Center for Western Weather and Water Extremes

SCRIPPS INSTITUTION OF OCEANOGRAPHY AT UC SAN DIEGO

IDI/Research IT Showcase:
Research Application: Big Data and Earth Sciences

Scott L. Sellars

scottsellars@ucsd.edu

May 17, 2017

Team: Phu Nguyen[2], John Graham[3], Joulien Tatar[3], Tom DeFanti[3] ,Larry Smarr[3], F. Martin Ralph[1] and Soroosh Sorooshian[2]

[1]Center for Western Weather and Water Extremes, UCSD, La Jolla, CA

[2]Center for Hydrometeorology and Remote Sensing, UCI, Irvine, CA

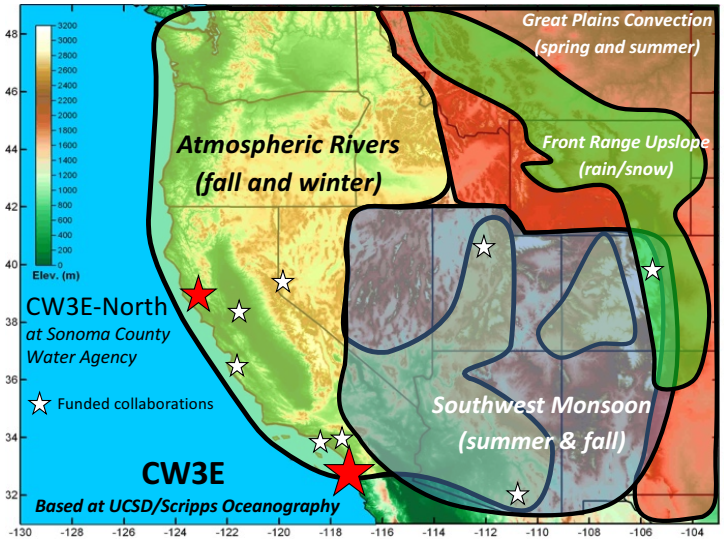
[3]California Institute for Telecommunications and Information Technology (Calit2), UCSD, La Jolla, CA

My Background

- BS Meteorology – University of Utah
- MA Climate and Society – Columbia University
- MS Civil Engineering – University of California, Irvine
- Ph.D. Civil Engineering – University of California, Irvine
- 2014/2015 California Council on Science and Technology Policy Fellow
  - Assemblymember Bill Quirk (D-Hayward)
- Postdoctoral Scholar – Scripps Institution of Oceanography
  - Forecast-Informed Reservoir Operations
    - Explore forecast accuracy and decision support logic for re-operations of a reservoir
  - Big Data

Outline

- Background on CW3E's research and IT needs
  - West-WRF, High Resolution Numerical Model
  - CONNected objECT (CONNECT) Algorithm



Key Phenomena Causing Extreme Precipitation in the Western U.S. (Ralph et al. 2014)

Center for Western Weather and Water Extremes

SCRIPPS INSTITUTION OF OCEANOGRAPHY AT UC SAN DIEGO

Director: F. Martin Ralph, Ph.D. **Website: cw3e.ucsd.edu**

Strategies: Observations, physical processes, modeling, decision support

Scope: A group of roughly 40 people with 10 major projects

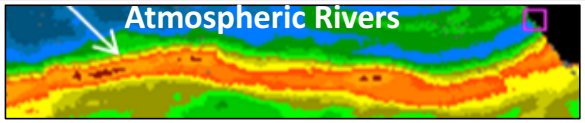
Partners: California DWR, Sonoma County Water Agency, CNAP, USGS San Diego Supercomputing Center

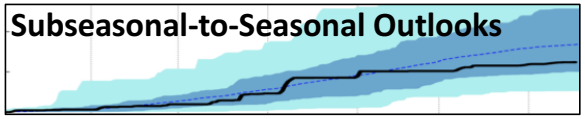
Sponsors: CA DWR, USACE/ERDC, NOAA, SCWA, NASA, USBR

Data is at the heart of what we do!

- High resolution numerical models
- Satellite images
- Ground based weather stations
- Weather radar
- Historical climate data

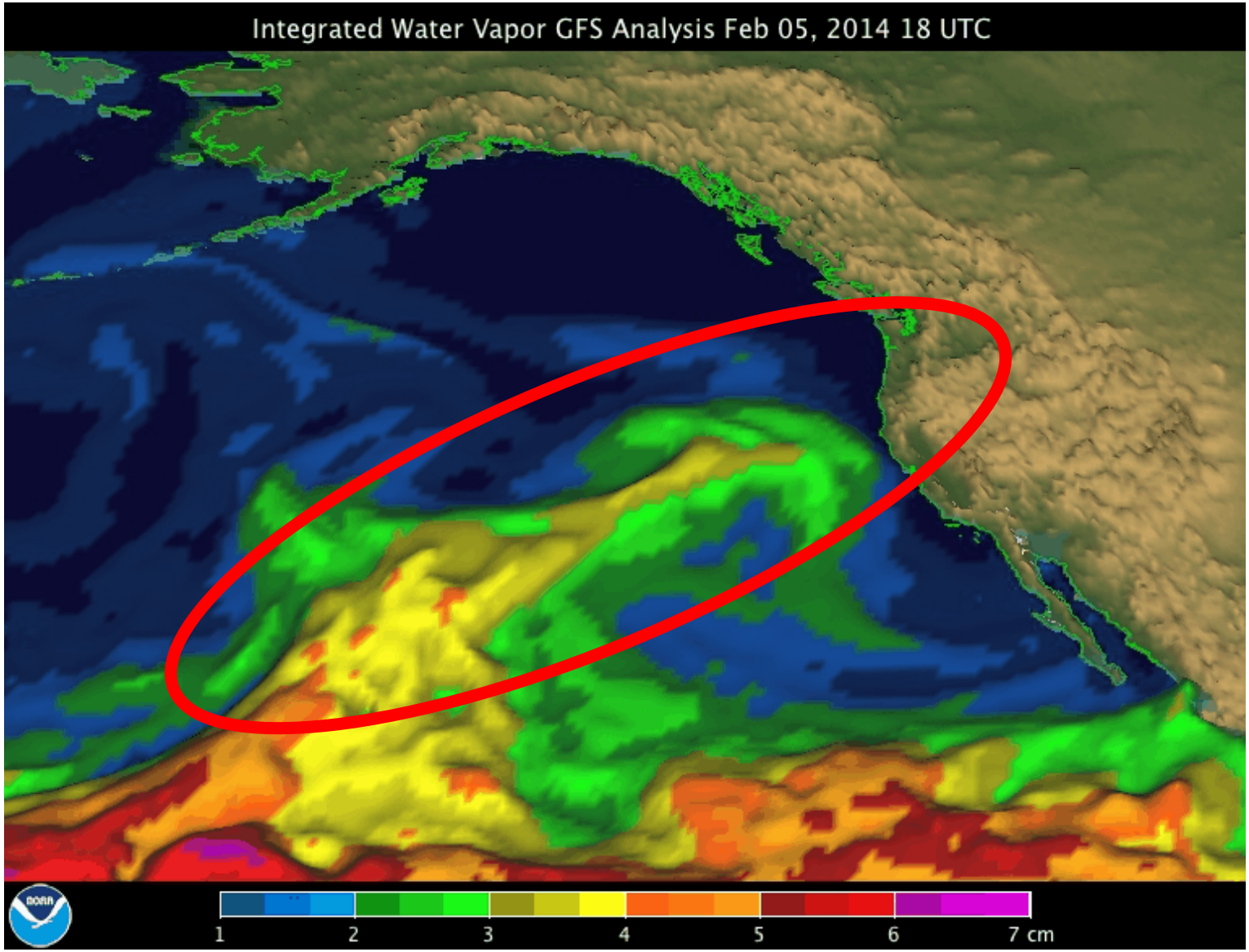
CW3E's Core Efforts





*NASA Apollo 17 crew traveling toward the Moon

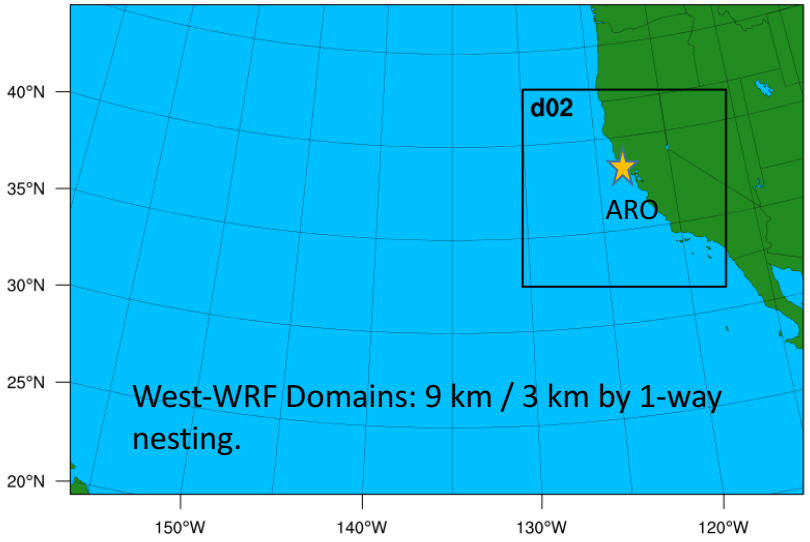
CW3E


Integrated Water Vapor GFS Analysis Feb 05, 2014 18 UTC
1
2
3
4
5
6
7 cm

What is West-WRF?

CW3E has developed West-WRF to:

1. Serve as a testbed for understanding physical processes and their relationship to forecast error.
2. Improve the accuracy of extreme event forecasts. In the western US, these events pose unique challenges (see table)



Unique Forecast Challenges Posed by Western US Extreme Events

Challenge	Primary NWP Shortcoming	References
AR Landfall Characteristics	Location and strength of water vapor flux	Wick et al. (2013) Ralph et al. (2017)
Extreme Precipitation Skill	Overprediction of light rain, Underprediction of extreme amounts	Ralph et al. (2010) Ralph and Dettinger (2012) Sukovich et al. (2014)
Snow level	Low precision, Biases near terrain	White et al., (2010) Neiman et al. (2014) Minder and Kingsmill (2013)

CW3E-SDSC Partnership

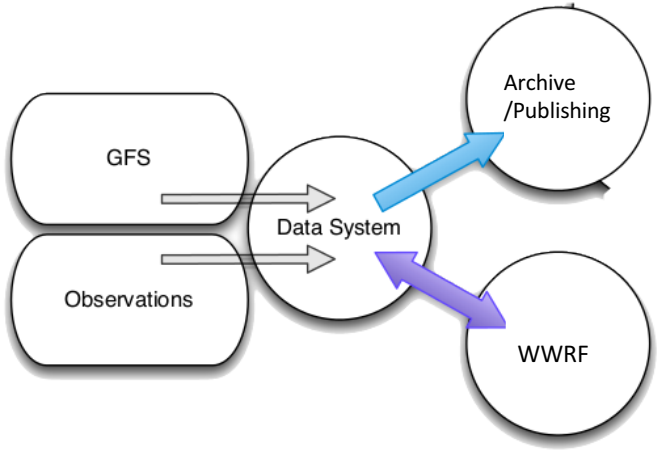
- Interdisciplinary team of SIO & SDSC Scientists, post-docs grad students and programmers.
- Working to an integrated *research* and *operations* plan
- SDSC Director and UCSD Physics Professor Mike Norman has been supportive of CW3E
- Computer time and disk storage on the *Gordon and Comet* supercomputers and 2 dedicated preprocessor and storage machines.

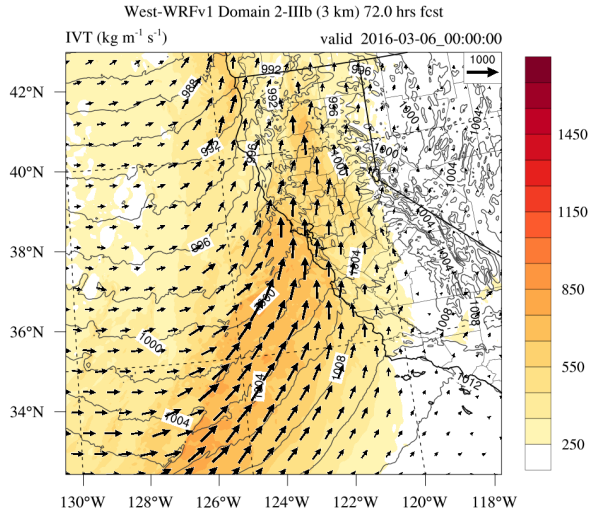
Extramural Support

- **NSF XSEDE**: (A. Martin, PI)
  Research and Production computing time
- **CA DWR Early Awards**: (F. M. Ralph, PI)
  Dedicated machines for preprocessing and storage
- **USACE FIRO**: (F. M. Ralph, PI)
  Research personnel at CW3E and subcontracted collaborators
- **CA DWR CA Info**: (F. M. Ralph, PI)
  Research and Production personnel at CW3E and SDSC

Real-Time CW3E West-WRF Forecasts

Automated Data Ingest, Archiving and Preprocessing System Developed at SDSC with Funding from DWR





To left: real-time forecast of IVT (kg m^{-1} s^{-1}), IVT vectors, and sea-level pressure produced by West-WRF and published at cw3e.ucsd.edu

Event depicted: March 6, 2016

Cool Seasons (Dec – Mar) 2015,2016,2017:

- forecasts issued daily
- Finest spatial resolution: 3 km
- Simulated radar and GPS occultation observations at key locations
- Millions of supercomputing units used
- 3-D atmospheric data produced: 4.5 Tb

- Data delivered in near-real-time
- Forecast products published real-time to cw3e.ucsd.edu

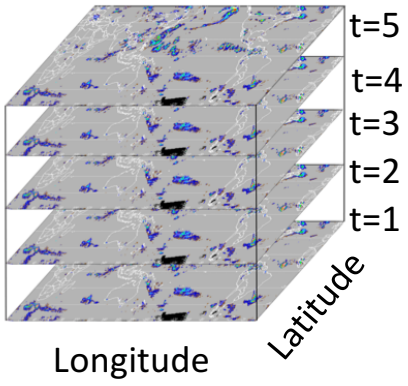
A. Martin, J. Helly, B. Kawzenuk

CW3E

Multi-Institutional Big Data Transformed Into Insight

- For Computational Earth Sciences (Sellars et al. 2013, 2015)
  - **CONNected objECT (CONNECT) Algorithm, developed at UCI-CHRS**
    - **Team: Wei Chu, Scott Sellars, Phu Nguyen, Xiaogang Gao, Kuo-lin Hsu, and Soroosh Sorooshian**

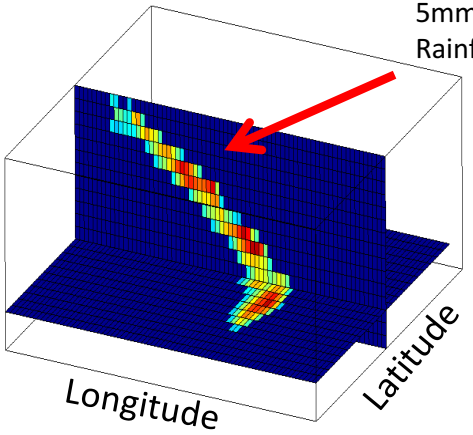
Data Hypercube:



Data

1. 60N-60S, 0-360 lat and long
2. Hourly time step
3. March 1st, 2000 to January 1st, 2011

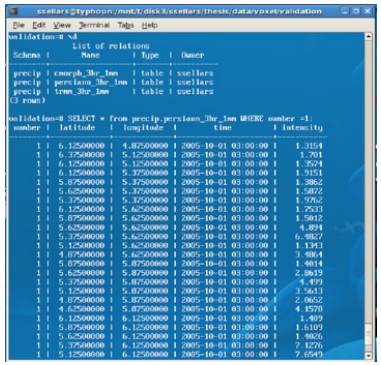
CONNECT: Object Segmentation



Set Object Criteria:

1. Each voxel must have 1mm/hr
2. Each object must exist for 24 hours
3. 6 voxel connections

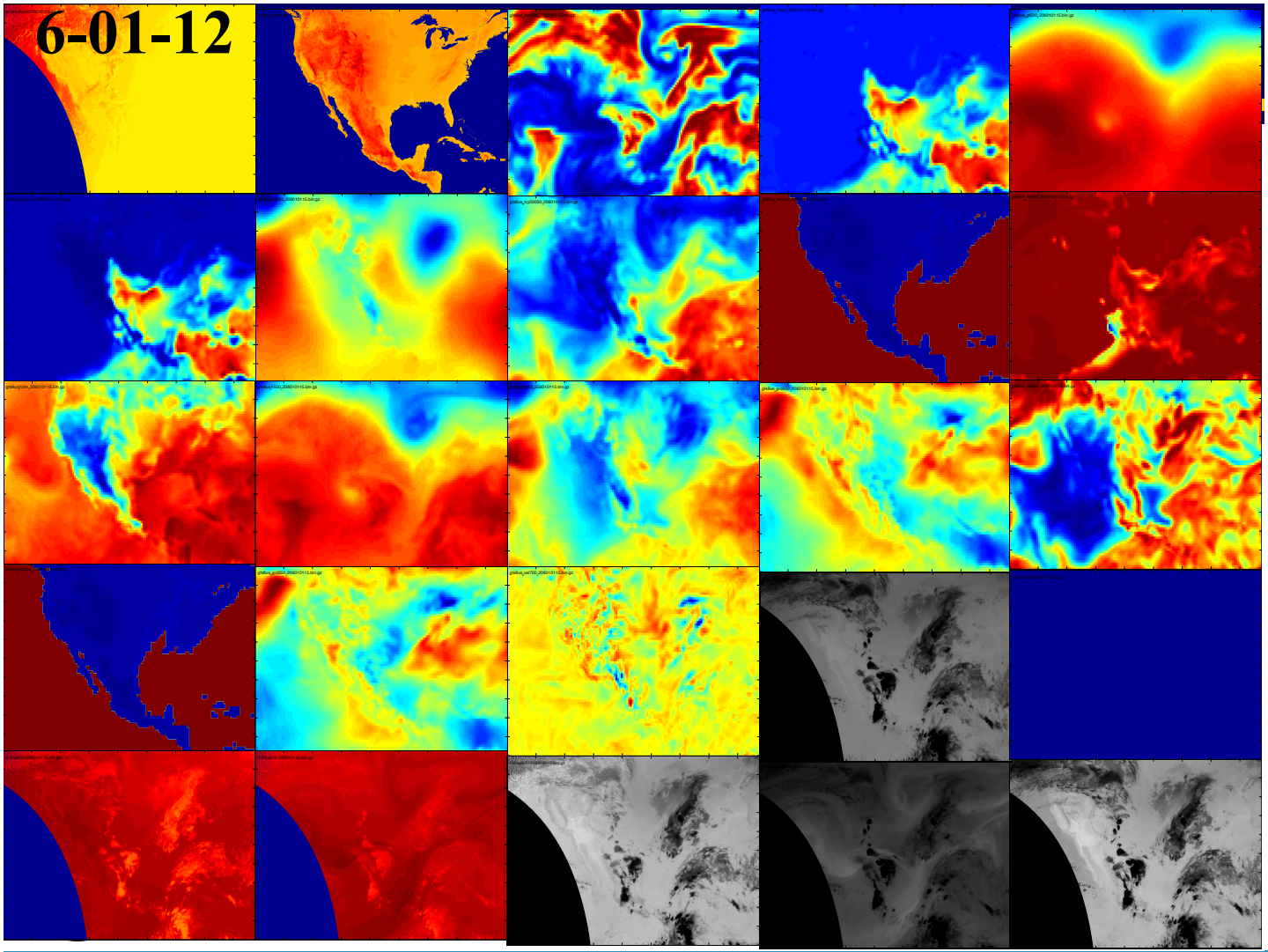
Object Storage (PostgreSQL)



Database Indexes:

1. Object ID Number
2. Latitude (of each voxel in objects)
3. Longitude (of each voxel in objects)
4. Time (hour)

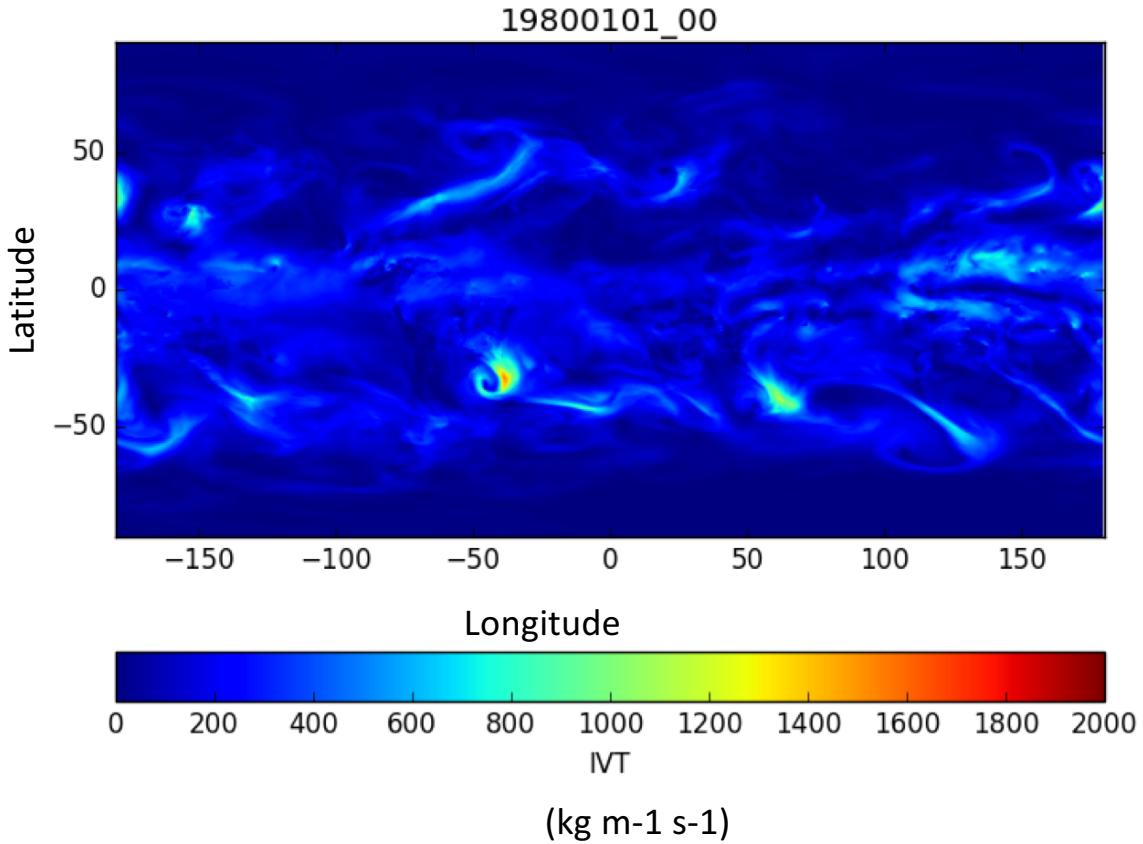
CW3E



Hyper-dimensional data from model reanalysis and satellite based sources for June 1, 2012 over the continental United States. Each image represents a specific atmospheric variable or feature.

CW3E

NASA MERRA2 IVT (kg m-1 s-1)



CW3E

Then I Talked to PRP and Research IT Engineers!

CENIC and Calit2's PRP

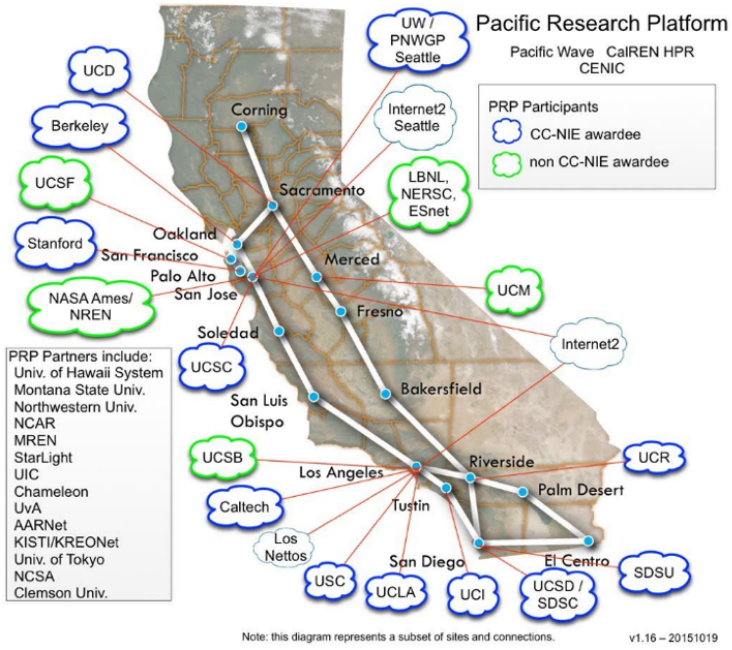


*image courtesy of Larry Smarr

Tools

FIONA – Flash I/O Network Appliance: Linux PCs Optimized for Big Data on DMZs

***SC 2015**

Working with:
John Graham and Joseph Keefe

CW3E

CW3E Usage Pacific Research Platform

- Our data-centric approach was quickly adapted and enhanced to take advantage of the PRP's end-to-end 10- 100Gb/s connections using PRP's FIONA technology

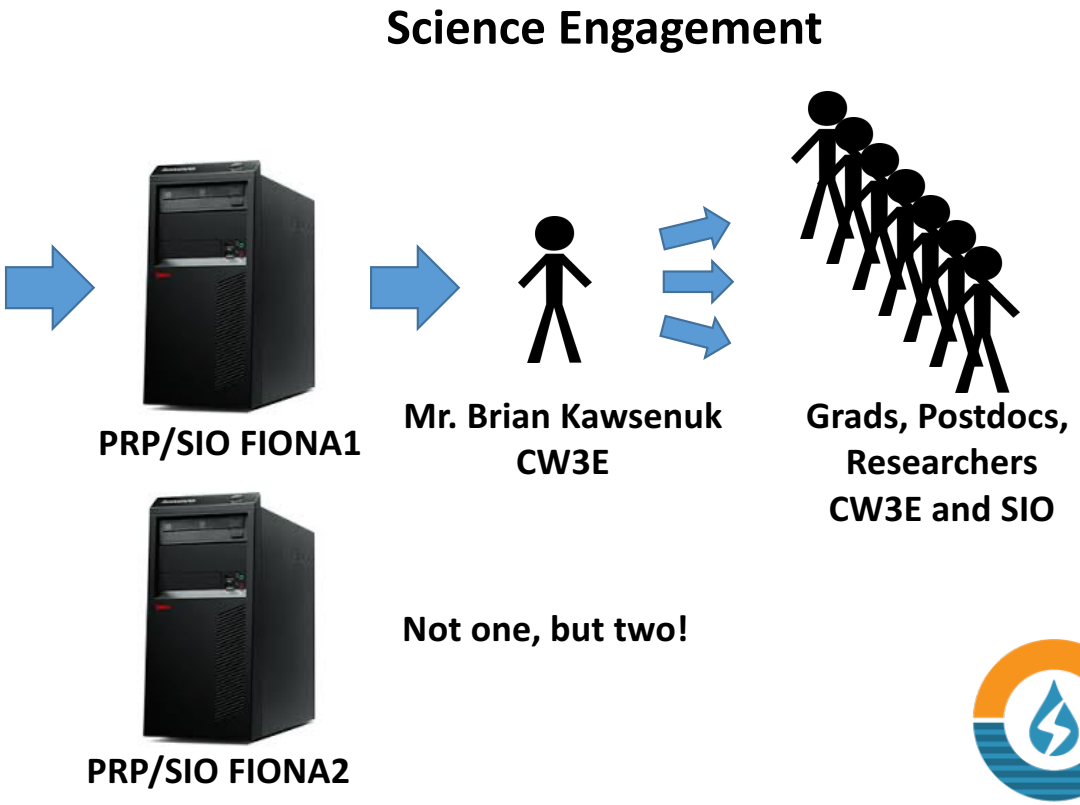
- Demonstrates the application of the multi-institutional rapid data access needs for applying an object characterization and segmentation approach CONNected objECT (CONNECT)

CW3E

Team

- PRP/CONNECT Pilot Project
  - Dr. Scott L. Sellars (Scripps-CW3E)
  - Dr. Phu Ngyuen (UCI-CHRS)
  - Dr. Joulien Tatar (UCI-OIT)
  - Mr. John Graham (QI)
  - Mr. Joseph Keefe (UCSD)
  - Dr. Ilkay Altintas (UCSD-SDSC)
  - Dr. Daniel Crawl (UCSD-SDSC)
  - Dr. Tom DeFanti (UCSD-Calit2)

  - Dr. Larry Smarry (UCSD-Calit2)
    - Director of Calit2
  - Dr. F. Martin Ralph (Scripps-CW3E)
    - Director of CW3E
  - Dr. Soroosh Sorooshian (UCI-CHRS)
    - Directory of CHRS



PAST: Big Data Analysis Pipeline: One Variable

2.4T
Time:
7d 10h 49min

NASA

Download

2.4T to 100GB
Time:
10d 5h 05min

CW3E

Data organization
Variable format

100GB to 50GB
Time:
~1d 14h 00m

CHRS

CONNECT
Segmentation

50GB to 100MB
Time:
~1d 5h 00m

C3WE

CONNECT
Characteristic Calc.

Total time: ~20d 11h 0m

- **Not including data visualization**
- **Not including data mining/machine learning jobs**
- **Assumes we know what we are doing**

CHRS

Data Visualization
And Search

CW3E

Data mining and
Discovery,
Machine Learning

CW3E

First, the Results, in Megabytes, to Date

- Before PRP (**7/16**) NASA **---->** local connection **---->** SIO
  - **10MB/s** download critical NASA data (2.4TB took over 7 days)

- Initial PRP test (**8/16**) NASA **---->** PRP connection **---->** SIO FIONA
  - **40MB/s (4x increase simply using the PRP and a FIONA DTN)**

- PRP/CONNECT (**4/17**) SIO FIONA **---->** via THREDDS **---->** UCI FIONA
  - **559MB/s (56x increase)**
  - **Could be as much as 1896MB/s between FIONAs using Globus (almost 200x better) based on local testing**

CW3E

Developing a new workflow!

Pacific Research Platform (10-100 Gb/s)

UC, Irvine

GPUs

Calit2's FIONA

SDSC's COMET

GPUs

Calit2's FIONA

UC, San Diego

CW3E

Conclusions

- None of this would be possible without engagement with UCSD IT staff and engineers
- Using PRP network via FIONA
  - **Download speed from NASA increased 4x** (40MB/s which is 4x faster than the 10MB/s standard connection previously being used by researchers at SIO).
  - **Removing data transfer as a limiting constraint changes everything!**
  - Researchers and scientists need to rethink how we are approaching data transfer and analysis
- PRP/CW3E/SIO "Big Data and the Earth Sciences: Grand Challenges Workshop"
  - May 31st to June 2nd, Four keynote lectures
  - 75+ registered participants

CW3E

PRP/CONNECT Pilot Project

Dr. Scott L. Sellars (Scripps-CW3E)
Dr. Phu Ngyuen (UCI-CHRS)
Dr. Joulien Tatar (UCI-OIT)
Mr. John Graham (QI)
Mr. Joseph Keefe (UCSD)
Dr. Ilkay Altintas (UCSD-SDSC)
Dr. Daniel Crawl (UCSD-SDSC)
Dr. Tom DeFanti (UCSD-Calit2)

Dr. Larry Smarry (UCSD-Calit2)
Director of Calit2
Dr. F. Martin Ralph (Scripps-CW3E)
Director of CW3E
Dr. Soroosh Sorooshian (UCI-CHRS)
Directory of CHRS

Thank you!

scottsellars@ucsd.edu

Funding Acknowledgements:
NSF The Pacific Research Platform Project #ACI-1541349
California Department of Water Resources
NASA Earth and Space Science Fellowship 2014 (NESSF)
Cooperative Institute for Climate Studies (CICS) (NOAA award NA09NES4400006)
U.S. Army Research Office (Award W911NF-11-1-0422)
NASA (award NNS09AO67G)

CW3E