OPEN SOURCE DATA TURBINE INITIATIVE

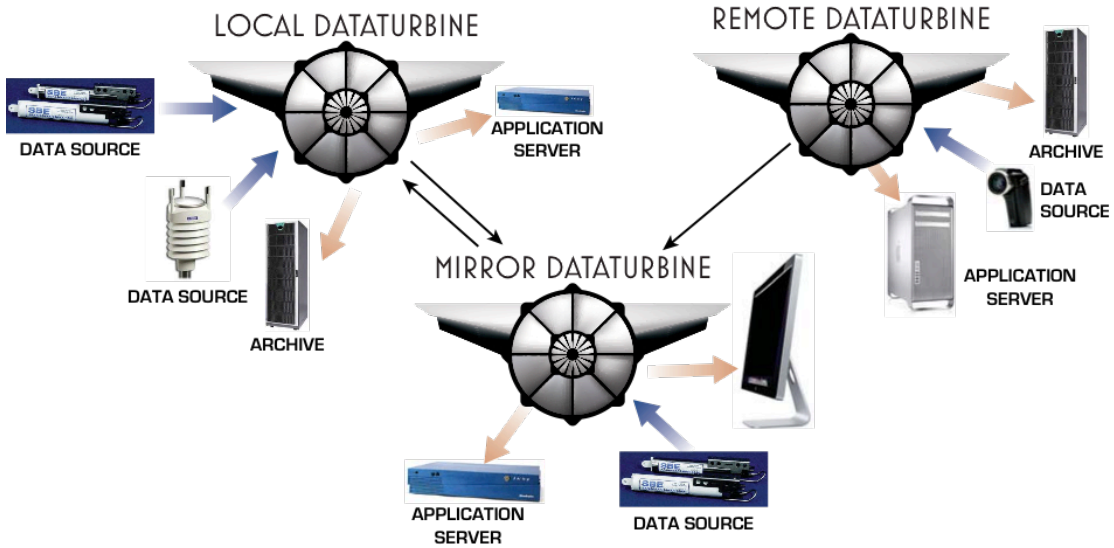
Empowering the Scientific Community with Streaming Data Middleware

Tony Fountain, University of California, San Diego (tfountain@ucsd.edu)

Environmental science and engineering communities are now actively engaged in the early planning and development phases of the next generation of large-scale sensor-based observing systems. These systems face two significant challenges: *heterogeneity of instrumentation* and *complexity of data stream processing*. Environmental observing systems are complex distributed systems. They incorporate instruments from across the spectrum of complexity, from temperature sensors to acoustic Doppler current profilers, to streaming video cameras, and to synthetic aperture radar. They operate under a variety of networking conditions, including wired and wireless, persistent and intermittent. They have stringent requirements on data timeliness and integrity. Managing these instruments and their data streams presents serious challenges in systems development and operations. The Open Source DataTurbine (OSDT) Initiative was launched in October 2007 with a two-year grant from the National Science Foundation Office of Cyberinfrastructure (award #OCI-0722067) to address these challenges through the publication, enhancement, and promotion of the DataTurbine streaming data middleware (www.dataturbine.org).

Streaming Data Middleware:

DataTurbine is a real-time streaming data engine. It is an open-source middleware product supported by NSF, NASA, and private industry. It has been successfully deployed in numerous environmental monitoring applications, from coral reefs to civil engineering to atmospheric science. The DataTurbine middleware satisfies a core set of infrastructure requirements that are common in environmental observing systems, including reliable data transport, a framework for integrating heterogeneous instruments, and a comprehensive suite of services for data management, routing, synchronization, monitoring, and visualization. From the perspective of distributed systems, the DataTurbine middleware is a "black box" to which applications and devices send and receive data. DataTurbine handles all data management operations between data sources and sinks, including reliable transport, routing, scheduling, and security. DataTurbine accomplishes this through the innovative use of flexible network bus objects combined with memory and file-based ring buffers. Network bus objects perform data stream multiplexing and routing. Ring buffers provide tunable persistent storage at key network nodes to facilitate reliable data transport.



DataTurbine Partners and Communities:

The Open Source DataTurbine Initiative benefits from the active participation from a diverse collection of international research groups. These groups contribute their own resources to advance the mission of the DataTurbine Initiative. The following are brief descriptions of some of the main community members and their interests and investments in DataTurbine. A key role of our UCSD research group is community support. This includes code management and publication, providing training materials and documentation, and coordinating activities across multiple groups. Through code developments and community support the Open Source DataTurbine Initiative provides essential cyberinfrastructure services to numerous science and engineering groups. Here are some of the key communities who participate in the Open Source DataTurbine Initiative:

GLEON www.gleon.org, is a grassroots network of limnologists, information technology experts, and engineers who have a common goal of building a scalable, persistent network of lake ecology observatories.

CREON www.coralreefeon.org, is a collaborating association of scientists and engineers from around the world striving to design and build marine sensor networks.

PRAGMA www.pragma-grid.net, is an NSF-sponsored open organization in which Pacific Rim institutions collaborate to develop grid-enabled applications throughout the Pacific Region.

PRIME prime.ucsd.edu, Launched by PRAGMA in 2004, the NSF-sponsored PRIME provides UCSD undergraduates the opportunity to participate in international research and cultural experiences that will better prepare them to participate in the 21st century's global workplace.

MoveBank www.movebank.org, is an NSF-sponsored open community with the common interest of remotely monitoring organisms in their habitats and a goal of developing and deploying technologies for gathering data on free-ranging organisms.

NEES www.nees.org, is an NSF-sponsored national network of 15 experimental facilities, collaborative tools, a centralized data repository, and earthquake simulation software, all linked by the ultra-high-speed Internet2 connections of NEESgrid.

MBARI www.mbari.org, is a privately funded, nonprofit oceanographic research Institution.

NCEAS www.nceas.ucsb.edu. The NSF-funded REAP project is a large initiative involving six institutions: National Center for Ecological Analysis and Synthesis (UCSB), San Diego Supercomputer Center (UCSD), Center for Embedded Network Sensing (UCLA), Genome Center (UCD), Oregon State University, and OPeNDAP, Inc.

CUAHSI www.cuahsi.org, is an organization of more than one hundred universities. Its mission is to foster advancements in hydrologic sciences.