false detections. Fig 6

shows the signal to noise ratio of each pulse that can be detected with 90% certainty as a function of the number of false positive detections. The 90% confidence intervals were generated using bootstrap sampling. These results show that adaptive methods (lower red and black curves) perform significantly better. These two adaptive approaches tune their sensitivity to each independent antenna on line, and thus are better able to ignore non-interesting noise and contaminant signals.

Another, more general example of the power of machine learning techniques for signal detection is illustrated in Figure 7. Here the vertical axis is the fraction of true events detected, and the horizontal axis is the number of false alarms generated. An ideal detection system would occupy the upper left corner of this plot. The semi-supervised matching learning algorithm is clearly more sensitive and more robust than other approaches.

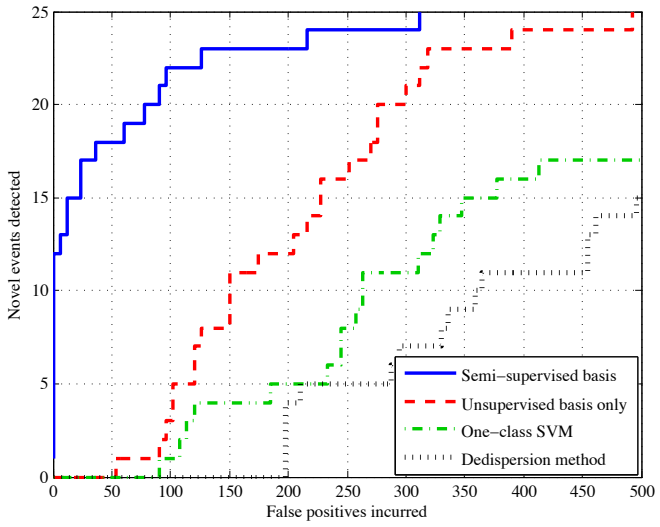

Figure 7 – Adaptive algorithm (blue curve) significantly improves signal detection performance [6].

Data triage can apply to many other situations, including the detection of instrumental anomalies in monitor and control data or real-time control of sensor properties in response to rapidly varying conditions. Intelligent algorithms are likely to be a major focus of future work in dealing with big data challenges.

4. Data Archiving and Mining

The result of most big data problems is, not surprisingly, a very large quantity of data to store, access, distribute, and mine for various types of information and understanding. The traditional approach, in which users download selected data from a central archive to analyze on their computers, will not work when the size of data sets makes data transfer an unacceptably slow process. Instead, analysis programs will have to run on computers closely linked to the archive storage infrastructure. Consequently users will need to be able to run their analysis programs and script on remote computing resources. This model raises several questions concerning security, robustness, and access.

JPL has invested in a scalable archive system that addresses these concerns. The Process Control System (PCS) was developed originally as an archive system for NASA's planetary missions, but its underlying software components have turned out to be applicable to many other large archive needs including the Climate Data Exchange, the James Webb Space Telescope, and the Early (Cancer) Detection Research Network (ERDN) [7-10]. The Object Oriented Data Technology (OODT) on which PCS is built was developed primarily by D. Crichton. It is scalable, hardware independent, database independent, and interoperable. Most importantly, OODT has a plug-in capability for user data processing tools and algorithms. OODT is the first NASA project to be distributed as open source software through the Apache Software Foundation.

Figure 8 shows the basic architectural components of PCS, and how they interact.

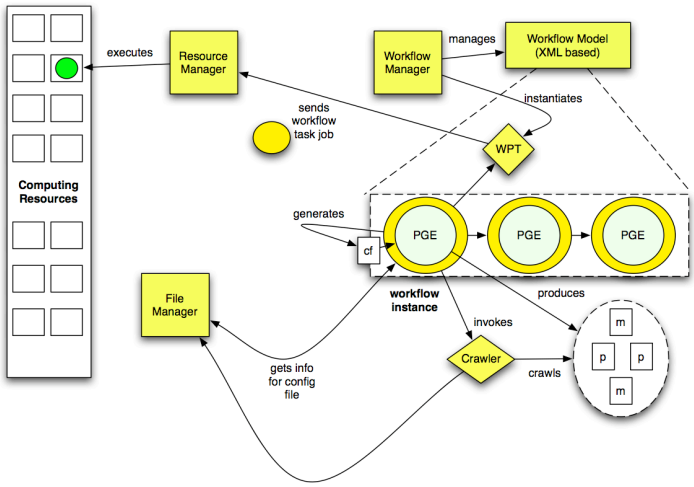

Figure 8 – The Process Control System is a set of reusable components from the open source Object Oriented Data Technology (OODT) framework [11].

Figure 9 shown an example PCS/OODT application based on a proposal to use this framework for an Expanded Very Large Array (EVLA) data pipeline and archive at the National Radio Astronomy Observatory.

The EVLA collaboration will begin with a demonstration of Apache OODT based on the EVLA summer school pipeline. In this demonstration, the Workflow Manager (WM in Fig 9) will ingest raw EVLA data via the File Manager (FM), dynamically create a script that runs standard EVLA data analysis programs to produce a set of radio images, and then stores the images and associated calibration and other meta-data. The stored results can be rapidly searched using meta-data elements to produce new data files and meta-data.

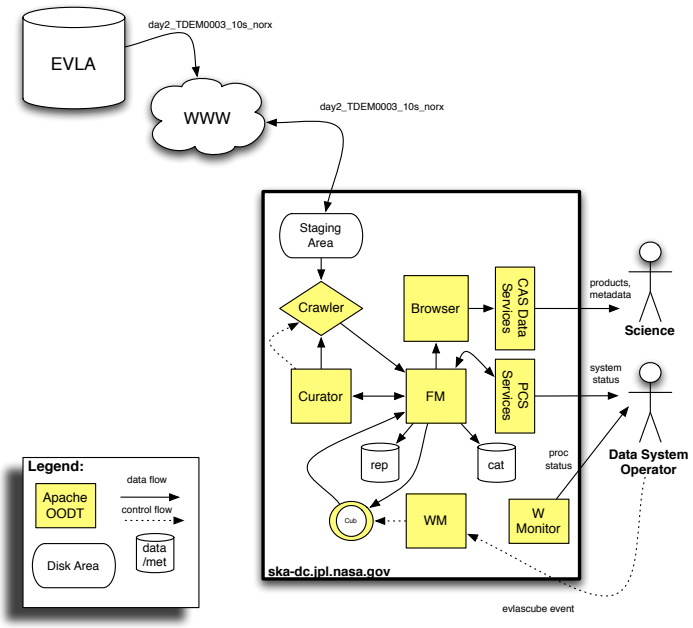

Figure 9 – A possible use of PCS/OODT for a future radio astronomy archive. EVLA is the Expanded Very Large Array in New Mexico, FM is the File Manager, and WM is the Workflow Manager.

A major goal of PCS/OODT is to provide a framework for sharing data across heterogeneous and distributed archives. As understanding more complex data continues to demand a greater degree of data assimilation and visualization, and thus access to ever larger and more heterogeneous data sets, the capabilities of PCS/OODT will be more widely useful.

5. Discussion

There are many aspects of big data problems that have not been discussed here. These include data compression and encoding techniques, long-distance data transport, very high I/O bandwidths within digital systems, high performance computing in general, and massively parallel computing architectures in particular. Each of these areas could be critical for a specific big data problem, but they have not been a focus of JPL's internal research to date because they are being addressed, with significant resources, by industry (long-distance broadband data transport, high I/O bandwidth hardware, high performance and parallel computing), or are more application specific (data compression and encoding, particularly for downlink of data from distant spacecraft).

Cloud computing is another area that often comes up in any discussion of difficult data processing problems. While this is a rapidly growing approach to large computational issues or storage of data, the nature of most big data problems is that they are data rate (I/O) limited, not computation limited. In this situation the goal is to minimize the need to move data once it has been stored, and this in turn requires close coupling between the computational hardware used for data analysis and the data storage media. We want to minimize the need to ever transfer very large data files to remote machines. Such data transfers can consume large amounts of time, power, and network bandwidth. Highly distributed computing and storage resources as epitomized by cloud computing may not be the optimal paradigm for the extreme I/O rates needed between computing resources and massive data archives.

The work was initially motived by challenges facing the SKA, and the technology being developed is still directly relevant for this project. Lower power data generation through careful consideration of data flow through ASICs has the potential to reduce one of the largest components of the SKA's predicted annual operating cost, paying for the generation of electric power. Power will be a significant operating expense independent of how it is generated.

Data triage algorithms can be used to extract additional science prior to unavoidable data averaging, and also might be able to reduce the raw data flow from antenna by adjusting the number of bits per sample in real time based on observed changes in properties of RFI. JPL has a long-standing interest in techniques that allow more data, or more optimally selected data, to be transmitted to Earth from distant planetary spacecraft to the Deep Space Network. Machine learning algorithms complement data compression and error-correcting codes to maximize the science return from missions.

Data triage algorithms may also be useful for detecting and characterizing anomalies in monitor and control data on faster time scales than monitor data is routinely logged. This is, in principle, related to the current practice of using machine learning techniques to classify variations in signals from astronomical objects (e.g., [12]).

Finally, data scale of the SKA data archive will preclude the routine transfer of data to remote user for analysis. It will thus be necessary for users to be able to run their software on computers located at the data archive site. In addition, very efficient searching and mining of the SKA data archive will be needed to maximize the scientific value of the data through combined analysis with data from other facilities. The open source OODT software tools may provide a viable framework to meet these requirements.

6. Conclusions

The need to address big data challenges is now widely recognized. One compelling reason in the world of radio astronomy is the SKA, which promises to produce data at rates and volumes orders of magnitude larger than existing facilities. Technology developed in the context of the SKA bid data problems will have much wider application. This extends to fields far beyond radio astronomy, including NASA and other agencies space missions, global and regional climate monitoring, biotech research, the financial industry, high energy physics experiments, all-sky optical monitoring programs (culminating in the Large Synoptic

Survey Telescope, whose continuous data rates will approach those of the SKA), telecommunications, and many others.

7. ACKNOWLEDGEMENTS

This research was done at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. The Very Long Baseline Array and the Expanded Very Large Array are facilities operated by the National Radio Astronomy Observatory, which is managed by Associated Universities, Inc., under a cooperative agreement with the National Science Foundation.

REFERENCES

[1] Larry D'Addario, "A Strawman Correlator for the SKA," URSI National Radio Science Meeting, Boulder, CO, January 5, 2011.

[2] Larry D'Addario, "Low-Power Correlator Architecture for the Mid-Frequency SKA," Square Kilometre Array Memo 133, March 21, 2011 (available from http://www.skatelescope.org/pages/page_memos.htm).

[3] Larry D'Addario, "Low Power Architectures for Large Radio Astronomy Correlators,". URSI General Assembly, Istanbul, August 16, 2011 (to appear in *IEEExplore*).

[4] Walter Brisken, Adam Deller, Walid A. Majid, David R. Thompson, Steven Tingay, Kiri L. Wagstaff, and Randall Wayth, "V-FASTR: Commensal Transient Detection with the VLBA," URSI National Radio Science Meeting, Boulder, CO, January 6, 2011.

[5] Randall Wayth, Walter Brisken, Adam Deller, Walid Majid, David Thompson, Steven Tingay, and Kiri Wagstaff, "V-FASTR: The VLBA Fast Radio Transients Experiment," Astrophysical Journal 735, 97, 2011.

[6] David Thompson, Kiri Wagstaff, Walter Brisken, Adam Deller, Walid Majid, Steven Tingay, and Randall Wayth, "Detection of Fast Transients with Multiple Stations: A Case Study Using the Very Long Baseline Array," Astrophysical Journal 735, 98, 2011.

[7] Chris Mattmann, Daniel Crichton, Nenad Medvidovic and Steve Hughes, "A software architecture-based framework for highly distributed and data intensive scientific applications," ICSE06, 721, 2006.

[8] Daniel Crichton, et al., "A distributed information services architecture to support biomarker discovery in early detection of cancer," e-Science, 44, 2006.

[9] Chris Mattmann, *et al.*, "Architecting Data-Intensive Systems," in Handbook of Data Intensive Computing, ed. B. Furth & A. Escalante, Springer Verlag, 2011.

[10] Chris Mattmann, A. Hart, Dayton Jones, Robert Preston, "SKA Data Processing Using Apache OODT," Science and Frontiers of Astronomy in the Era of Massive Datasets: The Promise and Challenges (Square Kilometre Array International Forum 2011, Banff, Canada), 2011.

[11] Chris Mattmann, *et al.*, "A Reusable Process Control System Framework for the Orbiting Carbon Observatory and NPP Sounder PEATE Missions," Third IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT 2009, Pasadena, CA), 165, 2009.

[12] Umaa Rebbapragada, K. Lo, C. Reed, Tara Murphy, Kiri Wagstaff, David Thompson, and Joseph Lazio, "Classification of ASKAP VAST Radio Light Curves," New Horizons in Time Domain Astronomy (International Astronomical Union Symposium 285), Oxford, UK, 2011.

BIOGRAPHY

Dayton Jones is a Principal Scientist at the Jet Propulsion Laboratory, California Institute of Technology. He has been involved in studies of space-based low frequency array mission concepts for the past two decades, most recently concepts for a large lunar-based array to observe neutral Hydrogen from early cosmic epochs. His research interests include high angular resolution imaging and high precision astrometry with very-long-baseline interferometry. He is an author on ~200 scientific publications. He has a BA in Physics from Carleton College, an MS in Scientific Instrumentation from the University of California, Santa Barbara, and MS and PhD degrees in Astronomy from Cornell University.