Funded by the National Science Foundation, Grant #2127548

OOI Approach to the Cloud

By Jeffrey Glatstein

The March Toward the Clouds: MF Perspectives

March 2, 2022
CI Compass Cyberinfrastructure for NSF Major Facilities Workshop

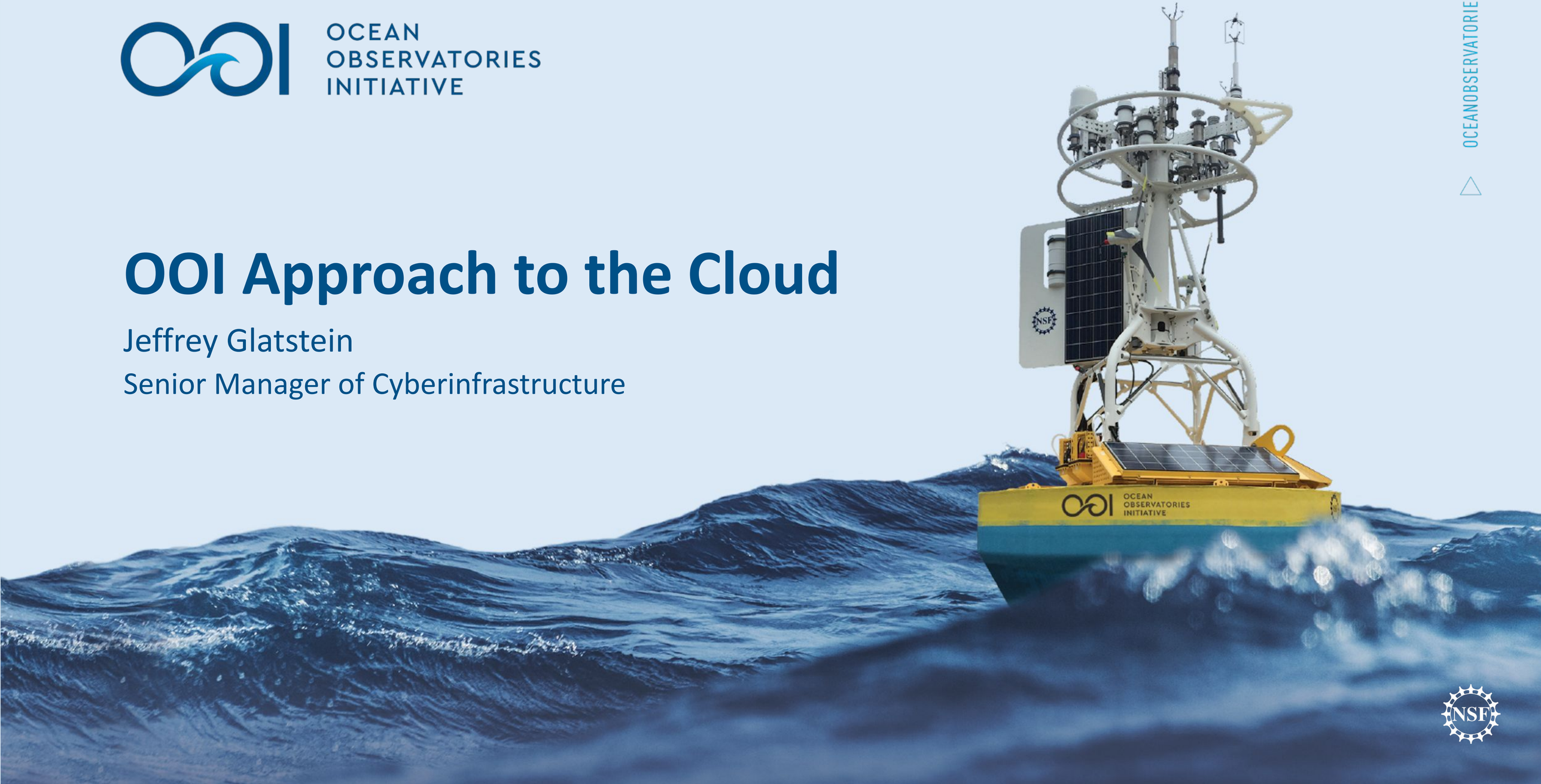
USC Viterbi School of Engineering Information Sciences Institute
INDIANA UNIVERSITY
UNIVERSITY OF NOTRE DAME
THE UNIVERSITY OF UTAH
TEXAS TECH UNIVERSITY
renci

OCEAN OBSERVATORIES INITIATIVE

OCEANOBSERVATORIES.ORG

OOI Approach to the Cloud

Jeffrey Glatstein
Senior Manager of Cyberinfrastructure



NSF

OOI Vision

- Real-time data from more than 800 instruments to enable research and education in Earth & Ocean sciences
- Marine arrays at three scales served by a common cyberinfrastructure
- Data freely available online
- 25-year lifetime
- Operated and maintained by WHOI, UW and OSU

- Sponsored by National Science Foundation:
  - Section Heads: Bauke Houtman & Lisa Clough
  - Division of Acquisition & Cooperative Support: Anna-Lee Misiano
  - Large Facilities Office: Ryszard Kaczmarek

OCEAN OBSERVATORIES INITIATIVE

NSF

OOI Data By The Numbers

- Data Collected
  - Cassandra 28 Nodes totaling 25TB
  - Postgres 350GB
  - 112 billion rows of numerical data to date with new data ingress every second
  - Raw data 765TB to date with expected growth rate of doubling every 3 years
  - HD video, digital still pictures, bio-acoustic sonar, Hydrophone acoustic sampling
- Data Delivered
  - 25 million data requests per month
  - These include 5.6 million data requests with 1000 rows or greater of data returned
  - External real-time systems interrogate OOI API every second to every 30 seconds

OCEANOBSERVATORIES.ORG

Initial Cloud Questions and Answers

Why should OOI consider using Cloud based processing and/or data access?

- Articles and anecdotal information suggests that Cloud processing can provide processing power, infrastructure flexibility and cost savings for an institution. It has become a popular method of running CI infrastructure across many domains. For these reasons, it is incumbent upon CI Management to explore it's features and any benefits that may be derived from use.

Can OOI migrate all processing to the cloud effectively?

- Research has shown that the current OOI infrastructure architecture, data process flow and data use patterns are barriers to a large scale move to the cloud due to the unpredictability of data egress costs as well as overall infrastructure deployment costs.

Can OOI prepare or better position the program for a migration to the Cloud?

- The current thought is yes – plan on implementing a hybrid solution which keeps the bulk of raw data local for user access and reduces the data processing footprint for migration to the Cloud.

How can OOI best utilize the Cloud **now** in a cost-effective manner given the answers to the above preliminary questions?

Note: Cloud Storage is utilized for disaster recovery but is not accessible by end users

Strategy Moving Forward

Levelling the “playing field” for using OOI data is the primary value the Cloud can provide for our users

- Eliminates the need to download data in order to interact with it, thereby removing the barrier of little to no access to highspeed internet connectivity
- Moves processing from local to Cloud, eliminating barrier of older and/or underperforming Researcher owned equipment
- Additional benefit of making access to large datasets more effective

…but the costs are still unpredictable and potentially high.

OCEANOBSERVATORIES.ORG

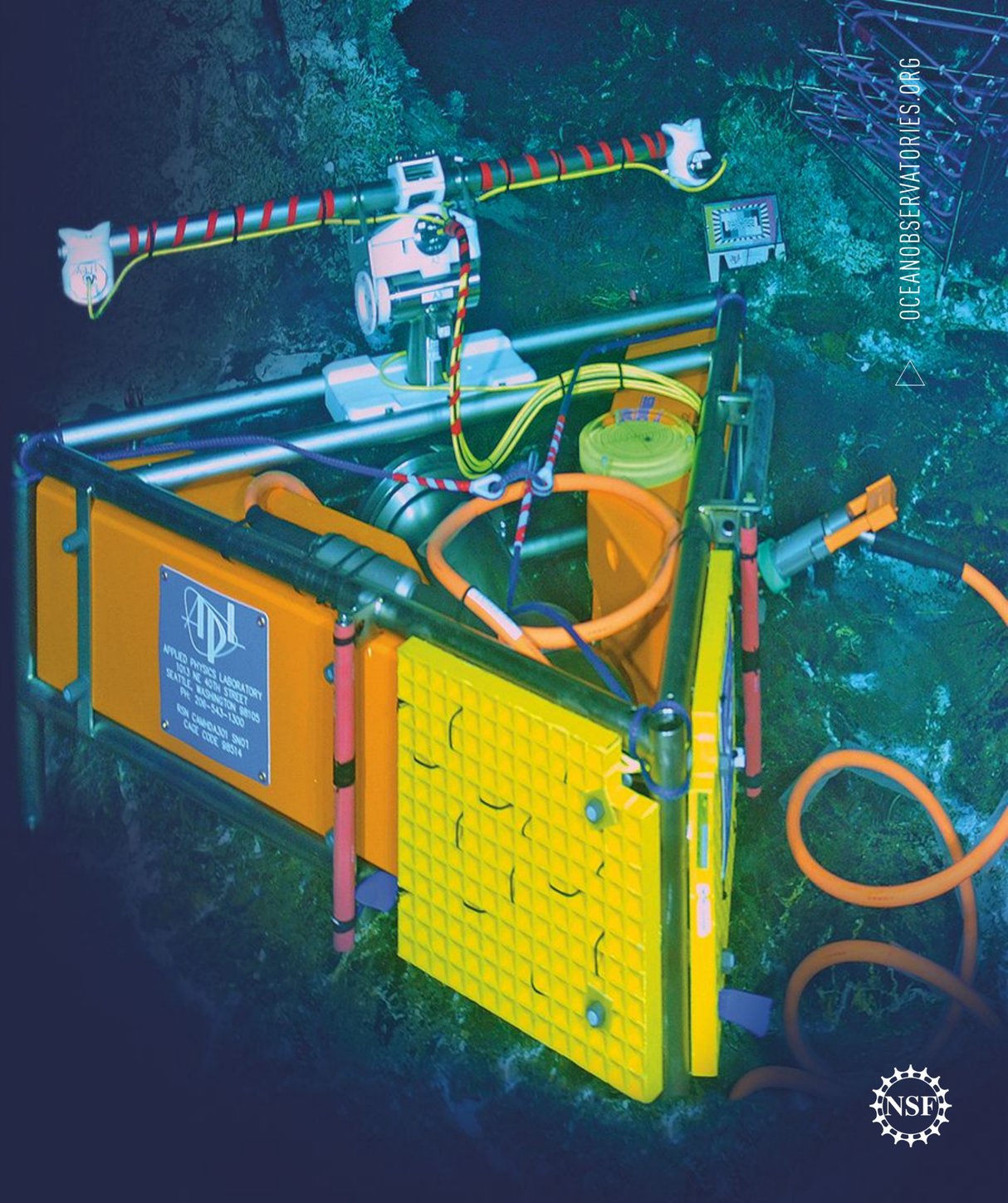
Proving Out The Strategy Through Proof of Concept(s)

Use small proof of concept (POC) projects to gather information that either proves or eliminates implementation ideas

- Build an OOI local environment to emulate a Cloud infrastructure without the budgetary unpredictability/risk of a traditional cloud provider
- Implement a cloud friendly architecture such as Jupyter Hub or Pangeo
- Make a subset of OOI data available in “cloud friendly” formats such as NetCDF or ZARR
- Organize data in a searchable format such as STAC (Spatio Temporal Asset Catalog)
- Implement software (Dell Live Optics) that measures on-prem system use for estimating potential Cloud considerations and costs

First POC implemented - Jupyter notebook on a local “Cloud” server, attached to local precomputed files utilizing STAC for data discovery was considered successful

OCEAN
OBSERVATORIES
INITIATIVE



OCEANOBSERVATORIES.ORG

NSF