End-to-End Cyberinfrastructure for Real-Time Environmental Decision Support

David Hill[1], Barbara Minsker[2], Yong Liu[1], Jim Myers[1]
(1) National Center for Supercomputing Applications, University of Illinois, Urbana, IL 61801
(2) Department of Civil and Environmental Engineering, University of Illinois, Urbana, IL 61801

Recent advances in sensor technology are facilitating the deployment of sensor networks into the environment that can produce measurements at high spatial and/or temporal resolutions. Through telemetry, these measurements can be delivered in near real time as data streams for use in real-time applications. These data streams can be instrumental in furthering our understanding of the environment, in monitoring and modeling the quality of the environments in which the sensors are deployed, and in providing real-time decision support for managing environmental systems. However, in order to realize these benefits, several challenges must be addressed, such as accessing heterogeneous data streams from multiple agencies; validating the data and fusing or transforming them for a desired application; providing high-dimensional models, visualizations, and decision support systems to explore data and forecast future conditions; and sharing the resulting knowledge. This poster will present an end-to-end cyberinfrastructure that can make these tasks possible in near real time, using a case study on real-time decision support for the combined sewer system in the Chicago Metropolitan Area.

Combined sewer systems (CSSs), such as Chicago's, exist in 746 communities located in 32 states; they are used in many of the nation's largest cities. Because these systems rely on a common infrastructure to convey both municiple wastewater and stormwater, intense storms can cause critical loading of the wastewater infrastructure resulting in combined sewer "overflow" (CSO) events (i.e., events where wastewater is discharged untreated into the environment in order to relieve pressure within the sewer system). A decision support system is being created that uses cyberinfrastructure to integrate heterogeneous sensor data from multiple agencies, to run models to predict the CSO occurrences, and to determine optimal storm-scale control sequences in real time that minimize the effects of CSOs on the environment.

The first challenge in developing the decision support system is accessing the relevant data. The Chicago CSS is owned by the municipalities within the Greater Chicago Area (e.g, the City of Chicago) yet managed by another agency, the Metropolitan Water Reclamation District of Greater Chicago (MWRDGC). Additionally, the weather data (particularly rainfall data) are collected in various forms and scales (e.g., rain gauge vs. radar data) by a number of different agencies such as the National Weather Service and the United States Geological Survey. Thus, data such as the physical structure of the system, system loading, and so forth must be retrieved from a number of different information storage systems. To support near-real-time use of the data, then, cyber-infrastructure must first facilitate the task of data integration (i.e., bringing the data from these multiple information systems to a central location and transforming or fusing them as needed). Numerous "digital watersheds" are being created to support various aspects of data retrieval. In our work, we are creating a semantically enhanced and integrated digital watershed that will facilitate the integration of multi-agency data as well as other analytical and visualization tasks that will be discussed shortly. The digital watershed system consists of semantic content management middleware (Tupelo), a semantically enhanced streaming data toolkit, virtual sensor management functionality, a workflow system, and a RESTful (REpresentational State Transfer) web service that can trigger workflow execution on demand or when particular events occur. This loosely coupled architecture presents a generic framework for constructing a Web 2.0–style digital watershed, through which users will be able to create virtual sensor data streams and become not only data consumers, but producers.

After the raw sensor data streams are integrated into the digital watershed, they are processed in two important ways. First, the data are filtered by an anomaly detection algorithm that detects data that deviate markedly from historical patterns. Anomalous data can be caused by sensor or data

transmission errors or by infrequent system behaviors that may be of interest to scientific and regulatory communities. When anomalies are identified, an alert can be generated; additionally, the data can be tagged with metadata indicating that the data were identified as anomalous. The second processing step involves repurposing (i.e., fusing and transforming) the data as needed for the particular real-time application. The repurposing is performed by virtual sensor middleware within the digital watershed environment. Using methods designed by domain scientists and published in workflows, the virtual sensor middleware performs spatial, temporal, and thematic transformations on the raw sensor data stream to provide a virtual sensor data stream. For example, the virtual sensor data may represent direct measurements that have been transformed to a different spatiotemporal scale, that are the product of merging direct measurements from multiple sensors at different spatiotemporal scales, or that are indirect measurements of conditions not being directly monitored and are created by combining observations from a group of heterogeneous physical sensors. In the CSO decision support system, virtual sensors will be essential for providing sewershed-scale measurements of rainfall, which are being created by repurposing data from the Next Generation Weather Radar (NEXRAD) system. The advantages of using virtual sensor cyberinfrastructure for these operations are that the results can be published continuously in near real time; that metadata can be automatically captured to record how the data were transformed; and that the algorithms can be shared, verified, and further customized for other applications by the research community.

Workflow systems are invaluable for performing such repeated operations. In addition to providing a visual platform for constructing models and data processing methods from model components, workflows also facilitate community validation of the results of these operations as well as the sharing of operations. Thus, the CSO decision support application uses a workflow system to manage the operations performed on the raw sensor data. Because many parts of the decision support system will employ legacy code created by different developers, it is important that the workflow system be able to integrate these components without having to rewrite them in a common language. Additionally, provenance tracking is necessary to keep a record of how the data have been transformed, to provide support for decision makers and for community review of the system construction and results. Cyberintegrator is one such workflow system used in this work. It provides support for the simple integration of methods implemented in different programming languages (e.g. Matlab, Fortran, etc.), as well as provenance tracking; distributed shared data access; remote execution; workflow functionality; model publication; metadata generation; and annotation of data, modules, and workflows.

Workflows will also be used to perform the modeling and optimization used to determine control strategies. The integrated raw and virtual sensor data will be used as inputs for forecasting the outcomes of decision strategies on the Chicago wastewater system. The workflow system will streamline data input as well as storage and visualization of the results. Furthermore, using a workflow for these processes will facilitate exploratory research on improved modules because old modules can be swapped out for improved versions, and, using the provenance tracking feature, the new workflow can be rerun using the exact same input as the old workflow, such that the results can be directly compared.

Finally, in order to communicate and explore the optimal decision sequences developed by the decision support system, high-dimensional visualization capabilities are necessary. The data and modeling results will represent complex spatiotemporal features at multiple scales; thus, the visualization should provide adaptive zooming capabilities as well as customized views that facilitate summary and interpretation of multiscale data. Additionally, because these data and modeling results are georeferenced, the visualization framework should allow for simple inclusion of georeferenced images and other types of geospatial data that provide a sense of place (e.g., road network and political boundaries) to the visualization.

This poster discusses these general cyberinfrastructure needs for real-time environmental decision support using a real-world case study. During this discussion, specific examples of cyberinfrstructure tools will be given, but other tools exist that may perform similar functions.