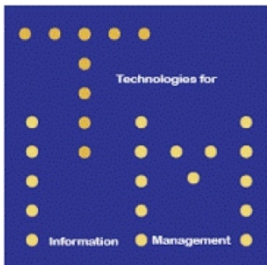
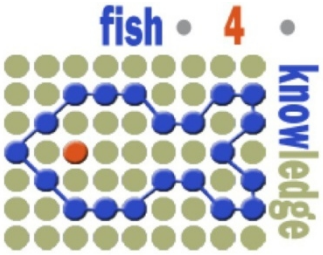




Fish4Knowledge Deliverable D4.1

Video and RDF Store, plus access

Principal Author: H.M Chou
Contributors: NARL
Dissemination: PU

Abstract: WP4.1 has two key objectives: to maintain a sustainable data capturing system which can provide high quality marine ecological observation videos continuously, and to build tera-scale data service platform consisting of repositories for data archiving and retrieving.

To fulfill the requirement of video quality with best signal-to-noise ratio for better image analysis and to balance with resources budget we target video quality to the resolution limit of CCTV cameras. Higher resolution also creates larger data volume; in order to accommodate the massive amounts of observation video captured daily a set of scale-out network-attached storage (NAS) with capacity up to 64TB is acquired and dedicated to the project. The speed of data retrieving is another issue which has significant influence on overall execution efficiency. The best configuration of network linkage, which is the key factor of transferring speed between storage and computation facilities, will be implemented to satisfy high-throughput transferring requirement. A relational table is created to record metadata of data collections, and a query interface is built to make data looking up easier.

Deliverable due: month 12

Table of Contents

1.Introduction

Using state-of-the-art information technology, NARL team established a cyberinfrastructure [1,2] for marine ecological research which integrates geographically distributed sensors, computing power, and storage resources into a uniform and secure platform [3,4]. A geographical map showing relative locations of observatories and NCHC facilities is presented in Figure2.1. The following sections will discuss challenges and resolutions regarding data capturing, storing and retrieving in response to the system for video store and access. For the RDF store as the development evolved, it was decided, with supercomputers involved described in D4.2, to generate in real time without saving for references. As a result, the RDF store will be reduced to minimum with special treatment. Section 2 will be on real time video capturing, which is critical to provide fresh video raw data, along with the archived one, and to well represent the real world coral reef fish ecological system for the locations of interest. Section 3 is then on data archiving, and retrieving over the collected raw data.

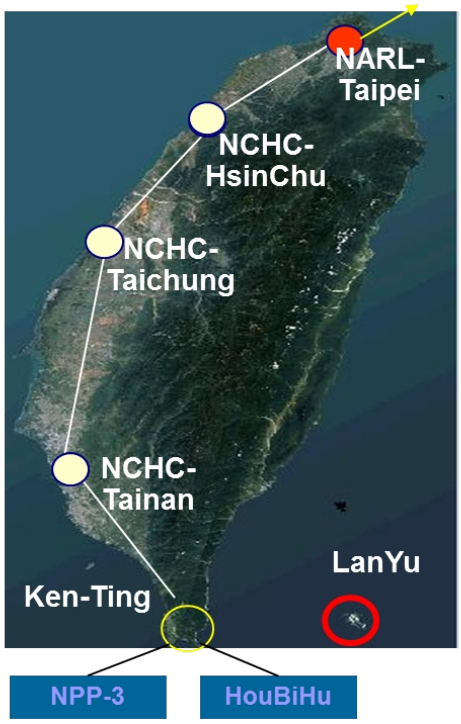
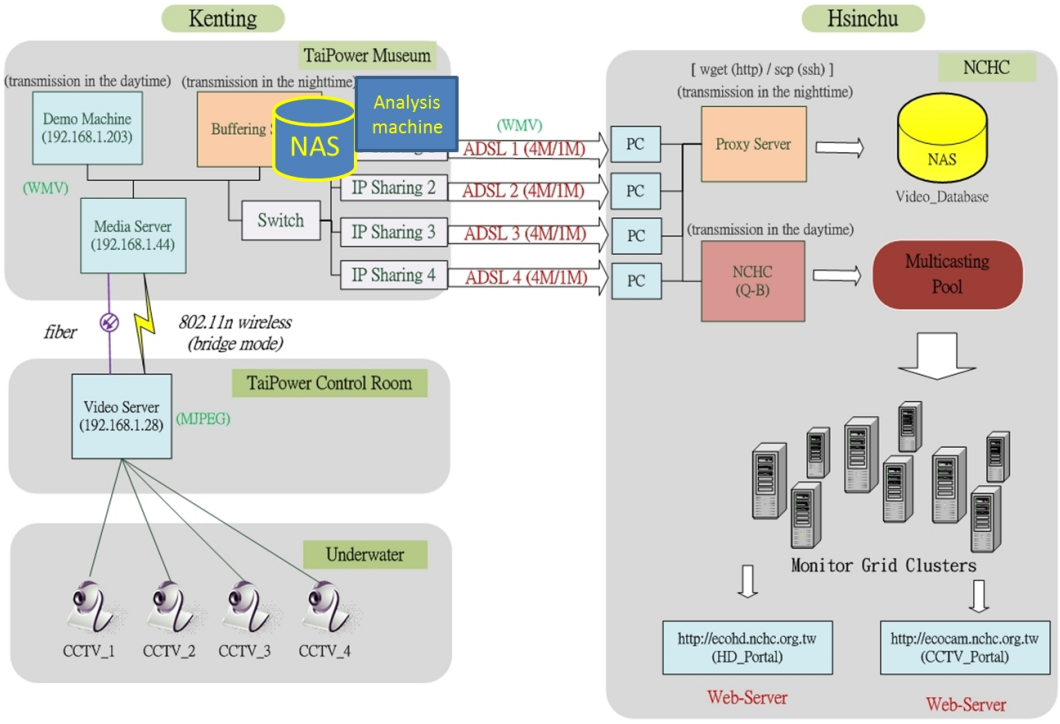


Figure2.1: geographical relation map of distributed sites

2. Video Capturing

The marine ecological observatories are distributed located in three sites in Taiwan, named NPP-3, HoBiHu, and Lanyu [4,5]. Each site is equipped with undersea cameras to continuously capture videos of marine ecosystem during day time and a video server to serve these videos over network. These monitoring cameras generate rich video contents that hold great potential to assist and enable scientific discovery. Given the fact that no well-established information infrastructure is available in the observatories, with only 4Mbps network bandwidth, video clips can’t be transmitted instantly to storage facilities located at NCHC resources center in Taichung where hosts massive storage and computing powers for video analysis. A buffering storage at observatories is deployed to store videos temporarily while they are queuing up for network channel. Besides from storage capacity, the buffer storage server also has enough computing power to do first filtering and bulk compression of the raw video data before transmission across the network. A schematic architecture of the video data capturing/transferring infrastructure is shown in Fugure2.2.



Fugure2.2: A schematic architecture of the video data capturing and transferring infrastructure. Takes NPP-3 site as an example.

The creation and analysis of high-quality data are core elements of scientific discovery. Yet, higher-quality data place heavier demands on storage capacity and network bandwidth. There are continues struggles for balancing constraints of resources budget to meet quality requirements best. As a solution, we plan to develop an efficient data compression scheme will be developed to suit data quality requirements under given resources constraints.

3. Video Archiving and Retrieving

3.1 Massive Data Storage System

The undersea observation project has accumulated 11TB video data since the date it started, and it's roughly adding 6TB more to the collection every month as we pushed to the upper limit of resolution. It has placed heavy demand on storage capacities. Table3.1 shows estimated data size in two different resolution scales.

Resolution	320x240 @ 8 fps	640x480 @ 24 fps
File Size (10 mins)	≈ 17MB	120MB ~ 300MB
Capturing Period	6:00 AM ~ 18:00 PM	6:00 AM ~ 18: 00PM
Data Size(per day/per camera)	17 MB * 6 * 13 = 1326 MB ≈ 1.3 GB	300 MB * 6 * 13 = 23400 MB ≈ 23GB
Data Size /Per Month	1.3 GB * 9 (cameras) * 30 (days) ≈351 GB	23 GB * 9 (cameras) * 30 (days) ≈ 6.21 TB
Data Size Generated Per Year	1.3 GB * 9 (cameras) * 365 (days) ≈4.3 TB	23 GB * 9 (cameras) * 365 (days) ≈ 75.6 TB

Table3.1: estimated data size in different resolution scale

The challenge to a scale-out storage system is to allow the user to grow storage resources in-line with data center demands as project needs change over time. This means that the storage system must expand but still maintain functionality and performance as it grows.

Beyond capacity, bandwidth is another resource that must be available to meet the project's intense performance requirement of data services. Without enough I/O bandwidth, connected servers and users can become bottlenecked, requiring sophisticated storage tuning to maintain reasonable performance.

To address the infrastructure challenges raised in this deliverable task a scale-out massive storage system is built to meet the requirements of storage capacity and data transfer performance. The system contains three levels of storage components to accommodate data throughout different stages of lifecycle: a data buffer up to 8TB at observatories to hold live data temporarily before they are transferred back to NCHC, a midterm storage pool up to 50TB at NCHC to store data in

analysis stage, and a set of long term storage up to 30TB at NCHC to preserve critical data. Both data buffer and midterm storage pool are in place for service, and we expect to add long term storage in the third quarter of 2012.

Figure3.1 shows the designed architecture of a scale-out infrastructure for midterm and long term storage. With this design, resource components can be dynamically added to the system to balance budget and performance requirements. As different resource components are added to the system the aggregate of each of these components in the system is upgraded simultaneously. For example, when there is demand for more storage capacity, sets of Network Attached Storage (NAS) can be added to NAS array and the capacity is extended correspondingly with minimum administrative efforts.

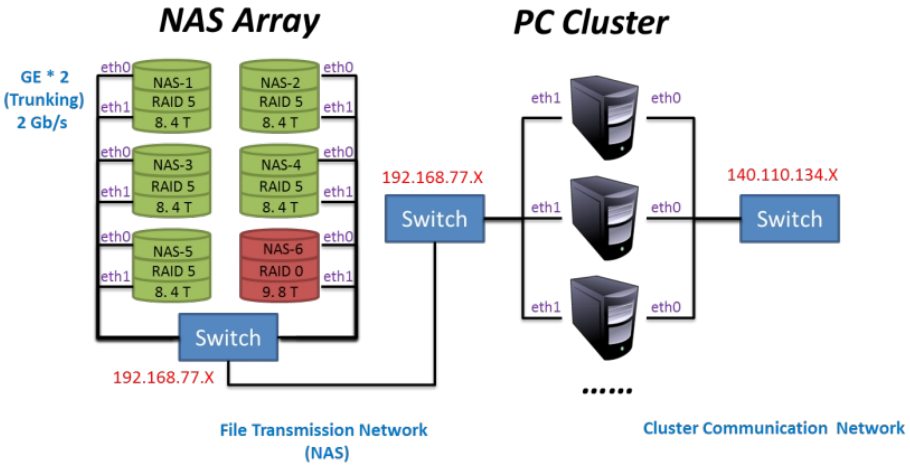


Figure 3.1: the architecture of scale-out infrastructure. It can be expanded by adding more nodes to the system. Cluster of Network Attached Storage (NAS Arrary) for storage capacity, PC for processing power, and network switch for I/O bandwidth.

3.2 Data Management

With the enormous amount of video data generated from observation cameras continuously, it requires efficient and effective mechanisms to store, access and retrieve these data. The first problem is the efficient

organization of raw video data. There has to be proper consistency in data in the sense that data are to be stored in a standard format for access and retrieval. A relational table with metadata tags of each video clip is created to manage video records, and a faster video query scheme is also required. Figure3.2 shows different design of query schemes. The prototype implementation of the dynamic query model demonstrated great improvement of efficiency when comparing with the query scheme without cache layers.

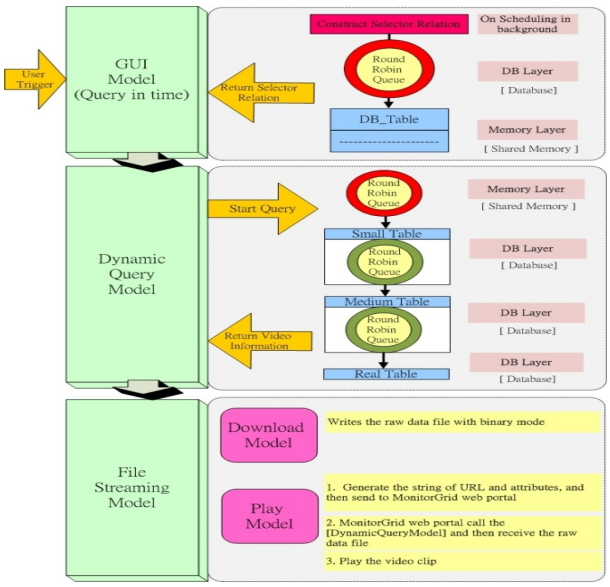


Figure3.2: cartoons of different query scheme. The first block shows a query scheme without cache layers, in this design frequently accessed data are stored in memory to shorten retrieval time, the overall performance is limited by memory capacity. The second block shows a query scheme with cache layers, in this design two extra cache tables, which are stored in faster drive, are added. With the memory plus cache layers, it can boost retrieval performance.

Through the query scheme, sorting by date, a list of available clips will be returned in xml format, including useful metadata of the clip, ex. camera id, resolution, size, etc. A short example of the XML returned from the query:

```
<MVList>
    <MVObject>

<UUID>b7e73c98c1300f108ccf24c61b9595f5#201103030100</UUID
>
        <capture_time>2011-03-03 01:00:00</capture_time>
        <camid>LanYu_Cam:1</camid>
        <resolution>640x480@8</resolution>
         <video_length>00:12:02</video_engths>
        <full_path>http://gad240.nchc.org.tw/EcoData/_NEW/Site_
        B/2011/03/03/01/00/Video1_2011-03-03-01-
        00.flv</full_path>
        <fdt_status>1</fdt_status>
    </MVObject>
</MVList>
```

In WP5, fish detection component extracts various features of video data from low-level features like shape, color, texture, and spatial relations and stored efficiently in relational database. As for the RDF store proposed in WP5, for the feasibility of web accessing, it has been revised to serving detection data in RDF dynamically, e.g. on the fly by an off the shelf RDB2RDF wrapper (currently D2R) that will just use the same SQL API as the rest of the system. A R2D server will be up and running 3 months after a stable RDB is in service.

3.3 Interface to Compute Components

A 3-tier architecture design of infrastructure, presents in Figure3.3, is adopted. With this design the system can be benefitted from the flexibility of scaling-out – the capacity is increased by adding more resources to different tiers. The issue here is data needs to be

transferred to computation tier in time for analysis. 10 Ggigabit Ethernet (10GE) network interfaces will be deployed for interconnections of storage and computation tiers to target best transmission rate. And also a high throughput data transfer scheme will be investigated.

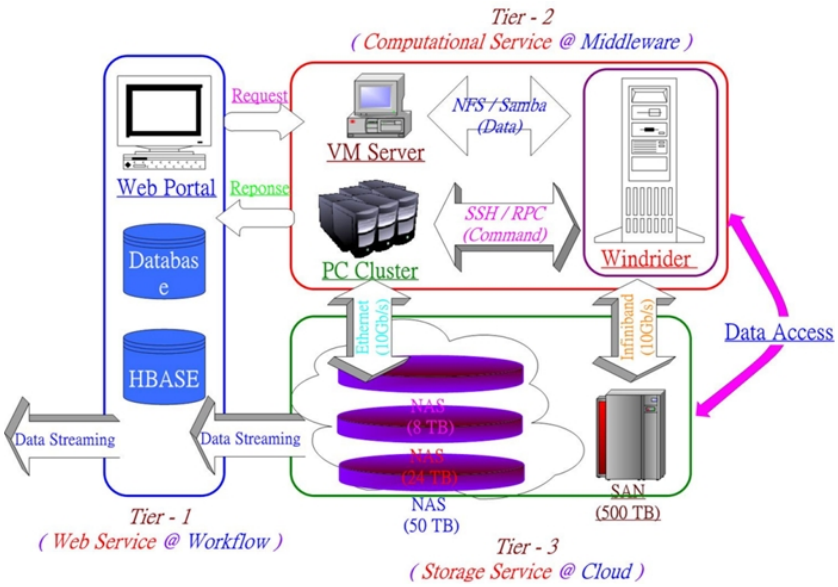


Figure3.3: The 3-tier architecture design of infrastructure. Tier-1 is

the service portal component which links users with computatioanl

service component, the Tier-2, and storage service component,

the Tier-3. Users submit request to the system through Tier-1.

Based on the request Tier-2 responds to it by initiate a processes

workflow, and data will be fetched from Tier-3 if a data analysis

process is included in the workflow.

4. Conclusions

Brief summarization of current developments:

- To fit the requirement of best signal-to-noise ratio for better image analysis video resolution is pushed to the limit of 640x480 at 25fps.
- A hierarchical storage system with capacity up to 64TB has been built to accommodate videos collections, and expecting to add 30TB tape storage to the system in third quarter of 2012.
- The best configuration of network linkage, which is the major factor of transferring speed between storage and computation facilities, will be implemented.
- A relational table is created to record metadata of video collection, and a layered query scheme is implemented to make data looking up effectively.

References

[1] Daniel Atkins, et al “Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.” National Science Foundation. January 15, 2003.

[2] Tony Hey, et al “Cyberinfrastructure for e-Science”, Science 308, 817 (2005).

[3] Whey-Fone Tsai, Weicheng Huang, Fang-Pang Lin, Bonita Hung, Yao-Tsung Wang, Steven Shiau, Shyi-Ching Lin, Chang-Huain Hsieh, His-En Yu, Li-Lun Pan and Chien-Lin Huang “The Human-Centered Cyberinfratructure for Scientific and Engineering Grid Applications”, J. of the Chinese Institute of Engineers, Vol. 31, No.7, pp. 1127-1139, 2008.

[4] Hsiu-Mei Chou, Yi-Haur Shiau, Shi-Wei Lo, Sun-In Lin, Fang-Pang Lin, Chia-Chen Kuo, Chuan-Lin Lai “A Real-Time Ecological Observation Video Streaming System Based on Grid Architecture” HPC Asia 2009, Kaohsiung, Taiwan.

[5] Nai-cheng LIN, Tung-Yung Fan, Fang-Pang Lin, Kwang-Tsao Shao, Tzong-Hwa Sheen, “Monitoring of of Coral Reefs at the Intake Inlet and Outlet Bay of the 3rd Nuclear Power Plant in southern Taiwan”, Annual Meeting of The Fisheries Society of Taiwan, 19-20, December, 2009.

