The Evolution Of Bits And Bottlenecks In A Scientific Workflow Trying To Keep Up With Technology: Accelerating 4D Image Segmentation Applied to NASA data

Scott L. Sellars*
Qualcomm Institute
University of California San Diego
La Jolla, California USA
ssellars@eng.ucsd.edu

John Graham
Qualcomm Institute
University of California San Diego
La Jolla, California USA

Dmitry Mishin
San Diego Supercomputer Center
University of California San Diego
La Jolla, California USA

Kyle Marcus
San Diego Supercomputer Center
University of California San Diego
La Jolla, California USA

Ilkay Altintas
San Diego Supercomputer Center
University of California San Diego
La Jolla, California USA

Thomas DeFanti
Qualcomm Institute
University of California San Diego
La Jolla, California USA

Larry Smarr
California Institute for Telecommunications and Information Technology
Department of Computer Science and Engineering
University of California San Diego
La Jolla, California USA

Camille Crittenden
CITRIS and the Banatao Institute
University of California, Berkeley
Berkeley, California USA

Frank Wuerthwein
Department of Physics
University of California San Diego
La Jolla, California USA

Joulien Tatar
Office of Information Technology
University of California Irvine
Irvine, California USA

Phu Nguyen
Center for Hydrometeorology and Remote Sensing
University of California Irvine
Irvine, California USA

Eric Shearer
Center for Hydrometeorology and Remote Sensing
University of California Irvine
Irvine, California USA

Soroosh Sorooshian
Center for Hydrometeorology and Remote Sensing
University of California Irvine
Irvine, California USA

F. Martin Ralph
Center for Western Weather and Water
Scripps Institution of Oceanography
La Jolla, California USA

***Abstract*—In 2016, a team of earth scientists directly engaged a team of computer scientists to identify cyberinfrastructure (CI) approaches that would speed up an earth science workflow. This paper describes the evolution of that workflow as the two teams bridged CI and an image segmentation algorithm to do large scale earth science research. The Pacific Research Platform (PRP) and The Cognitive Hardware and Software Ecosystem Community Infrastructure (CHASE-CI) resources were used to significantly decreased the earth science workflow's wall-clock time from 19.5 days to 53 minutes. The improvement in wall-clock time comes from the use of network appliances, improved image segmentation, deployment of a containerized workflow, and the increase in CI experience and training for the earth scientists. This paper presents a description of the evolving innovations used to improve the workflow, bottlenecks identified within each workflow version, and improvements made within each version of the workflow, over a three-year time period.**

National Science Foundation Award #1730158 and 1541349

***Index Terms*—Computer systems organization Cloud computing, Computer systems organization Cloud computing, Information systems Computing platforms**

978-1-7281-2451-3/19/$31.00 ©2019 IEEE
DOI 10.1109/eScience.2019.00016

IEEE computer society

I. Introduction

Over the last 50 years, advanced cyberinfrastructure (CI) has evolved greatly, moving from the use of the earliest spreadsheets and databases to supercomputers and dedicated servers, to, now, cloud-based and distributed architecture systems. New infrastructure and tools are constantly coming online and are providing ample opportunity to rapidly expand the frontiers of science and engineering. Traditional approaches to data storage, curation, modeling, and analyzing earth science data is being stressed due to the increased availability of data and demands on computational resources needed to analyze the ever-increasing volumes of data [1]–[3].

A team of earth scientists at the University of California, San Diego (UCSD) and University of California, Irvine (UCI), conducting high dimensional image segmentation on weather and climate data, hit computational limits in their research due to network, hardware, and software constraints. These constraints led to difficulties in processing massive amounts of climate and weather data (for this paper - high-resolution NASA Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA V2) earth science data) leading to long wall-clock times when running experiments and the inability to run multiple computational scenarios to compare results of the research being conducted.

In 2016, to solve this computational earth science problem, the team engaged computer scientists and engineers at the Qualcomm Institute (QI) - University of California San Diego's division of California Institute for Telecommunications and Information Technology (Calit2). Through this direct engagement, the earth science team was encouraged to adapt their CONNected objECT (CONNECT) workflow [4]–[6] for use on the National Science Foundation's (NSF) funded The Pacific Research Platform (PRP) and Cognitive Hardware and Software Ecosystem Community Infrastructure (CHASE-CI) cyberinfrastructure. The hope was that using this advanced cyberinfrastructure could solve their research challenges, such as the difficulty of transferring 10s of terabytes of data to and from different locations, and for rapid data processing with limited bandwidth and computational hardware.

It was clear that adapting the workflow to the PRP would provide dramatic positive impacts on the science. This is because of the combination of the PRP and CHASE-CI resources (referred to as the Nautilus) that included high-speed research network, dedicated network appliances (e.g., Data Transfer Nodes (DTNs)), flexible deployable computing environments (e.g., containerized applications and workflows), and access to accelerator hardware. Workflow process innovations were seen immediately and led to the current use of Nautilus, a distributed cyberinfrastructure with Kubernetes orchestration of hundreds of GPUs, 1000s of CPUs, and terabytes of memory, which reduced the total wall-clock time of the project from 19.5 days to 53 minutes.

Outline. This paper provides a background of the scientific motivation that led to this research and the cyberinfrastructure used in this project, a description of the data, and a walk through of each of the five versions of the workflow developed over the three-year period since the first engagement with the QI computer scientists and engineers. A deep discussion is presented for specific innovations found during each version of the workflow. Benchmark comparisons of wall-clock time are presented, followed by the next steps needed in the research, and wrapping up with concluding remarks that summarize the innovations presented, challenges that arose, and surprises that altered the course of the application of the technology to science.

II. Background

The motivation of our research is to use object-based approaches and image segmentation approaches to better understand climate and weather phenomena by characterizing them not just by their physics or specific weather observations, but by the statistical properties that arise from defining them as time and space propagating objects [4]. In doing so, characteristics of the objects can be data mined and used in Machine Learning for prediction of future weather and climate events.

Specialized algorithms are needed to accomplish the task of identifying, locating, and tracking earth science phenomena. A recent project to evaluate tracking algorithms for a global water vapor transport phenomenon, Atmospheric Rivers, has been developed called, The Atmospheric River Tracking Method Intercomparison Project (ARTMIP). ARTMIP is an international collaborative effort to understand and quantify the uncertainties in atmospheric river (AR) science based on the detection algorithm being used. There are many AR identification and tracking algorithms in the literature with a wide range of techniques and conclusions [7]. One such algorithm is CONNECT.

CONNECT is a "tracking" algorithm that uses a Lagrangian-style detection method and defines earth science phenomena as four-dimensional (4D) objects. The algorithm is run on entire volumes of time and space data to identify connected geospatial pixels as an object by using an instantaneous "footprint" and recognizing the sequential footprints at each time step from the same system with overlapped or connected areas. In other words, at each time step, an object consists of connected voxels (volumetric pixels) in direct neighborhood locations during that time step and in the previous and future time steps. In this way, earth science variables can be described as statistical 4D objects evolving in space (2D), time (1D), and intensity (1D). The algorithm accomplishes this based on a Union Find Method implemented in MATLAB. The approach is similar to blob analysis [8], Connected Component Labeling [9], and Flood-Filling Algorithms [10].

Organizing the data into 4D objects helps one to visualize the dynamical changes to the object in time and space, enabling empirical characteristics to be calculated for each object and studied, providing higher dimensional data and statistics and more advanced understanding of the phenomena than the pixel-level segmentation alone [11]. Large scale scalability

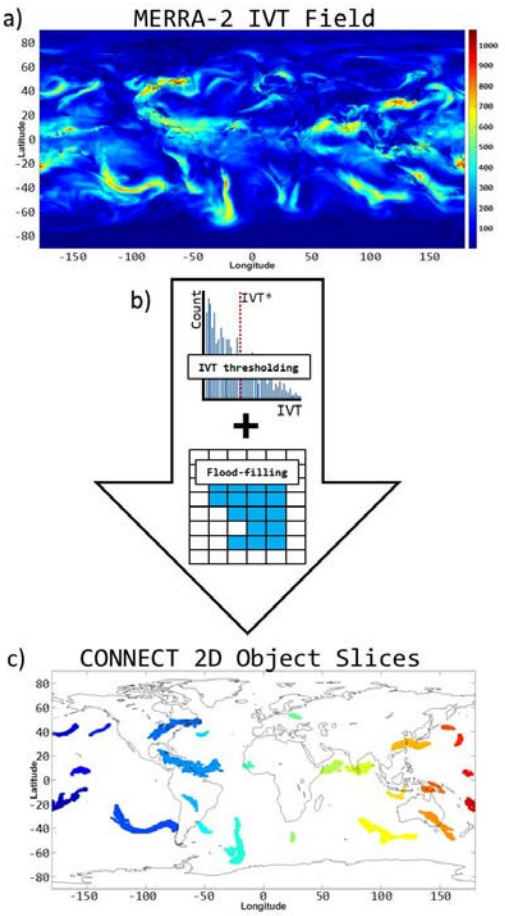


Fig. 1. NASA MERRA data and CONNECT segmentation results. Snapshot in time: January 1, 1980. a) The integrated water vapor transport (IVT) variable. IVT is a commonly used target variable for atmospheric river detection owing to its high correlation with AR precipitation caused by orography. b) For each 2-dimensional (spatial) slice of IVT, each pixel's value is extracted and compared to the user-inputed threshold. c) The masks of boolean trues are reapplied to the IVT field, creating a number of unique IVT "objects" with values greater than the threshold and a background of null values.

and experimentation was not possible with our single CPU MATLAB implementation.

A. *Cyberinfrastructure*

Computer scientists and engineers at the Qualcomm Institute (QI) - University of California, San Diego's division of Calit2 are leading experts in the deployment of state-of-the-art cyberinfrastructure and tools, which are based on the National Science Foundation's (NSF) funded The Pacific Research Platform (PRP) [award # 1541349] high speed network to connect Flash I/O Network Appliance (FIONA) Data Transfer Nodes (DTNs) used to transfer data. More recently, the computer scientists and engineers have deployed the Nautilus cluster through the PRP and the CHASE-CI project funded by the National Science Foundation [award # 1730158]. Nautilus is a distributed cyberinfrastructure with Kubernetes orchestration of hundreds of GPUs, 1000s of CPUs, terabytes of memory, and includes the ability to monitor computational processes using Grafana displays and metrics, integration cloud-based storage (CephFS, Rook, NextCloud, S3) for Machine Learning, and other applications [12], [13]. The distributed set of nodes are based on the PRP FIONAs and are distributed to over 40 institutions around the world.

The Pacific Research Platform (PRP) project design is driven by the high-speed networking needs of collaborative, big-data science. The PRP is a partnership of more than 20 institutions across the world, including the NSF/DOE/NASA supercomputer centers and is connected by deployed Data Transfer Nodes (DTNs) at dozens of partnering sites. Many research disciplines are increasingly multi-investigator and multi-institutional and, therefore, need rapid access to their ultra-large heterogeneous and widely distributed datasets. In response to this challenge, the Department of Energy's ESnet developed the Science DMZ model, a network system optimized for high-performance scientific applications rather than for general-purpose or enterprise computing. The PRP and partners established a high-speed cloud, connected on 10G, 40G and 100G networks using the ESnet Science DMZ [14] model as a basis for its architecture. It has enabled researchers to quickly and easily move data between collaborator labs, supercomputer centers, and data repositories, creating a big-data freeway that allows the data to traverse multiple, heterogeneous networks without performance degradation. The Science DMZ model consists of simple, scalable networks with a focus on with a focus on fast network throughput and high-performance computing. The main focus of the PRP project is to build a researcher-defined and data-focused network.

The CHASE-CI project takes advantage of infrastructure that was built by the PRP and allows a distributed network of appliances to put machine learning tools in the hands of researchers. PRP provided flexible storage options using the PRP's high speed network, allowing users to run their dynamic workflows across the network. To accomplish this, multi-tenant, "FIONA8" machines containing eight game GPUs were installed at various PRP sites, along with over a petabyte of storage (SSD and NVMe) for hosting scientific data with Kubernetes orchestration and a Ceph Object Store were provided by the CHASE-CI project. Using a containerized ecosystem allowed for an extensive hyper-converged system named "Nautilus" to emerge. Nautilus includes the use of CILogon Federated Authentication, Rook/Ceph Cloud-Native Storage, Kubernetes Container Orchestration, Network Monitoring perfSONAR MaDDash, Prometheus / Grafana Dashboards, and Federated Namespaces. The software backbone of Nautilus is Kubernetes, which is a container orchestration engine used for management and job deployment. It is a popular tool first open-sourced by Google in 2014 [15]. Containers provide many advantages including guarantees on environmental consistency, resource isolation, and portability across different networks and computational resources.

TABLE I
GENERAL DESCRIPTION OF EACH VERSION OF THE WORKFLOW, INCLUDING TOTAL WALL CLOCK TIME (IN HOURS).

Date	Iteration	Project	Hardware	Storage	Network	Wall Clock
6/1/16	Version 1	NA	MacBook Pro 2012	Ext USB2.0 HD 5TB	Local	463 hours
8/1/16	Version 2	PRP	MacBook Pro 2012, 2x FIONAs	FIONA storage 240TB	Local/PRP	300 hours
3/1/17	Version 3	PRP	2x FIONAs	FIONA storage 240TB	PRP	38 hours
1/15/18	Version 4	PRP, CHASE-CI	Nautilus	Ceph Volume Store	PRP	25 hours
2/15/19	Version 5	PRP, CHASE-CI	Nautilus	Ceph Volume Store	PRP	.86 hours

III. DATA

For the workflows described in this paper, 246GB of assimilated meteorological data is used. The data has a temporal resolution of 3 hours and is from NASA Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA V2) [16] from January 1, 1980 to May 31, 2018. This data was obtained, and additional variables are calculated. In this case, Integrated Water Vapor Transport (IVT) data was calculated from the raw NASA data archive. The entire MERRA V2 M2I3NPASM archive is 16TB, which includes 14 variables and 42 vertical levels in the atmosphere. The data has a temporal frequency of 3-hourly from 00:00 UTC (instantaneous), with a 3-D spatial grid at full horizontal resolution. The resolution is 0.5 x 0.625 in latitude and longitude (i.e., global resolution of 576x361 pixels). The data was downloaded from NASA's GES DISC data portal and is stored on a FIONA using Unidata's Thematic Real-time Environmental Distributed Data Services (THREDDS) [17] for rapid access to a THREDDS Data Server located on the PRP network (UCSD THREDDS - stores NASA MERRA data archives[1]). The THREDDS Data Server (TDS) is a web server installed on a FIONA and exposed to the entire PRP with access to the NASA data.

IV. SCIENTIFIC WORKFLOWS INNOVATION

Five workflows are described, with the original workflow developed in 2012 at the University of California, Irvine. Each workflow utilized computation and storage resources available to the researchers at the time of development and increase in sophistication as the deeper CI collaboration continued. Each version provided new innovations and were possible because of the skills gained throughout the collaboration. The first workflow was developed independently without CI expertise. The following four versions were developed in close collaboration with leading CI experts and engineers using the PRP and CHASE-CI projects over a three-year period, which included the availability of new technology that came online throughout the project. The code for Versions 4 and 5 can be accessed using Nautilus's GitLab repository[2].

A. Version 1, 2012-2016

The first version of the workflow starts out like most scientific problems with multi-institutional collaborations. The team begins with a hypothesis about a phenomenon and then determines which collaborator has computational and data storage resources to use for running experiments to test the hypotheses. The team would then download data to the local computational resource, run the algorithm, and then share the results with collaborators for further analysis to determine the correctness of the hypothesis. This type of collaboration can be challenging when dealing with 10s of terabytes of data.

In our case, the first step before running CONNECT was to decide on which variable would be used in the segmentation process from the NASA MERRA V2 archive of 14 variables. Here, the 4 atmospheric variables (out of 14 possible) at 4 atmospheric pressure levels (out of 42 possible) were needed, with each file size at around 250MB. To download this data, the second step required submitting a request to NASA's online data portal for access to download these four variables to a local machine. The GES DISC site[3] was used to download data from NASA to a MacBook Pro (2012) with an attached USB external 5TB hard drive connected to an ethernet port in an office at the University of California, San Diego's Scripps Institution of Oceanography (SIO). The total wall clock time from submission of the request to NASA to the completed data transfer to the MacBook Pro (2012) was 7.45 days. Broken down, it took 4.9 days for the data to be organized on the NASA side and 2.5 days to download the data. The transfer included 11,052 files (2.4T) at 11.7 MB/s.

Step four calculated our variable of interest on the MacBook Pro (2012), reducing the data volume from 2.4TB to 246GB, which took an additional 10.2 days. The MacBook Pro is used as the staging computer to download and run the algorithm. Step five shared the newly calculated IVT data (246GB) with colleagues at the University of California, Irvine (UCI). During step six, the colleagues at UCI run the CONNECT algorithm on the 246GB of data, performing 4D object segmentation using a research group's shared computing resource, and then share the results back with UCSD and others in step seven and eight. Total wall-clock workflow time for applying CONNECT to NASA MERRA v2 data in the Summer of 2016 for a single variable was 463 hours (19.3 days), as seen in Table 1. The team wanted this workflow to run on thousands of variables and seeing how time intensive one variable could be, it was not possible with the current workflow.

B. Version 2, late 2016

Due to the limited bandwidth potential of the MacBook Pro and barriers to using an External USB hard drive for

[1]https://thredds.nautilus.optiputer.net/thredds/catalog/catalog.html

[2]https://gitlab.nautilus.optiputer.net/connect

[3]NASA Data Portal: http://disc.sci.gsfc.nasa.gov

storing and transferring results, the team realized that specialized hardware was needed to conduct fast large volume data transfers. Therefore, the major innovation discovered and deployed in the second version of the workflow was the use of FIONAs, with one located at UCI and two at UCSD. With the high capacity network in place, instead of using NASA's data portal, the entire 16TB archive was downloaded directly to the FIONA located at UCSD, which included the four variables needed to calculate IVT, but also many others that could be used in additional research. The TDS provided rapid access to data over the PRP to any user, especially when using FIONAs, rather than relying on downloading data from the NASA data portal each time we wanted to run an experiment on a new variable or pressure level.

Simply using the FIONA to download the data directly from the NASA ftp server sped up the download speed by 4x (40MB/s). With this increased bandwidth, the team did not have to subset the data first using NASA's data portal and instead was able to download the entire M2I3NPASM_V5.12.4 archive (16TB). The same 2.4TB of data that was downloaded in Version 1, using the FIONA, dropped the wall clock time from 7.45 days to less than one day (17 hours). Harnessing the PRP and FIONAs also provided ease of transferring the data to and from UCI and UCSD as research experiments took place. The data download step was dramatically improved, however, constant data transfers between the two FIONAs became problematic and time consuming given the experiments the team wanted to perform. The total wall-clock time reduced from 463hrs (19.3 days) to 300hrs (12.5 days).

C. Version 3, mid 2017

It became clear that transferring data locally to a remote server or MacBook Pro, regardless of using FIONAs, was too time consuming and burdensome for processing big data. The innovation in Version 3 was the acknowledgment that FIONAs themselves could be compute nodes, not just DTNs.

The FIONAs each have upwards of 160TB of storage, and therefore, plenty of space to conduct the workflow computation on the FIONA machines, rather than using the PRP to transfer results to and from collaborating institutions to do the computation. The PRP had placed a FIONA at UCI, which we used as a compute machine, and a FIONA at UCSD, which ran the TDS and provided rapid access to the MERRA data. Not only did using the UCI FIONA as a compute machine allow for rapid CONNECT object segmentation, but it also allowed for multiple variables, multiple thresholds, and the generation of hundreds of thousands of objects to be generated. Current data transfers between UCI and UCSD FIONAs reached upwards of 230MB/s.

In the end, using the FIONA as a compute node, and having rapid access to the NASA data, shortened the total wall clock time from 12.5 days (300hrs) to essentially the time it took for the CONNECT algorithm to run at 1.6 days (38hrs) on the FIONA. It should be noted that it was important to have components at both UCI and UCSD locations. This was a multidisciplinary team, at multiple universities, and we wanted to demonstrate this multi-campus collaboration.

D. Version 4, late 2018

Accessibility of the Nautilus system allowed innovation to continue in 2017 into late 2018. Version 4 was a major overhaul of the workflow, including experimenting with a Machine Learning approach to do object segmentation. The QI team upgraded the original PRP FIONAs to FIONA8s, which included GPU compute nodes using Kubernetes (k8s) for container orchestration and started the use of Ceph Object Store for cloud-based storage. In addition, Nautilus has a GitLab instance, that is used to store code, build containers, and allow k8s to pull images for each step of the workflow.

The FIONA8 continues to have the ability to rapidly transfer data, but it also conducts large scale computations using GPUs. The team decided to experiment with a distributed data download procedure and other Machine Learning algorithms developed for GPUs acceleration to increase the object segmentation speed. Instead of using CONNECT's MATLAB functions (Versions 1 to 3), which use a single CPU to do the object segmentation, the Flood-Filling Networks (FFN) [18] algorithm was used. FFN is based on a 3D Convolution Neural Network (CNN), written in Google's TensorFlow, and is able to separate objects within a 3D volume of spatial data, or images, by using a deep stack of 3D convolutions. The network is trained to take an input object mask within the network's field of view to infer the boundaries of the objects. It was originally designed to segment 3D volumes of neurological data. FFN generated objects are much different than the objects produced by the original CONNECT algorithm. The FFN makes an inference using different information than CONNECT. CONNECT simply looks for directly connected voxels in time and space, whereas the FFN draws inference based on features that the 3D CNN uses, including geographical region, curvature, edge structure, and temporal evolution and lifetime. It is important to note that this object segmentation approach is different from previous CONNECT algorithm versions, although the goal of segmenting 4D earth sciences objects is the same.

This workflow was recently published in SNAC19 (See Altinitas 2019 for details). Summarizing the innovations, this workflow included the use of a distributed CI system and accelerated hardware (NVIDIA GPUs) managed by k8s provided by the CHASE-CI project. The project allowed the new workflow to be broken up into separate steps using docker images and k8s pods deployments. These steps included downloading data from TDS and data preparation (14 k8s pods using 42 CPUs and 256GB of memory), model training on historical data (1 k8s pod using 1 CPU, 1 GPU, and 15GB of memory), and distributing the inference job to 50 GPUs (50 k8s pods using 50 CPUs, 50 GPUs, and 600GB of memory). The entire 246 GB (576x361x112,249 or 2.3e10 voxels) is evenly distributed across the 50 GPUs and the total inference time is 18 hours 53 minutes (1133 minutes). In total, the wall-clock time for this version of the workflow, including

transferring data from TDS, input preparation, model training and inference was 25hrs (1.01 days).

E. Version 5, early 2019

Through all of these innovations, the importance of collaboration and interdisciplinary work cannot be overstated. This is most evident when Neuroscientists at Princeton University openly published a 3D Connected Component Labeling algorithm[4] that meets the criteria of the CONNECT algorithm and is written in Cython. Using this Cython optimized algorithm on the Nautilus system, rather than the CONNECT algorithm and version 4 FFN algorithm, the team was able to distribute 246 GB of data across 50 workers using 25 pods, which dramatically decreased the total workflow wall-clock time from 24.6hrs (using version 4 algorithm) to 52 minutes.

The main improvement came from dedicated programming of the segmentation approach using Cython. It is important to note that Version 5 innovations not only include the dramatically improved segmentation algorithm, but in combination with the previous experiences and lessons from Versions 2-4, provided an integration of the system as a whole, which produced the innovation in rapid object segmentation. For example, innovations to the previous versions include harnessing "worker nodes" to distribute the job across the Nautilus cluster and the inclusion of new NVMe storage drives for high-speed data access using Ceph Object Store, which now allows the entire workflow to be orchestrated in under one hour (52 mins). Each step, previously described, now uses multiple workers to download, process, and segment data.

With these rapid capabilities, multiple variables, geographical regions, and time ranges can be easily segmented and analyzed. Figure 2 shows a high dimensional volume that includes three variables in the MERRA archive that can be studied using object-based approaches.

TABLE II
NAUTILUS RESOURCE SUMMARY TABLE FOR ALL STEPS IN THE VERSION 5 WORKFLOW

Version 5	Download	Prep+Segmentation
# of Pods	14	25
# of CPUs	42	25
Data Processed	246GB	246GB
Memory	225GB	850GB
Total Time	37m	11m

V. PERFORMANCE MEASUREMENT

The expanding set of CI resources provided by the PRP and CHASE-CI allowed the earth sciences team to gain experience with and experiment with a variety of new technologies and machine learning methods, which consistently improved the total wall-clock time of the workflows described in this paper. However, using a variety of different technologies and methods also made a direct comparison of the five workflow versions challenging. Regardless of these challenges, two key elements of the workflows can be benchmarked and compared for a clear description of the improvements. These elements are: 1) remote data transfer and 2) data segmentation.

[4]https://github.com/seung-lab/connected-components-3d

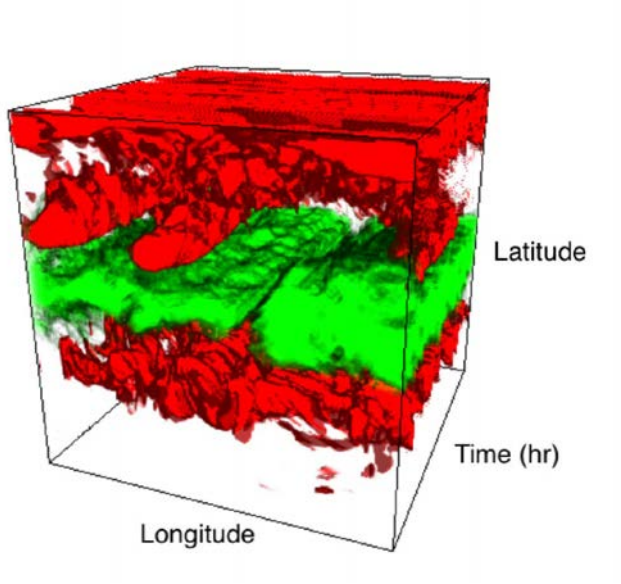


Fig. 2. A high dimensional data volume overlaying two variables - Sea Level Pressure objects (Red) and Tropical Moisture objects (Green). The cube is oriented with the x-axis representing longitude, the y-axis is latitude, and the z-axis is a 30 day time period at 3-hourly resolution.

It should be noted that the total wall clock time, hardware and software utilization for each workflow was calculated at the completion of each version. Version 1 and 2 performance was based on the Science DMZ network at UCSD in 2016 utilizing a MacBook Pro (2012) and access to a data server at UCI. These factors make it challenging to reproduce these results as the network and hardware has changed. With this said, the original algorithm and data are available. Versions 3, 4, and 5 can be reproduced on the current FIONAs and using Nautilus resources. All code is stored on the Nautilus GitLab repositories.

A. Remote Transfer

Data transfer is required for each workflow. Table 3 shows a comparison of data transfer speed (MB/s) and wall-clock time (hours) for each of the workflow versions described above.

Row one shows Version 1 workflow download process using the NASA GES DISC, which took a total of 179 hours to download 2400GB (2.4TB) of data. Simply using the PRP and a FIONA to directly access and download the data in Version 2, increased the data transfer speed by roughly 4x (from 10MB/s to 40MB/s), while also sidestepping the need to use the portal, reducing the total data transfer time to 17 hrs, as shown in row two.

The unexpected aspect of this for the UCSD team was not only the ability to increase the transfer speed, but also to be able to use the FIONA as a data store with the TDS, providing rapid access to the entire data set. Once using the PRP and FIONAs, the transfer rates greatly approved. When using the

PRP and FIONA to FIONA data transfer, the transfer rates reached 530MB/s. A test using Globus to transfer data from a FIONA to Ceph Storage Volume was performed. Current data transfers between UCSD FIONAs and Nautilus Ceph Object Store reached upwards of 230MB/s. An important note is that the raw trasfer of 2.4TB of data from the FINOA to CephFS did see an increase in total time (as seen in row 5 of Table 3). This increase is thought to be because of the file writing capabilities of CephFS. Finally, Version 4 and 5 harness the capabilities of k8s to have multiple worker pods (20 workers) perform multi-stream download using Aria2 from TDS to Ceph Volume Store of the 246GB processed dataset with a wall-clock time of 37 minutes.

TABLE III
COMPARISON OF REMOTE TRANSFER APPROACH WALL-CLOCK TIMES

	Protocol	Size (GB)	Speed (MB/s)	T (hrs)
NASA to MacBook	wget	2400	11.7	179
NASA to FIONA	wget	2400	40	17
FIONA to FIONA	https	2400	530	1.2
FIONA to CephFS	Globus	2400	230	2.9

B. Segmentation

Each of the approaches described in this paper accomplished object segmentation in 4D, making segmentation a good benchmark for workflow comparisons.

TABLE IV
COMPARISON OF SEGMENTATION APPROACH WALL-CLOCK TIMES

Segmentation Method	Data Size (GB)	Wall-Clock Time (hrs)
MATLAB CONNECT	246	38
Tensorflow FFN	246	24.6
Cython cc3d	246	.11 (7 mins)

A tremendous wall-clock time improvement is seen in Table 4 moving from a MATLAB single CPU implementation at 38 hours, to the GPU accelerated version at just over 24 hours, to finally Cython to implementation at 7 minutes. It is important to note that the 7 min time to completion is only possible because of the distributed resources available by Nautilus allowing for 10 pods to orchestrate 25 worker jobs running the Cython code on different chunks of data. These segmentation benchmarks do take into account post-processing needed to do final analysis.

C. Bottlenecks

In this section, we will discuss the bottlenecks uncovered within each workflow version as well as how CI resources and technologies helped the research team to decrease these bottlenecks and improve the workflow over time. Domain scientists and researchers do not have the CI knowledge and expertise needed to often identify the bottlenecks in their data processes when conducting research. Collaborations with CI experts to identify these bottlenecks and resolve them is important to improving these processes and research.

Version 1 bottlenecks were found at every workflow step. From using a MacBook Pro to orchestrate data downloads, transfers, and analysis to sharing the computing resource with an entire research group. The first bottleneck was the process of requesting data access from NASA, which took much longer than expected. Another bottleneck was downloading the data to an external hard drive attached to the MacBook Pro using a standard ethernet connection. In addition, the MATLAB single CPU implementation was not optimal, and finally, transferring the data to colleagues at UCI using ssh and cp commands added to the overall wall-clock time. Each of these bottlenecks was improved in the following versions.

Version 2 was able to eliminate one of the major bottlenecks in Version 1: NASA approval and data transfer; and did so by using a PRP optimized FIONA to access the NASA data directly and download the entire data set. However, Version 2 still had a bottleneck in the second step, which was transferring the data from the FIONAs to a local machine (MacBook Pro) and a remote server to run the algorithm.

Version 3 solved the two major bottlenecks that were found in Version 2: data transfer and workflow computation speed. This was accomplished by conducting all computations on the FIONA itself, as opposed to running computations on a local machine. The team installed MATLAB directly on the UCI FIONA and was able to run all segmentation scripts on the FIONA, to minimize local data transfer, however, this also sped up the overall workflow, eliminating the second bottleneck in Version 2. However, though it was improved, the algorithm remained the bottleneck in Version 3 as the computation speed was still not what the team wanted.

Version 4 saw the large change to the workflow with the emergence of Nautilus and the TDS FIONAs becoming GPU compute nodes, FIONA8s with k8s orchestration and having access to a distributed Ceph Object Store allowing rapid access to data, models, and results. Deep Learning was explored for object segmentation given the access to mutli-GPUs. Yet, even with access to accelerated hardware, the team ran into additional challenges not present with the original approach. The model training process and model prediction is time-consuming, with over 306 mins needed to do training on a small percentage of the 246GB and 1133mins to run the model on the entire dataset. Ongoing experiments with model settings are expected to improve the inference time and will be reported on in future publications. In addition, distributed training is an active area of research that should dramatically increase the speed of training and the team is working in this direction.

The use of the Cython algorithm in Version 5 eliminated the computational bottleneck from all previous versions, and the speed of object segmentation went from hours/days to minutes. However, this took the team full circle, leading to a bottleneck that had not been considered since Version 2, the ability to transfer data. The bottleneck originates from the current deployment of the TDS, which only uses a single k8s pod. Current research looks to scale the number of TDS pods, allowing for increased number of download streams, which we expect would increase the download transfer rates in the future.

VI. Future Steps

The future holds a wide range of possibilities and exciting innovations for this workflow. From a science perspective, since we have the ability to run the all of the steps in minutes, this gives us the ability to run several experiments at once on multiple variables using multiple thresholds. To fully accomplish this, there are a number of intermediate steps that can also be improved and typically involves preparing the variable for input into an algorithm or model by ensuring the volume characteristics (i.e., size, missing data, etc.) are understand on standardized so that the workflow can process in a distributed way. This would include finding network paths to additional earth science datasets to be connected to Nautilus so that they can be included in the experiments. We envision petabytes in size, nearly impossible to move to a local server for processing but would be available across the network for processing using this workflow. Work on this has already been started where a list of data files with specific variables has been prepared in order to start the download process.

Once all input data has been processed then using Kubernetes, we can start to run more experiments in parallel. They would all run the same code but have different input configuration so that each run is processing different input files. We envision a queuing system where experimental configuration are put inside of a work queue and worker nodes process these jobs and start a workflow run automatically. This would then give us a very high-level abstraction to this whole process where now the scientists only need to fill in a configuration and the experiment is started.

Another exciting future step in this workflow is integrating the machine learning workflow (Version 4) with Cython objects algorithm (Version 5) to train and validate new ML approaches to learning features of earth science phenomena. This is an integral step because the Cython object algorithm works so fast that a lot of training data can be generated. Before, we were the training is on a small set of training data, but with a larger amount of training data we can experiment more with ML approaches for describing and labeling objects using the vast array of GPUs available through Nautilus.

Distributed ML training is another future step that is necessary to decrease the wall clock time even more. Currently training is done on a single GPU, however as more training data is provided it will take more and more time to train if the training is not done in a distributed manner. TensorFlow allows the creation of distributed training so the code will have to be reworked into to accomplish this. This will then give us the ability to use a large amount of training data while still keeping the time to train very low.

Finally, there are many optimizations within the k8s orchestration that are currently being worked on. These include using advanced files systems, k8s workflow management, memory use and allocation, and optimized scaling of resources to further improved these efforts.

VII. Conclusion

CONNECT seeks to study hundreds of terabytes of earth science data using object-based approaches. Prior to adapting the workflow, the original CONNECT workflow (Version 1) would not be able to do this and only allowed slow and limited questions to be asked because of the time it took to generate results. The PRP provided additional expertise that laid the foundation for the team to rerun, adjust, and try new variables and algorithm settings without the need to download new datasets or subsets from traditional data portals. At the time, these wait times for data transfers were previously thought to be a normal part of doing science. This is obviously not the case, as demonstrated by these results. The original objective is now entirely possible, but so are other and potentially more important objectives which are currently being completed. Imaging real-time analysis of high-resolution phenomena, especially with NASA data that becomes available, can be rapidly processed, and results analyzed afterwards.

Through this collaboration of earth science researchers and computer scientists, a series of workflows were developed over a three-year period. Each workflow was the team's attempt to take advantage of the technologies available and to become proficient in these technologies, so that as the technologies evolved, the team could adapt previous work to the new capabilities. It is important to note that the workflows evolved just as much as the technology did. Considering the original workflow (Version 1), the improvements and enhancements were only possible because of the reliable support and expertise provided by the CI team, as well as their encouragement to adapt the original workflow, creating an environment where the team could rapidly experiment and try new methods. An important point is that Version 4 and 5 could not be run on the original MacBook Pro, nor would they be practical on a single FIONA. This conclusion highlights the capabilities now offered by the PRP and the CHASE-CI infrastructure built on the PRP.

Many innovations were discovered throughout the three-year project, including how the access to high speed networks and high-performance storage provided the capability to transfer large data sets many times, allowing for rapid exploration of the segmentation approaches and datasets downloaded. This advanced the development of a flexible workflow, which helped to scale the segmentation approach for the use of many variables and datasets. For a single variable, the total wall-clock time was reduced from 19.5 days to 52 minutes. This reduction in wall-clock time allows for the generation of millions of objects to be studied and analyzed in the near future. As additional technologies continue to come online, the expertise obtained will allow for further innovations to be discovered. The challenge is to keep up with technology and quickly update and take advantage of the capabilities that are provided by these technologies as the come online. This supports the flexible and dynamic environments provided by the PRP and CHASE-CI as essential to rapidly improving the

workflow with future innovations of both technology and new ways to study our Earth's phenomena. These innovations will all contribute to the main objective of this project: to have the capability to do rapid object segmentation of hundreds of terabytes, soon to be petabytes, of earth science data. In addition, teaching and training students on cutting edge data analysis hardware and software tools, methods, and technology reported in this paper are essential and efforts at UCSD are achieving this and building on the many related projects.

Looking forward, there are many new datasets to explore and relationships between variables to find. The authors expect many future science papers to be published using this workflow, and future workflows, which are provided by these innovations and technologies. Being able to study these new object-based datasets will help us to better understand the physical processes governing the hydrological cycle. This will take time, but the workflow and infrastructure laid out in this paper show that rapid object segmentation of large datasets is possible, thus leading to deeper exploration of earth science data, providing new characteristics and statistics to be studied.

ACKNOWLEDGMENT

The authors would like to acknowledge the National Science Foundation award #1730158 and 1541349, The Center for Hydrometeorology and Remote Sensing at the University of California, Irvine, The Center for Western Weather and Water Extremes at Scripps Institution of Oceanography, and Calit2's Qualcomm Institute for all of their support.

REFERENCES

[1] A. Kelbert, "Science and cyberinfrastructure: The chicken and egg problem," *Eos*, 2014.

[2] J. Cutcher-Gershenfeld, K. S. Baker, N. Berente, D. R. Carter, L. A. DeChurch, C. C. Flint, G. Gershenfeld, M. Haberman, J. L. King, C. Kirkpatrick, E. Knight, B. Lawrence, S. Lewis, W. C. Lenhardt, P. Lopez, M. S. Mayernik, C. McElroy, B. Mittleman, V. Nichol, M. Nolan, N. Shin, C. A. Thompson, S. Winter, and I. Zaslavsky, "Build it, but will they come? A geoscience cyberinfrastructure baseline analysis," *Data Science Journal*, 2016.

[3] S. L. Sellars, "Grand Challenges in Big Data and the Earth Sciences," *Bulletin of the American Meteorological Society*, 2018.

[4] S. Sellars, P. Nguyen, W. Chu, X. Gao, K.-l. Hsu, and S. Sorooshian, "Computational Earth Science: Big Data Transformed Into Insight," *Eos, Transactions American Geophysical Union*, vol. 94, pp. 277–278, 2013.

[5] S. L. Sellars, X. Gao, and S. Sorooshian, "An Object-Oriented Approach to Investigate Impacts of Climate Oscillations on Precipitation: A Western United States Case Study," *Journal of Hydrometeorology*, p. 150119122635001, 2015. [Online]. Available: http://journals.ametsoc.org/doi/abs/10.1175/JHM-D-14-0101.1

[6] S. L. Sellars, B. Kawzenuk, P. Nguyen, F. M. Ralph, and S. Sorooshian, "Genesis, Pathways, and Terminations of Intense Global Water Vapor Transport in Association with Large-Scale Climate Patterns," *Geophysical Research Letters*, 2017.

[7] C. A. Shields, J. J. Rutz, L. Y. Leung, F. Martin Ralph, M. Wehner, B. Kawzenuk, J. M. Lora, E. McClenny, T. Osborne, A. E. Payne, P. Ullrich, A. Gershunov, N. Goldenson, B. Guan, Y. Qian, A. M. Ramos, C. Sarangi, S. Sellars, I. Gorodetskaya, K. Kashinath, V. Kurlin, K. Mahoney, G. Muszynski, R. Pierce, A. C. Subramanian, R. Tome, D. Waliser, D. Walton, G. Wick, A. Wilson, D. Lavers, Prabhat, A. Collow, H. Krishnan, G. Magnusdottir, and P. Nguyen, "Atmospheric River Tracking Method Intercomparison Project (ARTMIP): Project goals and experimental design," *Geoscientific Model Development*, 2018.

[8] T. Blaschke and G. J. Hay, "Object-Oriented Image Analysis and Scale-Space: Theory and Methods for Modeling and Evaluating Multiscale Landscape Structure," *International Archives of Photogrammetry and Remote Sensing*, 2001.

[9] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, and Y. Chao, "The connected-component labeling problem: A review of state-of-the-art algorithms," *Pattern Recognition*, 2017.

[10] S. V. Burtsev and Y. P. Kuzmin, "An efficient flood-filling algorithm," *Computers and Graphics*, 1993.

[11] P. Nguyen, S. Sorooshian, A. Thorstensen, H. Tran, P. Huynh, T. Pham, H. Ashouri, K. Hsu, A. A. Kouchak, and D. Braithwaite, "Exploring trends through "rainsphere": Research data transformed into public knowledge," *Bulletin of the American Meteorological Society*, 2017.

[12] L. Smarr, "CHASE-CI: A Distributed Big Data Machine Learning Platform, Opening Talk With Professor Ken Kreutz-Delgado," Qualcomm Institute University of California, San Diego, 2018.

[13] J. Graham, "Building the Pacific Research Platform: A Workshop Towards Deploying a Science-Driven Regional Big Data Freeway'," Calit2, San Diego Supercomputer Center, and CITRIS, San Diego, 2015.

[14] Fasterdata.es.net, "Science DMZ Architecture," 2019. [Online]. Available: https://fasterdata.es.net/science-dmz/science-dmz-architecture/

[15] Kubernetes.io, "Kubernetes Documentation," 2019. [Online]. Available: https://kubernetes.io/docs/home/

[16] M. M. Rienecker, M. J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G. K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. Da Silva, W. Gu, J. Joiner, R. D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C. R. Redder, R. Reichle, F. R. Robertson, A. G. Ruddick, M. Sienkiewicz, and J. Woollen, "MERRA: NASA's modern-era retrospective analysis for research and applications," *Journal of Climate*, 2011.

[17] B. Domenico, J. Caron, E. Davis, R. Kambic, and S. Nativi, "Thematic Real-time Environmental Distributed Data Services (THREDDS): Incorporating interactive analysis tools into NSDL," 2002.

[18] M. Januszewski, J. Kornfeld, P. H. Li, A. Pope, T. Blakely, L. Lindsey, J. Maitin-Shepard, M. Tyka, W. Denk, and V. Jain, "High-precision automated reconstruction of neurons with flood-filling networks," *Nature Methods*, 2018.