# Data Management at Kenting's Underwater Ecological Observatory

Ebbe Strandell[‡], Sameer Tilak[†], Hsiu-Mei Chou[‡],
Yao-Tsung Wang[‡], Fang-Pang Lin [‡], Peter Arzberger[†], Tony Fountain[†]
Tung-Yung Fan[#], Rong-Quen Jan[§], and Kwang-Tsao Shao[§]

[†]San Diego Supercomputer Center
University of California, San Diego,
La Jolla, CA 92093-0505

[‡]National Center for High-Performance
Computing,
Number 7, R&D Rd. VI, Hsinchu Science Park,
Hsinchu, Taiwan

[#]National Museum of Marine Biology and
Aquarium
2 Houwan Road, Checheng,
Pingtung, 944, Taiwan

[§]Biodiversity Center at Academia Sinica,
128 Academia Road Sec. 2, Nankang,
Taipei 115 Taiwan

### Abstract

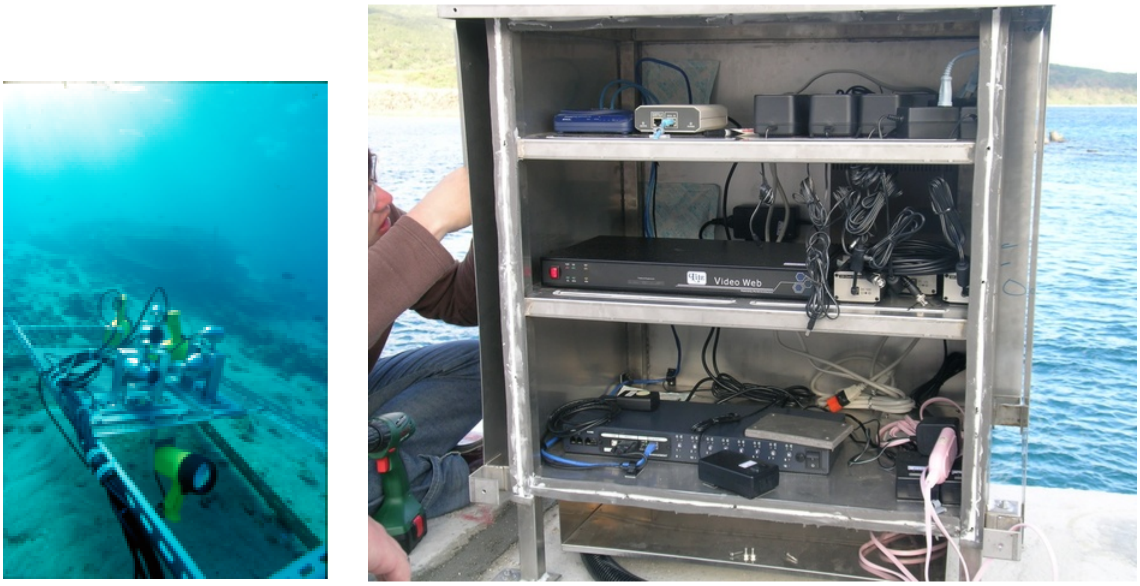
*The management of real-time streaming data in large-scale collaborative applications presents major processing, communication and administrative challenges. To that end, an open-source RBNB DataTurbine provides an excellent basis for developing robust streaming data middleware. The current RBNB DataTurbine streaming data middleware 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, and a comprehensive suite of services for data management, routing, synchronization, monitoring, and visualization. As a part of PRAGMA telescience group, in collaboration with the National Center for High-Performance Computing (NCHC) Taiwan, researchers at the San Diego Supercomputer Center (SDSC) deployed RBNB DataTurbine-based system to acquire data from underwater cameras (in the ocean) at Kenting. More specifically, we describe a system that integrates sensors (underwater video cameras) with computing and storage Grids to create a complete fabric for conducting e-Science. The system is currently used for observation by marine research scientists at the Biodiversity Research Center of Academia Sinica in Taiwan. The described system increased performance and availability of the captured videos and we are, so far, pleased with its results.*

## 1. Introduction

The vision of large-scale sensor-based Observing Systems to address the National Research Councils Grand Challenges for environmental science[1] relies on robust cyberinfrastructure. Environmental challenges include invasive species, infectious diseases, climate and land-use change. Environmental observing science is undergoing dramatic changes through revolutions in information and other technologies including sensor networks and grid computing. Wireless sensor networks capable of measuring environmental variables in an in-situ fashion and at unprecedented temporal and spatial granularities are being deployed to get tremendous insight into complex physical world. Cyberinfrastructure (CI) investments, developments and deployments are being conducted worldwide. They are bringing resources (computers, data storage facilities, equipment for experiments or observations, other researchers) to researchers, and eliminating distance as a roadblock to usage. Computational and data grid technologies are capable of delivering teraflops of CPU cycles and petabytes of storage space required to process and collect the massive amounts of data collected by large-scale sensor networks. Therefore, there is an enormous benefit by Grid-enabling sensors and instruments. The management of real-time streaming data in large-scale collaborative applications presents major processing, communication and administrative challenges. These applications must provide scalable and secure support for data acquisition, instrument and data stream management, and analysis and visualization. Most applications address these issues by building custom systems that are inevitably complex and difficult to support. Extensibility, scalability, and interoperability are often sacrificed under this approach. To address these cyberinfrastructure challenges in a principled manner, we built a solution around an open-source streaming data middleware – RBNB DataTurbine.

## 2. Motivation

The Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) is an open organization in which Pacific Rim institutions collaborate to develop grid-enabled applications and deploy the



(a) Underwater Cameras at Kenting.

(b) Video server.

**Fig. 1:** Kenting setup: Underwater camera deployment and video server

needed infrastructure throughout the Pacific Region to allow data, computing, and other resource sharing (http://www.pragma-grid.net/) [2]. PRAGMA provides an opportunity for member institutions to work together to address applications and infrastructure research of common interest. PRAGMA provides an international testbed for developing and evaluating cyberinfrastructure middleware and is currently extending this testbed to include sensor networks.

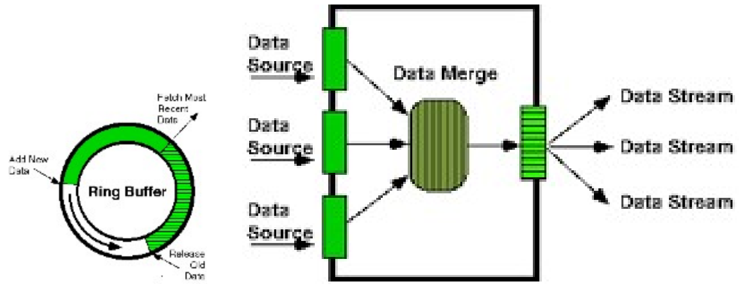
As a part of PRAGMA telescience group, in collaboration with the National Center for High-Performance Computing (NCHC) Taiwan, researchers at the San Diego Supercomputer Center (SDSC) deployed RBNB DataTurbine to acquire data from underwater cameras (in the ocean) at Kenting. Kenting's national park is located on the southernmost tip of Taiwan. The area has a tropical climate with diversified terrain, plentiful wildlife and a dazzling coral reef which is a subject to active research. A system of underwater cameras facilitates marine researchers to monitor the coral reef and the life around it. In this paper we describe our effort to integrate sensors (more specifically the underwater video cameras) with computing and storage Grids to create a complete fabric for conducting e-Science. The system is currently used for observation by marine research scientists at the Biodiversity Research Center of Academia Sinica in Taiwan. To the best of our knowledge, this is the first real-world deployment that acquires data in a network of underwater cameras.

## 3. Background

### *A. RBNB DataTurbine:*

RBNB DataTurbine was developed and is owned by Creare Inc  [3]. After several years of collaboration, executives at Creare Inc. have released RBNB DataTurbine into open-source (under Apache license version 2.0 [4]) in collaboration with the San Diego Supercomputer Center (SDSC). An open-source RBNB DataTurbine is a tremendous asset to the Observing Systems community, and the open-source announcement has generated considerable interest. RBNB DataTurbine provides an excellent basis for developing robust streaming data middleware. The current RBNB DataTurbine streaming data middleware 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, and a comprehensive suite of services for data management, routing, synchronization, monitoring, and visualization. RBNB DataTurbine provides scientists and system users with richer control over data streams, sources, and sinks. RBNB DataTurbine has been tested in a variety of real-world streaming data applications [5], [6], [7]. It facilitates the development of complex distributed streaming data applications, including real-time virtual observatories and telepresence collaboratories [8].

At the core of the RBNB architecture are Ring Buffer Objects (RBO) and Network Bus Objects (NBO). The Ring Buffers function as virtual data servers, managing and archiving data from local and remote clients (Figure 2(a)). The Network Bus Objects provide access to the data for local and remote users, including remote RBNBs(Figure 2(b)). The ability to distribute and mirror data in this manner provides maximum flexibility in terms of managing processing loads, connections and bandwidth utilization. Dynamic Ring Buffers link data source sources with the balance of the RBNB environment. Their critical ”ring buffering” function enables downstream monitors to request contiguous segments of historical or ”most recent” data for applications that require more than



(a) Ring buffer.

(b) Network bus.

**Fig. 2:** RBNB internal details

single-point "current value tables". The Ring Buffers can be configured to provide a combination of RAM memory for high-speed gapless access to a specified amount of most-recent data, plus slower disk memory for playback of historical data. In addition, the Ring Buffers can function as the primary long-term data storage mechanism simply by configuring an appropriately large disk file.

From the perspective of distributed systems, the RBNB DataTurbine is a "black box" from which applications and devices send data and receive data. RBNB DataTurbine handles all data management operations between data sources and sinks, including reliable transport, routing, scheduling, and security. RBNB accomplishes this through the innovative use of memory and file-based ring buffers combined with flexible network objects. Ring buffers are a programmer-configurable mixture of memory and disk, allowing system tuning to meet application-dependent data management requirements. Network bus elements perform data stream multiplexing and routing. These elements combine to support seamless real-time data archiving and distribution over existing local and wide area networks. Ring buffers also connect directly to client applications to provide TiVo-like services including data stream subscription, capture, rewind, and replay. This presents clients with a simple, uniform interface to real-time and historical (playback) data. RBNB's Java implementation language lends it wide platform flexibility.

### B. *Storage Resource Broker:*

The SDSC Storage Resource Broker (SRB) supports shared collections that can be distributed across multiple organizations and heterogeneous storage systems [9]. The SRB can be used as a Data Grid Management System (DGMS) that provides a hierarchical logical namespace to manage the organization of data (usually files). This system also allows for new storage resources to be added dynamically.

### C. *International Communities*

*GLEON:* The Global Lake Ecological Observatory Network (GLEON) 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 [10]. Data from these observatories, including The Long Term Ecological Research (LTER) Network sites, will allow us to better understand key processes such as the effects of climate and landuse change on lake function, the role of episodic events such as typhoons in resetting lake dynamics, and carbon cycling within lakes.

*CREON:* The Coral Reef Environmental Observatory Network (CREON) is a collaborating association of scientists and engineers from around the world striving to design and build marine sensor networks [11]. Benefits of collaborating with CREON are enormous as we attempt to understand the stresses that are shaping the marine world. In particular coral reefs are exhibiting signs of decay around the world as global warming; over fishing and pollution have an impact. Based on our interaction with the GLEON and the CREON communities, we believe that both the communities will greatly benefit by adapting RBNB based solution for managing their real-time data streams. To that end, collaborations with the CLEON and CREON communities is our next step.

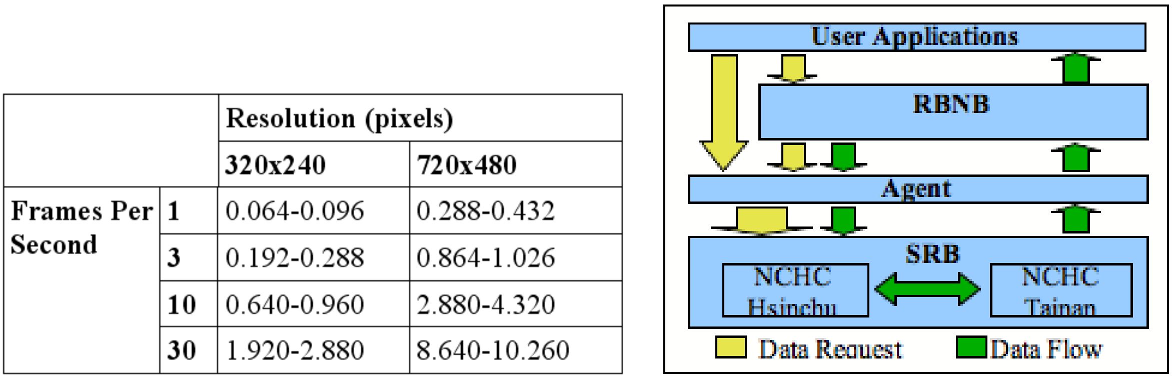
## 4. System Architecture

In this section we first describe the hardware setup and then describe our data management strategy and conclude the section by presenting performance results.

### A. *Hardware Setup*

The ecological observatory in Kenting includes a total of 10 underwater video cameras (Figure 1(a)) located on three different sites inside a fairly large lagoon. Each site is equipped with a video server that converts analog video signals into digital MJPEG video streams. The video servers are installed within a steel casing located on the shore (Figure 1(b)). As demonstrated in (Figure 4) the video streams are transmitted to a local monitoring station using a wireless connection from where they can be accessed from NCHC.

The only available network between the monitoring station and NCHC is a dual 512kbps ADSL Internet connection. Limited bandwidth is a common issue of remote observatories, and it has significant impact on the management of the image data captured in Kenting. Table 3(a) presents estimated bandwidth requirements for each camera at different resolutions and frame



| | | Resolution (pixels) | |
|---|---|---|---|
| | | 320x240 | 720x480 |
| Frames Per Second | 1 | 0.064-0.096 | 0.288-0.432 |
| | 3 | 0.192-0.288 | 0.864-1.026 |
| | 10 | 0.640-0.960 | 2.880-4.320 |
| | 30 | 1.920-2.880 | 8.640-10.260 |

(a) Estimated transfer rate, in mbps, per camera for different configurations.

(b) DataTurbine and SRB data flow within the system.

**Fig. 3:** Kenting setup: Underwater camera hardware details and overall system dataflow

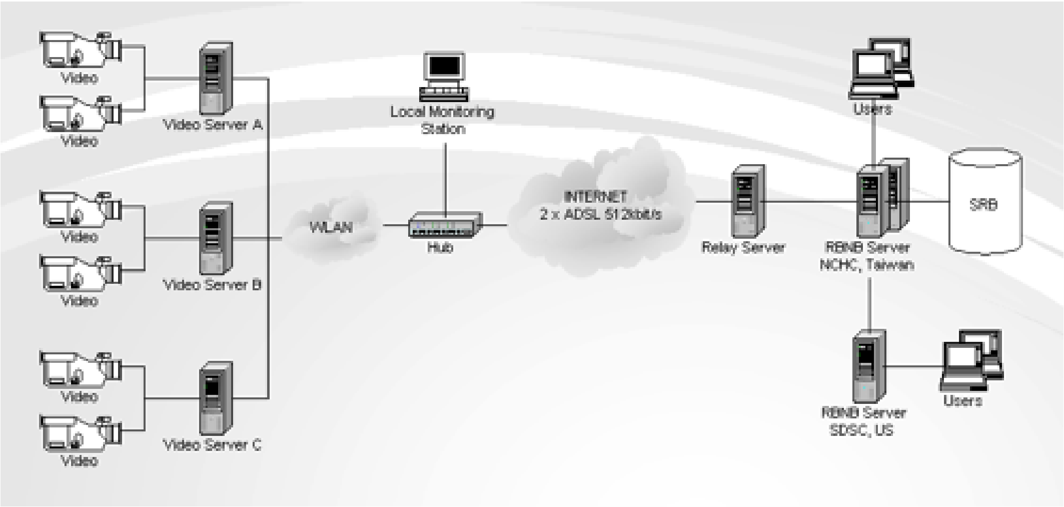


**Fig. 4:** System Architecture and Deployment details

rates. presents estimated bandwidth requirements per camera at different resolution and frame rate. At full resolution and frame rate the bandwidth requirement of 10 videos streams would exceed 100mbps. Clearly, this is not possible in our case. Therefore resolution of each video stream is re-sampled into 320x240 pixels and the frame rate is decreased to 10 fps. The effective transfer rate, however, is no more than 2-3 frames per second.

Support for appropriate and efficient network topologies is a significant consideration when designing a data management and distribution architecture. Via intelligent selection of network topologies, a sensor network designer can improve power efficiency, bandwidth utilization, or system response time. To that end, RBNB DataTurbine can be easily configured to support a broad variety of network topologies. In this deployment, we configured RBNB DataTurbine in a tree topology with parent-child routing. Video servers at Kenting streams the data to a relay server at NCHC. The relay server then converts MJPEG video streams into a format required by the RBNB DataTurbine. At NCHC, the RBNB setup includes two PC's running as dedicated servers in a parent/child relationship,thus utilizing resource sharing while providing a single point of data access. Each RBNB channel is pushed onto one of the servers and immediately becomes accessible for user applications to monitor or analyze its data. In addition, in the current configuration, one RBNB server is located at NCHC and the the server is located at SDSC (Figure 4). Geographical replication of data between SDSC and NCHC servers provides a high reliability in the case of catastrophic failures.

### B. *Historical Data and Persistent Storage*

Given the real-time nature of the experiments, to ensure high performance, each RBNB server is configured to keep a small amount of the most recent data in memory. In addition, historical data of each video channel is provided through a local data archive. This archive holds a predetermined (and configurable) amount of the most recent data, stored on a local disk as a set of binary files. In the current setup the archive holds 4 days of data in files containing 10 minutes of video each. Approximately 15-20 GB of data is acquired per day and in total 110 GB of data is kept in local archives.

SRB is used to tackle the challenge of long-term storage [9]. The SRB system spans over two of NCHC's main sites: Hsinchu and Tainan. To prevent data loss, data inserted at either of these sites is automatically replicated at the other site.

We have implemented an external, Java-based agent to manage data sharing between RBNB and SRB. The agent is responsible of saving data from the local RBNB archive onto SRB along with set of metadata. The metadata makes it possible to search and retrieve data using standard database queries which can include spatial or temporal attributes. The agent is also responsible of handling user requests for historical data stored on SRB. On such requests the agent downloads any data that matches the user query and pushes them onto the RBNB server as a temporary data channel. Currently a few weeks of Kenting data is stored using SRB, occupying roughly 25% of the total storage capacity of 4 TB. An interface to access legacy data stored on SRB is under development.

### *C. Client Software*

We have used RDV [12], an open-source RBNB DataTurbine client program, which we describe now. The Real-time Data Viewer (RDV) provides an interface for viewing and analyzing live or archived time-synchronized data either locally or streamed across a network from a Data Turbine server. RDV is capable of displaying textual and numerical data, still images, and video. The playback rate can be adjusted so data is presented slower or faster than real time to aid in analysis. Figure 5 shows that case where the RBNB server is running with all 10 data channels and a constant archive of about 4 days (or approximately 80GB) archived MJPEG.

### *D. Performance*

To address larger networks for ECO-science the most important question is scalability. The Kenting observatory is a good example of a small, manageable setup which could be one site in a network of tens or maybe hundreds of sites. In our tests we have been able to push more than 70 HD videos (1280x1080, 1 fps, 1.9mbps) onto a single RBNB server (Dual 3.4GHZ, 2GB RAM) with total input rate of 129 mbps. Using DV quality (720x480, 1-2 fps, at 1.2mbps) the corresponding figure was 100 sources and a rate of 122mbps. The data rate of the Kenting video streams is much smaller and one RBNB server should be able to handle a couple of hundred such video streams, at least.

Another test was performed to find the limitations of the parent node (the node through which all the traffic was going through). Using a number of RDV clients we were able to subscribe to 140 DV video streams without any noticeable loss of image update rate. At that time the outgoing data rate exceeded 100mbps (total rate 217mbps).

## 5. Future Work

Future system extensions will include integration of tools for analyzing the data though advanced image processing. This system needs to support both real-time and historical data. We will also collaborate with the GLEON and the CREON community members to help them deploy our system at various GLEON and CREON sites. This will ensure broader applicability of the developed system and would help ecologists around the globe to acquire, transport, and manage large-scale scientific data in real-time manner.

## 6. Conclusion

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, recently open-sourced, 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 RBNB DataTurbine is suitable to manage a larger number of video streams. Test results also indicate that it would be fairly straight-forward to set up a DataTurbine-based solution to manage data from observatories orders of magnitude larger than the one in Kenting. Large scale, world-wide networks that include hundreds and thousands of video streams will be the primary focus of our future research.

## 7. Acknowledgements

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 grant (TPC-73-92-07309) from Taiwan Power Company.

## References

[1] "National research council (nrc) grand challenges in the environmental sciences," 2001.

[2] D. Abramson, A. Lynch, H. Takemiya, Y. Tanimura, S. Date, H. Nakamura, K. Jeong, S. Hwang, J. Zhu, Z. hua Lu, C. Amoreira, K. Baldridge, H.-C. Lee, C.-W. Wang, T. M. Horng-Liang Shih, W. Li, and P. Arzberger, "Deploying scientific applications to the pragma grid testbed: Strategies and lessons," in *Sixth IEEE International Symposium on Cluster Computing and the Grid*, 2006.

[3] "Creare incorporated." [Online]. Available: http://rbnb.creare.com/rbnb/WP/WebWP/rbnbwp.html

[4] "Apache license version 2.0," January 2004. [Online]. Available: http://www.opensource.org/licenses/apache2.0.php

[5] "Nees: The network for earthquake engineering and simulation." [Online]. Available: http://www.nees.org

[6] "The global lake ecological observatory network (gleon)." [Online]. Available: www.gleon.org

[7] "Namma: The nasa african monsoon multidisciplinary analyses program." [Online]. Available: http://namma.msfc.nasa.gov/

[8] P. H. et al., "Using a network ring buffer for science and outreach," 2005, http://www.phfactor.net/sc05/.

[9] C. Baru, R. Moore, A. Rajasekar, and M. Wan, "The sdsc storage resource broker," in *CASCON '98: Proceedings of the 1998 conference of the Centre for Advanced Studies on Collaborative research.* IBM Press, 1998, p. 5.

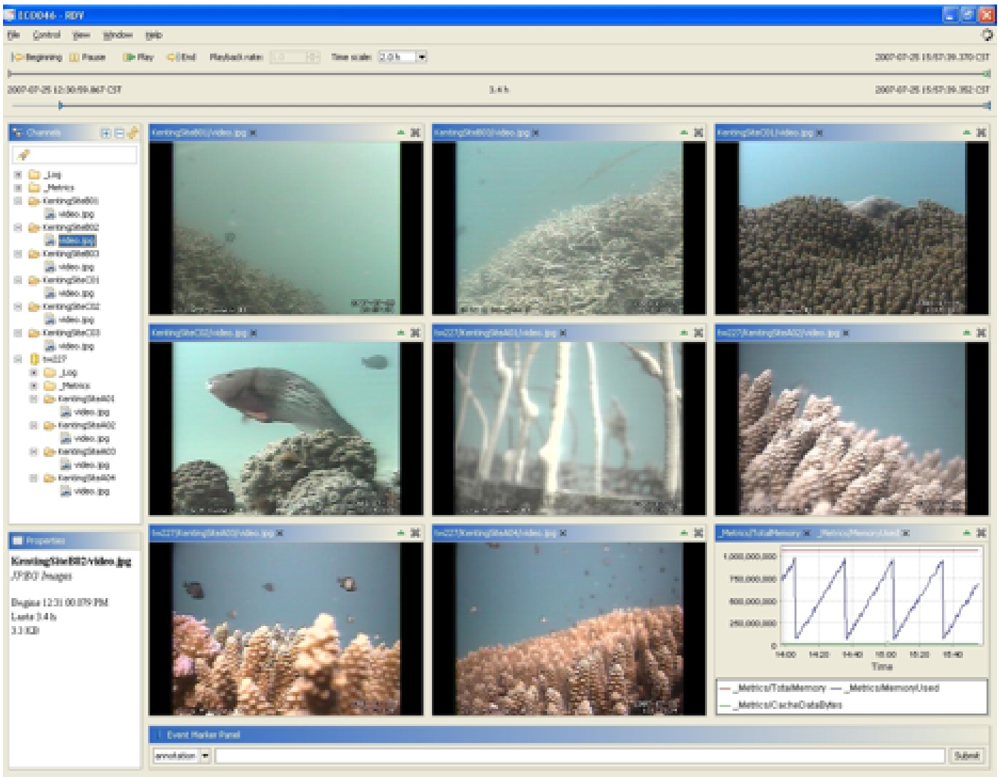


**Fig. 5:** Screen shots of video streams from 8 underwater cameras and RBNB DataTurbine performance metrics (right bottom corner) displayed using RDV

[10] K. Timothy, P. Arzberger, B. Benson, C.-Y. Chiu, K. Chiu, L. Ding, T. Fountain, D. Hamilton, P. Hanson, Y. H. Hu, F.-P. Lin, D. McMullen, S. Tilak, and C. Wu, "Towards a global lake ecological observatory network," in *Publications of the Karelian Institute*, 2006.

[11] "Creon: The coral reef environmental observatory network." [Online]. Available: http://www.coralreefeon.org/

[12] "Rdv (real-time data viewer)." [Online]. Available: http://it.nees.org/software/rdv/