VDC
VIRTUAL DATA COLLABORATORY

Towards a Smart Streaming-Based Cyber-Infrastructure Framework for Scientific Data Observatories

Yubo Qin, Anthony Simonet, Ivan Rodero

Rutgers Discovery Informatics Institute (RDI²)

NSF

1766

RUTGERS
Discovery Informatics Institute

The Future of Large Facilities: Data and Science

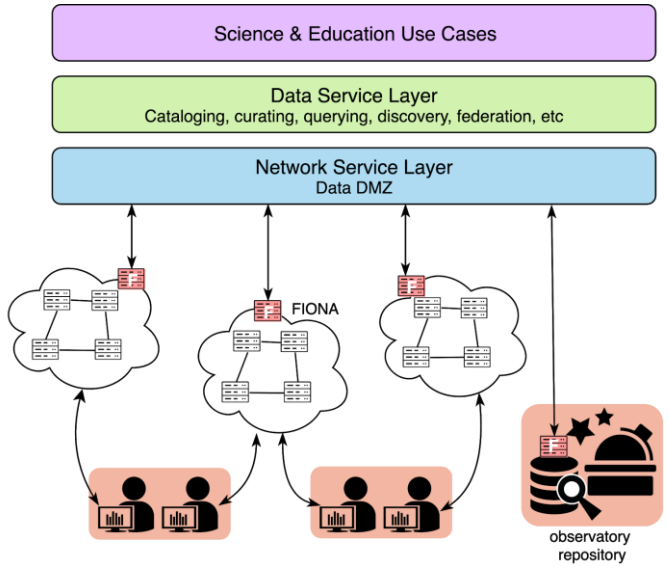
Data and services provided by large-scale instruments and observatories have become **important enablers** of scientific discoveries.

This research explores how an understanding of the **science usage patterns** coupled with emerging cyber-infrastructures solutions can improve the performance, usability and science impact of data and services provided by facilities.

SYSTEM PLANE	Physical infrastructure (compute, storage, network, FIONAS), operating system and networking. Virtual infrastructure management: VMs, containers, etc.
DATA PLANE	Cross-repository data indexing and discovery, provenance records and other data-related services.
KNOWLEDGE PLANE	Data analytics, cross-repository data fusion, in-transit processing.
USER PLANE	Productivity tools, streaming-based interfaces, advanced caching and prefetching strategies.

Our vision is broken down into **4 planes** that each address specific needs.

Each **planes** can be mapped directly to the VDC architecture.



VDC
VIRTUAL DATA COLLABORATORY

Work in Progress: a Smart Cache with Prefetching

A framework to improve the performance of data delivery on top of the VDC.

The framework will implement a cache with placement and **prefetching** guided by the knowledge of previous **data queries**, leveraging data analytics and Machine Learning.
Prefetching will be performed when a high level of certainty is achieved and offer **suggestions** in other cases (like Amazon does for product recommendations).

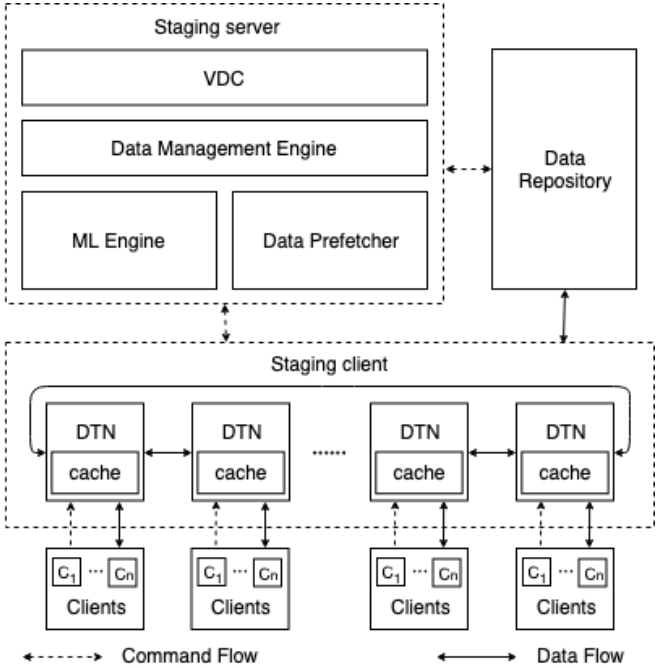
Smart cache
Prefetching model based on
- associative rule mining
- historical records

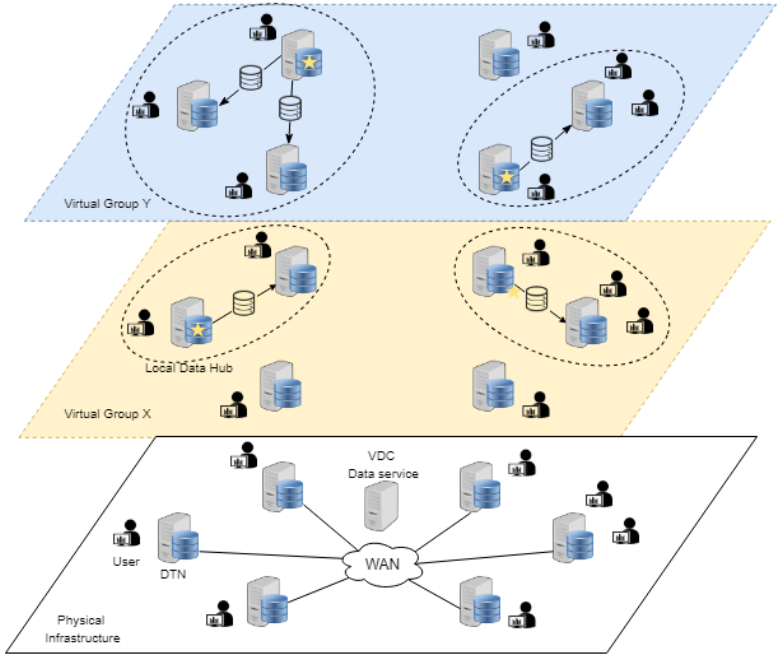
Cache placement
- Virtual group
- Local data hub

Analysis on Ocean Observatory Initiative (OOI)

User classification
- Interactive (humans)
- Regular programs
- Faulty programs
- Real-time programs

Request correlation
- Data objects
- Location
- Time

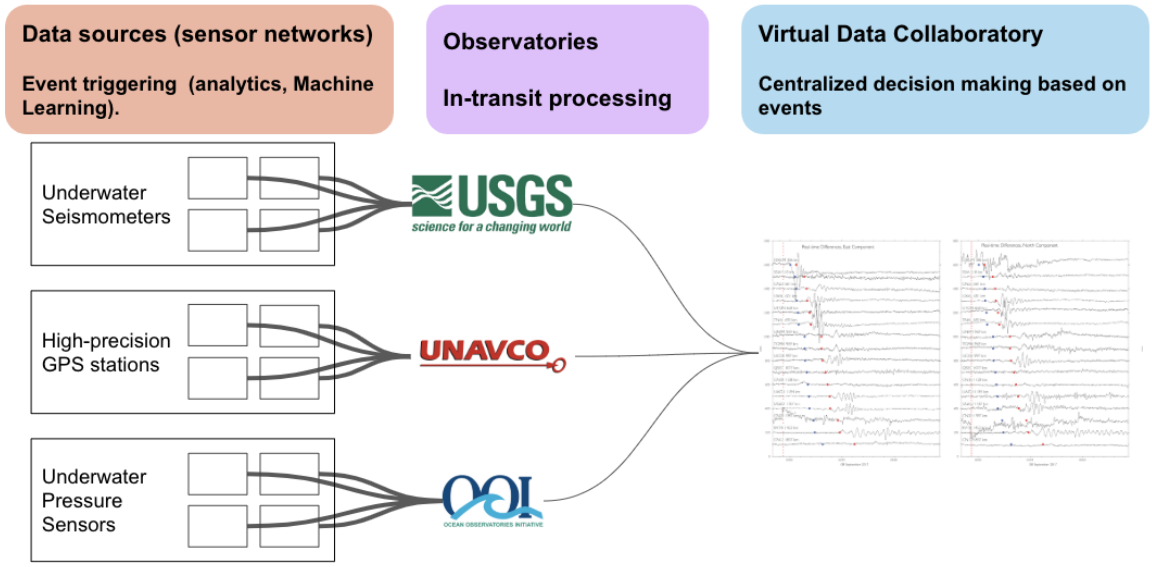




Use-case: Tsunami Early Warning

Increase precision and delay for Tsunami warning by analyzing multiple geo-graphically-distributed data sources simultaneously, in collaboration with UNAVCO.

To issue Tsunami Early Warnings, earthquakes must first be characterized (magnitude, location, speed of displacement, etc.).
Seismometers are good for the ***smaller* earthquakes** (< 6.5), **high-precision GPS** are good for ***larger* earthquakes**.
Goal: combining multiple data sources to improve the precision and delay to issue warnings by covering the **whole spectrum** of events.



# 60s MTS	GPS (# Events)	Seismic (# Events)
Magnitude < 5	7,718 (170)	1,038 (349)
$5 \leq$ Magnitude < 6	3,859 (85)	None
$6 \leq$ Magnitude < 7	991 (4)	266 (4)
$7 \leq$ Magnitude < 8	432 (6)	249 (6)
Magnitude > 8	265 (4)	133 (4)
Total	13,265 (269)	1,686 (363)