The Case for Mobile Devices in Environmental Observing Systems

Sameer Tilak, Kenneth Chiu, Gregory Quinn, Tony Fountain, and Chaitan Baru

***Abstract*— To address challenging issues such as global warming, invasive species, infectious diseases, desertification, salination, pollution, and land use; scientists have been embedding cyberinfrastructure in the form of large scale observatories into the environment. This cyberinfrastructure also extends into the lab, providing the scientist with tools to analyze and curate data at unprecedented spatio-temporal scales. To date, however, little work has focused on extending the cyberinfrastructure to the scientist while in the field, thus allowing them to interact with the embedded cyberinfrastructure, such as the sensor network. To bridge this gap, we argue that an integrated, management and analysis platform should be extended out to the field through the use of mobile devices such as PDAs and smart phones. We identify several key use cases and user bases, and explain how such a system can give scientists, field technicians, and other users precise control over sensors while in the field, thus enhancing their ability to collect high quality data. The use of mobile devices would thus allow them to circumvent the existing cumbersome and costly process of sensor deployment, (re-) configuration, and data collection.**

I. Introduction

Issues such as global warming, invasive species, infectious diseases, desertification, salination, pollution, and land use all present difficult and challenging scientific questions. Rapid advancements in the development and deployment of cyberinfrastructure and its wireless extensions to sensor networks (embedded cyberinfrastructure) are playing a key role in addressing these grand scientific challenges. To that end, existing and emerging large-scale observing systems comprised of hundreds to thousands of sensors will gather data at the prescribed spatial and temporal granularity. EarthScope [1], and NEON [3] are examples of such large-scale observatories which would consist of thousands of sensors, tens of thousands of data streams, and hundreds of thousands of end users. These observing systems would include two main components namely, cyberinfrastructure– large-scale distributed system comprised of computational and storage resources and embedded cyberinfrastructure – numerous sensors embedded deeply into our physical world.

We have started to embed our cyberinfrastructure into ecosystems for measurement, and we are developing cyberinfrastructure for the scientist in the lab. But we have very limited, if any, cyberinfrastucture for the scientist in the field that can help them interact with the embedded cyberinfrasture (the sensor network). Also, currently, if there is any cyberinfrastructure support for the scientists in the field to interact with the sensor network, it is primarily limited to data downloading. Rather than using these devices just for downloading the data during field visits (an easy to imagine usage), we argue that a much broader user base and a much richer usage needs to be considered. We identify several key use case scenarios and user bases.

In order to bridge this gap, in this position paper we argue to extend an integrated, management and analysis platform out to the field through the use of mobile devices such as PDAs and smart phones. When scientists, field technicians, and other users are out in the field, this would give them a more precise control over sensors and enhance their ability to collect high quality data. The use of mobile devices would allow them to circumvent the existing cumbersome and costly process of sensor deployment, configuration, data collection.

In addition, we discuss in detail the following three goals and their associated software challenges: (1) build a service-oriented architecture for in-situ dynamic resource discovery and sensor tasking capability, (2) provide visualization, especially of environmental datasets, on handheld devices and (3) provide lightweight machine learning algorithms on handheld devices.

II. State-of-the-art

Current sensor networks for ecological systems provide an embedded cyberinfrastructure for sensing. But to date, when scientists are out in the field, they do not have mature, easy to use cyberinfrastructure tools to interact with this embedded cyberinfrastructure. There, the scientist must rely on traditional, ad hoc systems comprising a motley assortment of notebooks, laptops, and other tools.

We now briefly describe a typical sensor network setup for environmental and ecological monitoring applications. Typically, a digital or analog sensor is connected to a datalogger device, which acts as a simple computer to drive the sensor and store the results. The datalogger may be equipped with a wireless radio, allowing it to be periodically polled for data transfer. Some dataloggers are stand-alone, which means that a scientist must periodically visit the site to download data to a laptop or PDA. At present, typically, scientists do not have access to handheld devices for in-situ operations and when they are out in the field, they do not have access to any significant computational tools. The field-deployed sensors are capable of collecting large amounts of data, which are transferred to a base station (mostly WHAT DOES "mostly" MEAN HERE? the datalogger is equipped with wireless radio) and the reported data is stored in databases installed on machines located in the base station (BS). It is important to make sure that the collected data meets the required quality standards. At present the algorithms that ensure data quality assurance and control (QA/QC) of the incoming data run on machines that are installed in the base station.

Typically, a scientist visits the field, to (re) configure the network and then returns to the base station to ensure the quality of incoming data. If the incoming data does not meet the required quality standards, he analyzes the problem and visits the field to fix it. Sometimes, a number of back-and-forth trips between the base station and the field are required before the data collection is successful. This is clearly a cumbersome and costly process and therefore developing tools, which can be made available on handheld devices to be used during the field visits would be a very useful addition. For example, these tools would allow them configure the network and ensure that the data collection is happening smoothly (e.g,. ensuring that the data quality meets the required standards) during their field visits.

In addition, the current generation of sensor networks is application-specific and is exposed only to a limited set of users. Despite the relatively recent emergence of sensor networks as a field of study, already a large number of sensor hardware platforms and software elements (operating systems, networking protocols, data base systems, etc.) have emerged. Therefore available sensors are likely to be heterogeneous in terms of their capabilities and software elements. They may provide different types of services and allow different configurability and access. Handling this heterogeneity is a daunting task. Currently, there is very little, if any, software infrastructure, which would allow scientists to visit the field, discover available sensors dynamically, and task them with the experiments they want to conduct.

III. Motivating Scenario

To help motivate our work, we now present this concrete scenario in the context of lake monitoring project at North Temperate Lake (NTL) [4] site in Northern Wisconsin, which serves as our test site[1]: A scientist has a particular hypothesis she wishes to test. This hypothesis will help her develop and improve the model she is currently working on. During the experiment planning stage, the scientist needs to decide appropriate sensor parameters such as sampling frequency. That needs to be decided based on past data and data models. For example, in order to adjust depths of the temperature sensors immersed inside the lake, she visualizes data obtained from the temperature chain in last six months at various depths, at fixed time intervals. This data consists of three dimensions, namely, lake temperature, sensor location, and reading timestamp. Another crucial factor that might influence the placement of sensors is circulation patterns. She uses appropriate models to predict future circulation patterns and then uses this information in conjunction with the visualization of historical data to decide the appropriate locations for sensor placement. All this happens inside the base station using existing cyberinfrastructure tools on powerful hardware.

In order to deploy the sensors the scientist now goes out to the field. In the current state, the datalogger is instrumented with 20 sensors. She adds a new sensor to the data logger. Therefore, the program on the data logger needs to be updated so that data can be acquired from the newly attached sensor. However, it may be the case that something can go wrong during the auto configuration. Therefore it is important to make sure that sensor network configuration in terms of sensor placement, sensor reporting frequency, sampling variables etc. is correct. Several factors including software bugs, hardware malfunctioning, human error (improper/loose connections) can disrupt the sensor network functionality. Manual debugging of sensor network health is very time-consuming and error-prone.

If the handheld device carried by the scientist is equipped with machine learning tools that would enable him to acquire data streams, analyze them and classify them into various categories such as operational, errors, software errors, hardware errors, biotic errors, abiotic errors etc.- that would be helpful. Also, the handheld devices with visualization capabilities would be very

[1]NTL is just a vehicle to demonstrate concrete problems, however, these problems also occur in other environmental observing sytems.

useful since it would enable scientists to observe a wide range of conditions including state-of-the health of the sensor network to understanding results of an ongoing experiment without going back to the lab. For example, if the scientist can look at various system states in the form of visually informative graphs, he can easily understand the state of the network as well as that of an ongoing experiment. Examples of such maps include network energy map (available energy at all parts of the sensor network) layered on top of the terrain map and sensor network connectivity graph (map of sensor network along with links color coded as a function of radio signal strength). Finally, once the sensor network is operating in a desirable state, the data collection process begins. The field-deployed sensors are now capable of collecting large amounts of data, which is stored in databases installed on machines located in the base station. Scientists can use this gathered data to gain key insights into complex physical ecosystems.

IV. Proposed Approach

To improve the collection of data, we propose to extend an integrated, powerful management and analysis platform out to the field through the use of mobile computing. This will allow scientists to have a powerful facility in the field to make decisions and conduct experiments in an efficient manner. In order to make use of the platform, we will develop visualization, and machine learning tools to help them analyze the data.

Mobile computing through the use of devices such as laptop computers, Tablet PCs and smaller, less powerful, handheld devices has become a part of a new revolution in data availability and access. These devices will allow scientists to interact with sensors directly when they are in the field. Powerful visualizations will provide highly informative displays of historical and real-time data. It would allow scientists to easily observe trends in data and would provide facilities to select a subset of important and relevant datasets so that they can understand dynamic relationships between various system variables. It would also provide easy to use navigation and search facilities so that scientists can search the history for interesting or similar operating conditions and to compare them with ongoing operations. Due to physical limitations, sensors might drift considerably over long time periods and to that end, visualization capability can be used for real-time in-situ sensor calibration.

We argue that automated sensor deployment, sensor network diagnostics, and event detection are possible if machine learning algorithms can be applied in the field, though mobile computing. However, PDA and handhelds are resource-constrained devices. This prohibits the use of traditional heavyweight (in terms of processing power and memory consumption) machine learning algorithms on these devices. To that end, drawing upon our considerable experience in machine learning application [5]–[7] we propose to adapt traditional machine learning and feature selection algorithms for handheld devices. This will make the mobile computing more effective and will help researchers make decisions in the field.

V. Goals

A. Goal 1: Build a services oriented architecture for in-situ dynamic resource discovery and sensor tasking capability

In contrast to the current generation of sensor networks that are application-specific and exposed only to a limited set of users, heterogeneous sensor networks will start to be used by users in their surrounding environment dynamically. We envision the second generation with commodity sensors and sensor networks providing standard services that can be composed by various clients into a wide range of applications. For example, a given sensor network deployment might be shared by many scientists to conduct their own experiments. Despite the relatively recent emergence of sensor networks as a field of study, already a large number of sensor hardware platforms and software elements (operating systems, networking protocols, data base systems, etc.) have emerged. Therefore available sensors are likely to be heterogeneous in terms of their capabilities and software elements. They may provide different types of services and allow different configurability and access. The goal is to provide software infrastructure that will allow scientists to visit the field, discover available sensors dynamically, and task them with the experiments they want to conduct. Increasingly advanced and accessible applications will be enabled as sensor networks are shared dynamically and ubiquitously, allowing clients to unite disparate components on demand into virtual sensor networks [8].

Providing a modular, flexible, and effective service oriented framework that hides these individual sensor network differences and provides a consistent, unified view of a single virtual sensor network to its client is a challenging task. Critical components in realizing such a vision are dynamic resource discovery and tasking of sensors. The nature of resource discovery requires modular standards that allow interoperation among heterogeneous software and hardware systems. To that end we argue employing a services oriented architecture [9] to develop light-weight and energy efficient protocols for dynamic resource discovery and sensor tasking, which can be run on handheld devices. Therefore, when scientists are in the field, it will enable them to dynamically discover and task sensors using handheld devices. Within our architecture, we will explore various system level tradeoffs such as the one between interoperability and energy efficiency.

B. Goal 2: Provide visualization on Handheld devices

It is now not at all unusual to find field researchers and technicians using a tablet PC, and more recently Pocket PC devices, in the context of their work. The field scientist would greatly benefit from in-situ visualization and machine learning tools for monitoring various system states in the form of informative graphs through which they can easily understand the state of the network as well as that of an ongoing experiment. Examples of such maps include network energy map [10](available energy at all part of the sensor network) layered on top of the terrain map and sensor network connectivity graph (map of sensor network along with links color coded as a function of radio signal strength).

The powerful combination of connected handheld devices and highly capable mobile operating system enables field access to complex data visualizations of the kind hitherto associated with fully-fledged PCs. Here we propose the use of advanced mobile 2D and 3D graphics APIs such as Direct3D Mobile [11]and OpenGL ES [12] to create an application that will enable the presentation of streaming time-variant sensor data to field operatives.

Data will be accessed using cellular wireless internet connectivity and, in some cases, 802.11 field station signals, and will be presented to the operator in a highly visual and interactive format. Where available, the application will utilize hardware graphics acceleration in the Smart Phone, to significantly enhance data rendering capability and end user experience. To date, this kind of usage has been somewhat restricted by limited availability of visualization APIs and development tools for handheld devices. This scenario is now changing dramatically, as PDA functionality becomes incorporated with cellular phones in so-called Smart Phone devices.

The Nokia N93 [13] would be of particular interest for the purposes of both data visualization in the field and also remote data collection. Since this handset uses the TI OMAP2420 platform [14], with its powerful video collection and 3D rendering capability, coupled to the extremely capable S60 Symbian 9.1 operating system, we believe that this will be an excellent choice of development platform for the proposed research.

C. Goal 3: Provide machine learning algorithms on Handheld devices:

Handheld devices are resource-constrained devices. This prohibits the use of traditional heavyweight (in terms of processing power and memory consumption) machine learning algorithms on these devices. To that end, in this paper, with handheld devices as the target platform, we discuss the following three activities.

(1) Automated sensor deployment strategies can help optimize sensor deployment for the field scientist. For example, if the number of sensors is limited and a particular classification task is required (e.g., having 10 sensors to classify lake metabolism patterns), the value of a newly placed sensor can be evaluated by measuring the resulting increase in classification accuracy. This can be done by measuring the feature score of the variable represented by the sensor, through a feature score formula such as Golub/Furey's Precision Strength formula [15], t-test, or Battacharyya Distance [16]. Examples of the underlying classification algorithm would be Support Vector Machines [17] or Bayesian Classifiers [18]. We believe that developing libraries, which contain the aforementioned feature selection and machine learning algorithms for resource constrained handheld devices would greatly facilitate the deployment process.

(2) Expert System: At NTL, in the current system design a scientist often visits the field and reconfigures the network (e.g., he either adds/removes a sensor or changes the sensor parameters such as sampling frequency). Typically, as a result of this reconfiguration, the program on the data logger needs to be updated. It is often the case that, something can go wrong during this process. Several factors including software bugs, hardware malfunctioning, human errors (improper/loose connections) can disrupt the sensor network functionality. Therefore it is important to make sure that sensor network configuration in terms of sensor placement, sensor reporting frequency, sampling variables etc. is correct. Manual debugging of sensor network health is very time-consuming and error-prone. These issues can be mitigated by machine learning assistance. To that end, we believe that an expert system that can be used to diagnose and help troubleshoot the hardware or software problems [19]. It would assist the scientist to classify the current state of the network into various categories such as operational, error state because of (i) software error, (ii) hardware error, (iii) error due to biotic factors, or (iv) due to abiotic factors etc. This expert system would be designed keeping in mind the small amount of memory and computation power provided by PDAs and Smart Phones. Calibrating sensors and checking the quality of

the sensor data stream at the field will eliminate the need for extra trips. Automatic calibration and quality check of the data stream can be assisted using statistical quality control methods [20].

(3) Event Detection: When the scientist is in the field, various time series analysis techniques from statistics and machine learning can help him detect the abnormal scientific events and ensure quality of the sensor data [20]–[22]. To that end, one can apply, evaluate, and refine a variety of algorithms ranging from statistical methods such as an AutoRegressive Moving Average (ARMA) and Cumulative Sum (CUSUM) to more advanced machine learning techniques such as Kalman filter, Bayesian Networks [19], and Support Vector Machines (SVM) [23]. However, we would like to point out that to develop and rigorously test the accuracy of these machine learning algorithms, a close interaction with the domain scientists is crucial.

VI. Discussion

The goal of this position paper is to present several use case scenarios for mobile devices in the context of an environmental observing system. Rather than using these devices just for downloading the data during field visits (an easy to envisage scenario), we argue that a much broader user base and a much richer usage needs to be considered. For example, field technicians, scientists, and end users can use the mobile devices for a variety of operations ranging from checking (and visualizing) state of the health of the network, checking (and visualizing) status of running experiment, sensor network (re) configuration, and abnormality detection to name a few. We hope that these use cases would be directly applicable to large-scale environmental observing systems, which are poised to become the dominant means for studying a variety of natural phenomena [1]–[3]. Addressing these issues would require a multi-pronged approach and an interdisciplinary team of domain scientists, sensor network researchers, experts of mobile phone platforms, and machine learning experts. We believe that a discussion of these use case scenarios, associated technical challenges, and possible approaches would be invaluable to the sensor network community and would assist with observing system design and engineering [1]–[3].

References

[1] "The earthscope project," http://earthscope.org.
[2] "The ocean research interactive observatory network," http://www.orionprogram.org.
[3] "The national ecological observatory network," http://www.neoninc.org.
[4] "The north temperate lakes long term ecological research program," http://lter.limnology.wisc.edu/.
[5] A. Bulut, A. Singh, P. Shin, T. Fountain, H. Jasso, L. Yan, and A. Elgamal, "Real-time nondestructive structural health monitoring using support vector machines and wavelets," in *12th The International Society of Optical Engineering Annual International Symposium, Smart Structures/NDE*, 2004.
[6] D. Pennington, H. Jasso, P. Shin, and T. Fountain, "The effect of landscape heterogeneity on classification accuracy: a comparison of classifier prediction in sub-optimal sampling conditions," in *in the Seventh Workshop on Mining Scientific and Engineering Datasets, SIAM International Conference on Data Mining*, 2004.
[7] H. Jasso, P. Shin, T. Fountain, and D. Pennington, "Using wavelets for the classification of hyperspectral images," in *Fourth European Conference on Ecological Modelling and Fourth International Workshop on Environmental Applications of Machine Learning*, 2004.
[8] S. Tilak, K. Chiu, N. Abu-Ghazaleh, and T. Fountain, "Dynamic resource discovery for wireless sensor networks," in *IFIP International Symposium on Network-Centric Ubiquitous Systems*, 2005.
[9] N. Cotofana, L. Ding, P. Shin, S. Tilak, T. Fountain, J. Eakins, and F. Vernon, "An soa-based framework for instrument management for large-scale observing systems (usarray case study)," in *ICWS*, 2006.
[10] R. M. et al, "Prediction based energy map for wireless sensor networks," in *Ad Hoc Networks Journal*, 2005.
[11] Http://msdn2.microsoft.com/enus/library/ms172504.aspx.
[12] Http://www.khronos.org/opengles.
[13] Http://www.nokia.com/nseries/index.html?loc=inside,main_n93.
[14] Http://focus.ti.com/pdfs/wtbu/TI_omap2420.pdf.
[15] T. Golub and et al., "Molecular classification of cancer: Class discovery and class prediction by gene exp ression monitoring," in *Science*, 1999.
[16] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," in *Bull. Calcutta Math. Soc.*, vol. 35, 1943.
[17] N. Cristianini and J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods," in *Cambridge University Press*, 2000.
[18] P. Langley, W. Iba, and K. Thomson, "An analysis of bayesian classifiers," in *Proceeding of National Conference on Artificial Intelligence*, 1992.
[19] S. O. Fang Liu, Plamen K. Nikolov, "Parametric fault diagnosis for analog circuits using a bayesian framework," in *24th IEEE VLSI Test Symposium*, 2006.
[20] E. L. Grant, "Statistical quality control," in *McGraw-Hill*, 1988.
[21] C. Chatfield, "The analysis of time series: An introduction," 2004.
[22] W. Wong, A. Moore, G. Cooper, and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," in *In Proceedings of the 20th International Conference on Machine Learning*, 2003.
[23] W. Hu, Y. Liao, and V. R. Vemuri, "Robust anomaly detection using support vector machines."