

# Data stream management in global-scale ecological observatory networks

Ebbe Strandell, Hsiu-Mei Chou, Yao-Tsung Wang, Fang-Pang Lin  
National Center for High-Performance Computing, Hsinchu, Taiwan

Sameer Tilak, Tony Fountain, Peter Arzberger, Arcot Rajasekar  
San Diego Supercomputer Center, University of California, San Diego, USA

## Introduction

Ecological Observatory Networks, such as GLEON, CREON and Ecogrid, are playing an important role to drive new scientific discovery. One of the current key issues is that the size of the collected data is growing exponentially as these networks expand, and over time diverse data sources can accumulate huge amounts of data. Methods to access, manage, and archive this data effectively has emerged to a challenging field in IT development.

In this work, we demonstrate a robust data stream management system to bridge users, sensors and persistent storage pools in a scalable manner. The system is designed to work with any type of streaming data (i.e numerical, audio, video) and allows users to subscribe to two-way data channels on the fly, either for sources of dynamic data or for legacy data. Parallel data streaming from two ecological observatories in Taiwan is used as proof-of-concept, thus allowing us to demonstrate how the system can be effectively used for sharing data amongst global science and IT communities.

## Method

Our streaming data management system incorporates three main components: RBNB DataTurbine, Storage Resource Broker and an external agent to manage internal data transfers. These are briefly explained below:

**Ring Buffered Network Bus (RBNB)** DataTurbine is a product of Create Inc. that has been released as open-source in collaboration with the San Diego Supercomputer Center (SDSC). The RBNB DataTurbine provides excellent basis for developing robust streaming data middleware. The system satisfies a core set of critical infrastructure requirements including reliable data transport, the promotion of sensors and sensor streams to first-class objects, a framework for the integration of heterogeneous instruments as well as a comprehensive suite of services for data management, routing, synchronization, monitoring, and visualization.

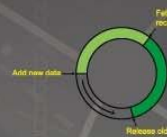


Figure 1. Ring buffer.

## Kenting's underwater observation site

Kenting National Park is located on the southernmost tip of Taiwan and is known for its dazzling coral reef. A system of 10 underwater cameras facilitates marine researchers to closely monitor the reef and to track changes in the environment. The observation system is divided into three different sites, each equipped with a video server to convert analog video signals into digital MJPEG streams. A wireless connection is used to transmit each video stream to a local monitoring station from where it can be accessed from the outside world (Figure 4).



Due to bandwidth limitations, video resolution is set to 320x240px at a fixed frame rate of 10fps. A relay server is used to pull data from Kenting to NCHC from where the data streams can be accessed over a high-speed network and be inserted into RBNB.

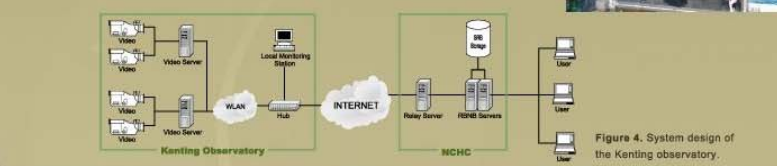


Figure 4. System design of the Kenting observatory.



## Yuan-Yang Lake

Yuan-Yang Lake (YYL) has been established as a natural preserve and has been a long-term ecological study site in Taiwan. Sensors in the area measure factors such as wind speed, temperature and dissolved oxygen in the lake. A 900MHz radio system is used to transmit numerical sensor data to a remote storage point from where it can be inserted into RBNB (Figure 5a). Such automatic transmission mechanisms are necessary to record data during severe weather conditions (Figure 5b).

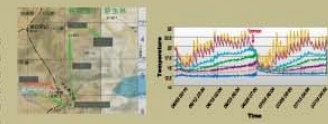


Figure 5. RBNB DataTurbine details.

## More Information :

**Software:**  
RBNB DataTurbine: <http://www.dataturbine.org/>, <http://rtnb.create.com/rbnb/>  
SRB: [http://www.sdsc.edu/srb/index.php/Main\\_Page](http://www.sdsc.edu/srb/index.php/Main_Page)  
RDV: <http://l.nes.org/software/rdv/>

**Networks:**  
PRAGMA: <http://www.pragma-grid.net/>  
GLEON: <http://www.gleon.org/>  
CREON: <http://www.coralreefleon.org/>  
ECOGRID: <http://ecogrid.nchc.org.tw/>

## Results

Our prototype has been successfully deployed and is currently collecting numerical and video data from our sites in Kenting and Yuan Yang Lake at a rate of 7mbps. Three RBNB servers are used in a tree topology with parent-child routing. This strategy allows us to share resources (memory, CPU, network capacity) between nodes while providing a single point of access. In addition, data mirroring is used to minimize the network load when streaming data between NCHC and SDSC via the PRAGMA connection (Figure 3). The archive on SRB is growing with roughly 20GB per day and contains a few weeks of video data (using about 0.7 of totally 4TB os storage capacity). The screenshot in figure 7 shows 8 of our Kenting video streams being displayed simultaneously, in real-time, using Realtime Data Viewer (RDV).

To address larger networks for ECO-science the most important question is scalability. In our test bed more than 70 HD videos has been pushed onto a single RBNB node with total input rate of 129 mbps (Table 1). In a similar test, a number of RDV clients were connected to a parent node from which up to 140 DV video streams were subscribed. Despite a small loss of data rate per channel there were no noticeable loss of image update rate (Figure 6). At that time the outgoing data rate exceeded 100mbps (total rate 217mbps).

Input Type	Resolution	Update Rate (FPS)	No. Sources	Data Rate/Source (kbps)	Total Data Rate (kbps)
High Definition Video	1280x1080	1	70	1870	129000
NTSC DV	720x480	1	60	1950	115000
DV	720x480	1	100	1220	122000
Low Res. Video	600x400	2	200+	220	46000+

Table 1. Max number of sources per child-node (dual 3.4GHz, 2GB RAM, 1000Mbit NIC) depending on video type, resolution and frame rate.

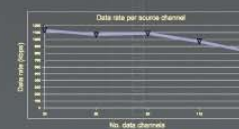


Figure 6. Data rate per channel.



Figure 7. A number of Kenting video streams being displayed simultaneously, in real-time, using RDV.

## Conclusions

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. In all of these systems, streaming data has a central role. The RBNB DataTurbine presents a compelling solution by addressing a core set of cyberinfrastructure requirements common across several environmental observing systems initiatives. Our deployment at Kenting shows that the DataTurbine middleware provides a modular and robust solution to manage streaming video data. In addition, our laboratory tests indicate that it would be fairly straight-forward to set up a DataTurbine-based solution to manage data from observatory orders of magnitude larger than the one in Kenting. Large scale, worldwide networks that include hundreds and thousands of video streams will be the primary focus of our future research.

## Acknowledgments

This work was partially supported by a grant from the Gordon and Betty Moore Foundation and the following NSF grants OCI 0722067 and 0627026. NCHC's research in Kenting is conducted under a cyberinfrastructure initiative called The Knowledge Innovation National Grid (KING) in Taiwan. The coral reef monitoring application is funded by grants TPC-73-92-07309 and TPC-073-94-07310 from Taiwan Power Company.

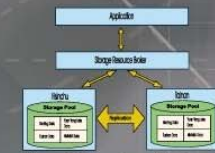


Figure 2. SRB running on two of our main sites: Hsinchu and Taiwan.

**Storage Resource Broker (SRB)** is a data grid middleware developed at SDSC. SRB provides a hierarchical logical namespace to manage large amounts of data across multiple sites. Through SRB, applications are presented with a single file hierarchy to access distributed data. SRB also provides functionalities to replicate data between remote storage pools (Figure 2). In our construction, SRB is used in conjunction with RBNB to provide persistent storage for sensor data. An external agent has been developed to manage internal data transfers (Figure 3).

Figure 3. RBNB DataTurbine is used as base for our streaming data middleware. Sensor data is pushed onto one of NCHC's RBNB servers and immediately becomes available for user applications. Data mirroring is used to minimize the network load when streaming the data to SDSC in the US via the PRAGMA connection. SRB provides persistent storage to the data and an external agent is used to transfer data between RBNB DataTurbine and SRB.

