National Science Foundation Ocean Observatories Initiative (OOI) Cyberinfrastructure White Paper

Ivan Rodero and Manish Parashar

Rutgers Discovery Informatics Institute (RDI^2), {irodero, parashar}@rutgers.edu

INTRODUCTION

The NSF Ocean Observatories Initiative (OOI) is a networked ocean research observatory with arrays of instrumented water column moorings and buoys, profilers, gliders and autonomous underwater vehicles within different open ocean and coastal regions. OOI infrastructure also includes a cabled array of instrumented seafloor platforms and water column moorings on the Juan de Fuca tectonic plate. This networked system of instruments, moored and mobile platforms, and arrays will provide ocean scientists, educators and the public the means to collect sustained, time-series data sets that will enable examination of complex, interlinked physical, chemical, biological, and geological processes operating throughout the coastal regions and open ocean.

The seven arrays built and deployed during construction support the core set of OOI multidisciplinary scientific instruments that are integrated into a networked software system that will process, distribute, and store all acquired data. The OOI has been built with an expectation of operation for 25 years. This unprecedented and diverse data flow is coming from 89 platforms carrying over 830 instruments which provide over 100,000 scientific and engineering data products.

The OOI is funded by the National Science Foundation and is managed and coordinated by the OOI Program Office at the Consortium for Ocean Leadership (COL). Implementing organizations, subcontractors to COL, are responsible for construction and development of the different components of the program. Woods Hole Oceanographic Institution (WHOI) is responsible for the Coastal Pioneer Array and the four Global Arrays, including all associated vehicles. Oregon State University (OSU) is responsible for the Coastal Endurance Array. The University of Washington (UW) is responsible for cabled seafloor systems and moorings. Rutgers, The State University of New Jersey, is implementing the Cyberinfrastructure (CI) component. The OOI data evaluation and education and public engagement team is co-located with the Cyberinfrastructure group at Rutgers University.

OOI CYBER-INFRASTRUCTURE SERVICES

The primary functions of the OOI CI are data acquisition/collection, storage, processing and delivery. The overall architecture of the OOI CI network is shown in Figure 1.

(a) Data Collection and Transmission to the OOI CI: Data is gathered by both cabled and un-cabled (wireless) instruments located across multiple research stations in the Pacific and Atlantic oceans. Once acquired, the raw data (consisting mostly of tables of raw instrument values – counts, volts, etc.) are transmitted to one of three operations centers: Pacific City, directly connected via fiber optic cable to all cabled instruments in the Cabled Array; OSU, an Operational Management Center (OMC) responsible for all un-cabled instrument data on the Pacific coast; and WHOI, the OMC for Atlantic coast-based un-cabled instrument data. The data from the operations centers is transferred to the OOI CI for processing, storage and dissemination.

(b) Data Management, Storage, and Processing: Two primary CI centers operated by the Rutgers Discovery Informatics Institute (RDI^2) are dedicated to OOI data management: the West Coast CI in Portland, OR, and the East Coast CI, at Rutgers University. While data from the Cabled Array components are initially received at the Shore Station in Washington, it is the East Coast CI that houses the primary computing servers, data storage and backup, and front-facing CI portal access point, all of which are then mirrored to the West Coast CI over a high-bandwidth Internet2 network link provisioned by MAGPI (Mid-Atlantic GigaPOP in Philadelphia) on the east coast and PNWGP (Pacific-Northwest GigaPOP) on the west coast. The data stores at the OMCs at OSU and WHOI are continuously synchronized with the data repositories located at the East and West Coast CI sites.

(c) Data Safety & Integrity: Data safety and protection is ensured in two ways: data security and data integrity. Data security is addressed through the use of a robust and resilient network architecture that employs redundant, highly available next-generation firewalls along with secure virtual private networks. Data integrity is managed through a robust and resilient information life-cycle management architecture.

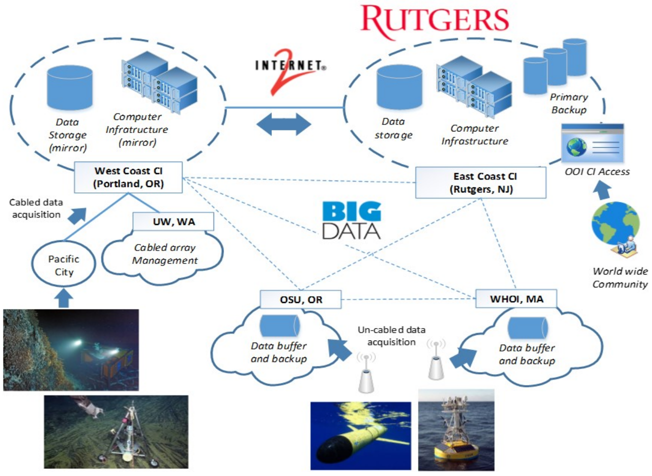

Fig. 1. OOI CI Network Architecture

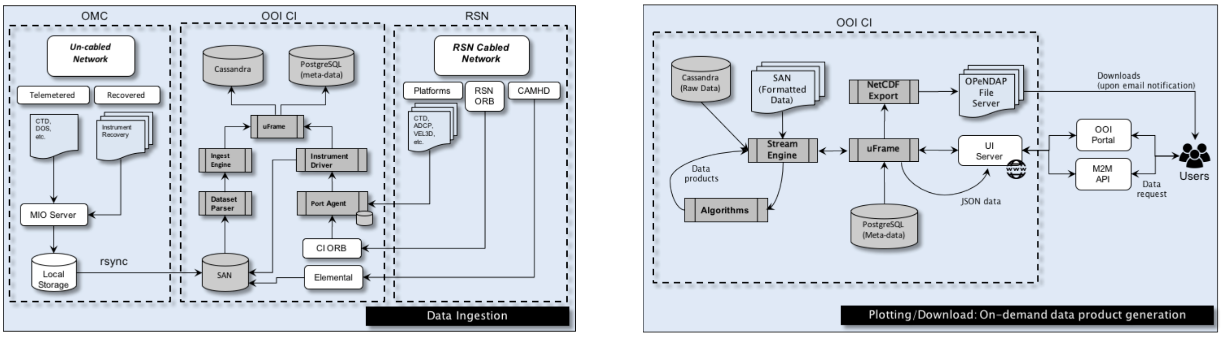

Fig. 2. UFrame-based OOINet software data workflow (left: data ingestion, right: data plotting/download)

(d) Public Data Access: The OOI CI software ecosystem (OOINet) employs the uFrame software framework that processes the raw data and presents it in visually meaningful and comprehensible ways in response to user queries, which is accessible over the Internet through the CI web-based portal access point. A machine-to-machine (M2M) API provides programmatic access to OOINet through a RESTful API. In addition to the portal and API, OOI CI provides the following data delivery methods: (1) *THREDDS Data Server*: delivers data products requested through the CI portal (i.e., generated asynchronously); (2) *Raw Data Archive*: delivers data as they are received directly from the instrument, in instrument-specific format, and (3) *Alfresco Server*: provide cruise data, including shipboard observations. OOI CI software ecosystem permits 24/7 connectivity to bring sustained ocean observing data to a user any time, any place. Anyone with an Internet connection can create an account or use CILogon and access OOI data.

DESING AND IMPLEMENTATION ISSUES

The OOI CI design and implementation principles are based on industry best practises for the different aspects of the CI. The approach is based on a decentralized but coordinated architecture, which is driven by requirements, e.g., data storage capabilities, system load, security, etc.

(a) Redundancy and resiliency: The OOI CI is a mirrored infrastructure for high availability, disaster recovery and business continuity. It implements a resilient information life-cycle management architecture that integrates redundant enterprise storage area network (disk-based) and a robotic library (tape-based). Redundancy is implemented at different layers, for example, an enterprise-level storage network of multiple hard drives managed by an intelligent device manager, reduces the data footprint by reducing data duplication while maintaining data integrity and access performance through storage redundancy, and tape storage, a "last tier" storage that is not dependent on power or cooling, supports longer-term backup and archiving, disaster recovery, and data transport.

(c) Service-oriented Architecture: The core of the OOI CI software ecosystem (Uframe-based OOINet, see Figure 2) is based on a service oriented architecture, a set of data dataset, instrument, platform drivers and data product algorithms, which plug in to the uFrame framework. Uframe-based OOINet uses latest generation technologies for big management data such as Apache Cassandra, which is a state-of-the-art, scalable and highly available distributed database management system designed to handle large amounts of data. Uframe-based OOINet services are exposed through a RESTful API and are available as the M2M interface for external access through a secure endpoint. The use of a well-defined API based on standard protocols enables other systems to interface and interact with OOI CI programmatically.

(c) Cyber-security: The system is based on a multi-tier security approach with dedicated and redundant (highly available) appliances at the CI perimeter. The OOI CI implementation supports encryption of traffic, network traffic segregation, multi-layer traffic filtering, multi-layer access control and comprehensive monitoring. Further, data delivery to external users is implemented through dedicated and distinct storage appliances (i.e., physical and logical isolation from core storage infrastructure) In addition to implementing industry best practices, the OOI CI cyber-security effort includes a comprehensive cyber-security program based on engagement with the NSF Center for Trustworthy Scientific Cyber-Infrastructure. This program encompasses a set of policies and procedures. Regular vulnerability scans/audits (internally and externally) are also performed to the OOI CI.

CONCLUSION

OOI CI has initiated its operational phase and data (including science, engineering and data products) flowing from those instruments is freely available to users. The OOI CI portal provides all data, metadata and data processed via conventional algorithms or direct retrieval from OOI storage or data archives. Data quality and data management will utilize generally accepted protocols, factory calibrations and at sea calibration procedures.

During its early operation (1.5 years), OOI community has been growing every day and is made up of a diverse set of users from 180 different organizations from around the world. At least 500 people has already registered on the OOI Data Portal, which has over 3,000 unique visitors each month[1].

[1] OOI is a NSF-funded effort and involves teams from Consortium for Ocean Leadership, Woods Hole Oceanographic Institution, Oregon State University, University of Washington, Rutgers University, and Raytheon. This document summarizes the contributions from these teams. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.