"High Performance Cyberinfrastructure for Data-Intensive Research"

Distinguished Lecture

UC Riverside

October 18, 2013

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology

Harry E. Gruber Professor,

Dept. of Computer Science and Engineering

Jacobs School of Engineering, UCSD

Abstract

With the increasing number of digital scientific instruments and sensornets available to university researchers, the need for a high performance cyberinfrastructure (HPCI), separate from the shared Internet, is becoming necessary. The backbone of such an HPCI are dedicated wavelengths of light on optical fiber, typically with speeds of 10Gbps or 10,000 megabits/sec, roughly 1000x the speed of the shared Internet. We are fortunate in California to have one of the most advanced optical state networks, the CENIC research and education network. I will describe future extensions of the CENIC backbone to enable a wide range of disciplinary Big Data research. One extension involves building optical fiber "Big Data Freeways" on UC campuses, similar to the NSF-funded PRISM network now being deployed on the UCSD campus, to feed the coming 100Gbps CENIC campus connections. These Freeways connect on-campus end users, compute and storage resources, and data-generating devices, such as scientific instruments, with remote Big Data facilities. I will describe uses of PRISM ranging from particle physics to biomedical data to climate research. The second type of extension is high performance wireless networks to cover the rural regions of our counties, similar to the NSF-funded High Performance Wireless Research and Education Network (HPWREN) currently deployed in San Diego and Imperial counties. HPWREN has enabled data-intensive astronomy observations, wildfire detection, first responder connectivity, Internet access to Native American reservations, seismic networks, and nature observatories.

My Previous Lecture at UC Riverside Was in 2003- This is a Decade-Later Update

UC San Diego's Larry Smarr to Give Distinguished Lecture on Building the Digital University



Prof. Larry Smarr of the department of computer science and engineering at UC San Diego. Smarr will give a lecture entitled "Building the Digital University -- The Emerging Knowledge Infrastructure" on April 16, 2003, at UC Riverside.

UCR | Today
New News, New Look
GO >>

The Data-Intensive Discovery Era Requires High Performance Cyberinfrastructure

- Growth of Digital Data is Exponential
  - "Data Tsunami"
- Driven by Advances in Digital Detectors, Computing, Networking, & Storage Technologies
- Shared Internet Optimized for Megabyte-Size Objects
- Need Dedicated Photonic Cyberinfrastructure for Gigabyte/Terabyte Data Objects
- Finding Patterns in the Data is the New Imperative
  - Data-Driven Applications
  - Data Mining
  - Visual Analytics
  - Data Analysis Workflows



SDSC

Source: SDSC

it²

The White House Announcement Has Galvanized U.S. Campus CI Innovations

News

White House launches big data R&D effort

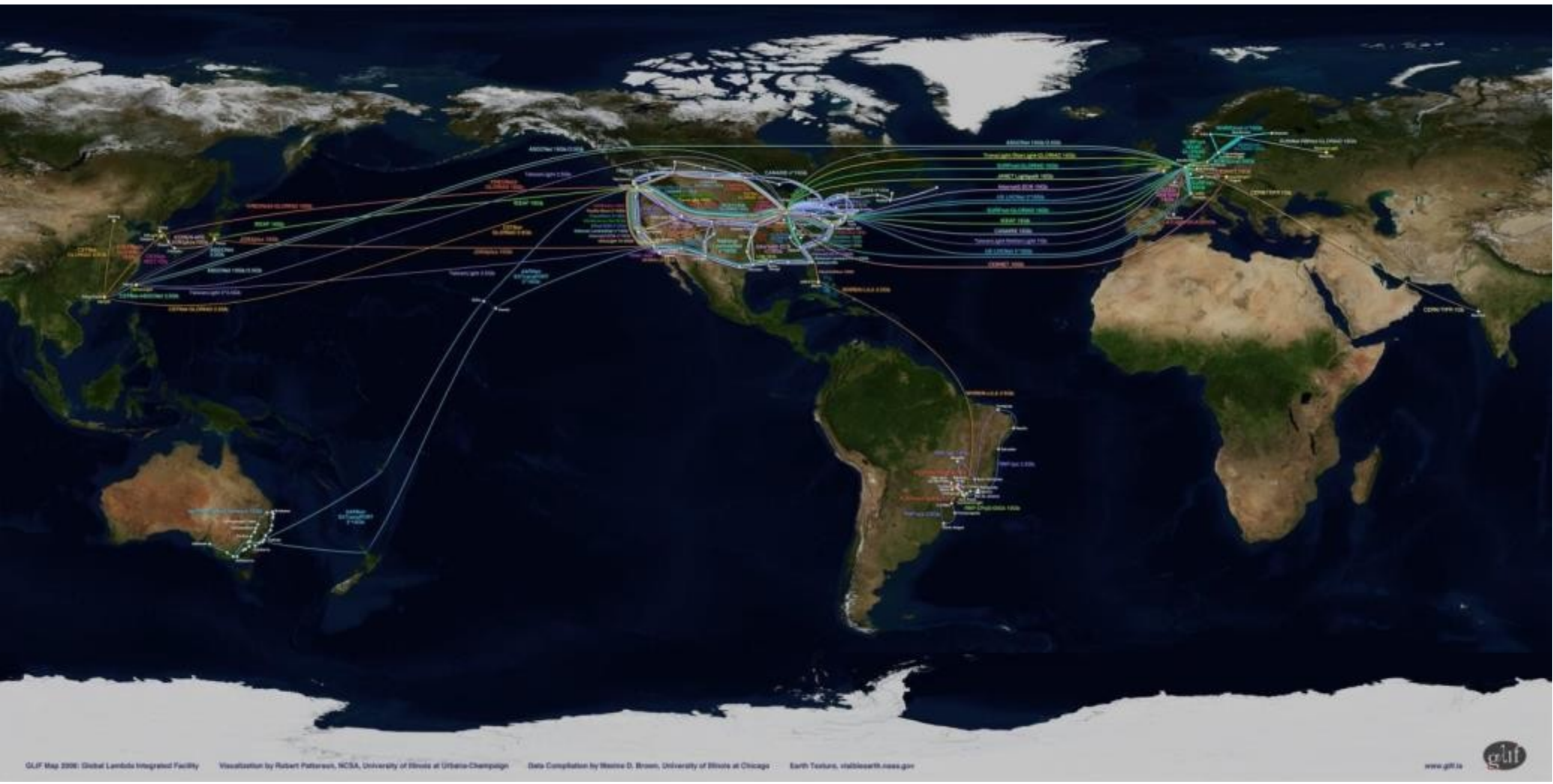
NSF and DARPA are among the agencies looking to harness the power of large volumes of data

By Grant Gross

March 29, 2012 12:34 PM ET

5 Comments + Briefcase What's this?

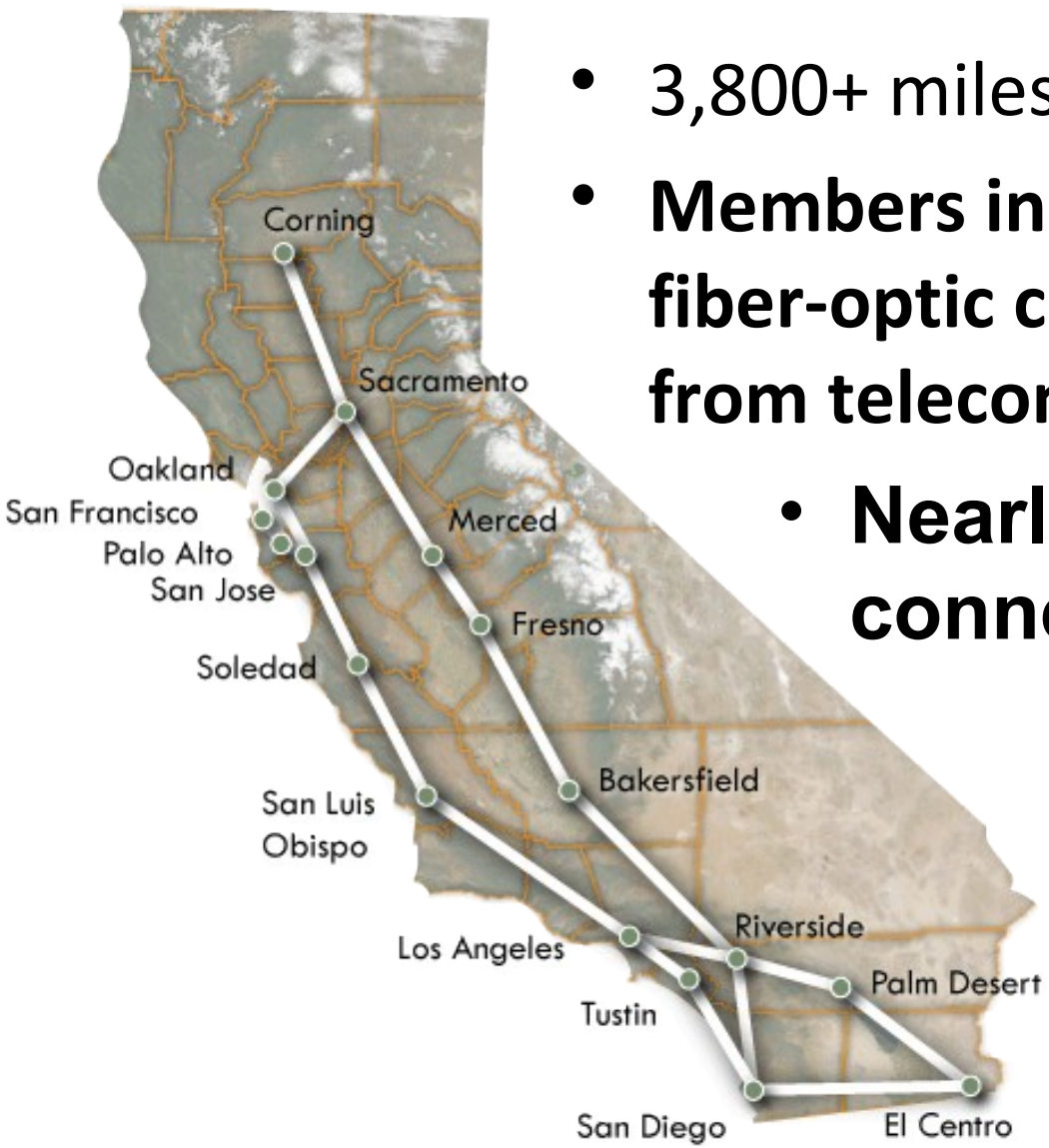
Global Innovation Centers are Being Connected with 10,000 Megabits/sec Clear Channel Lightpaths



100 Gbps Commercially Available;
Research on 1 Tbps

Source: Maxine Brown, UIC and Robert Patterson, NCSA

Corporation For Education Network Initiatives In California (CENIC)



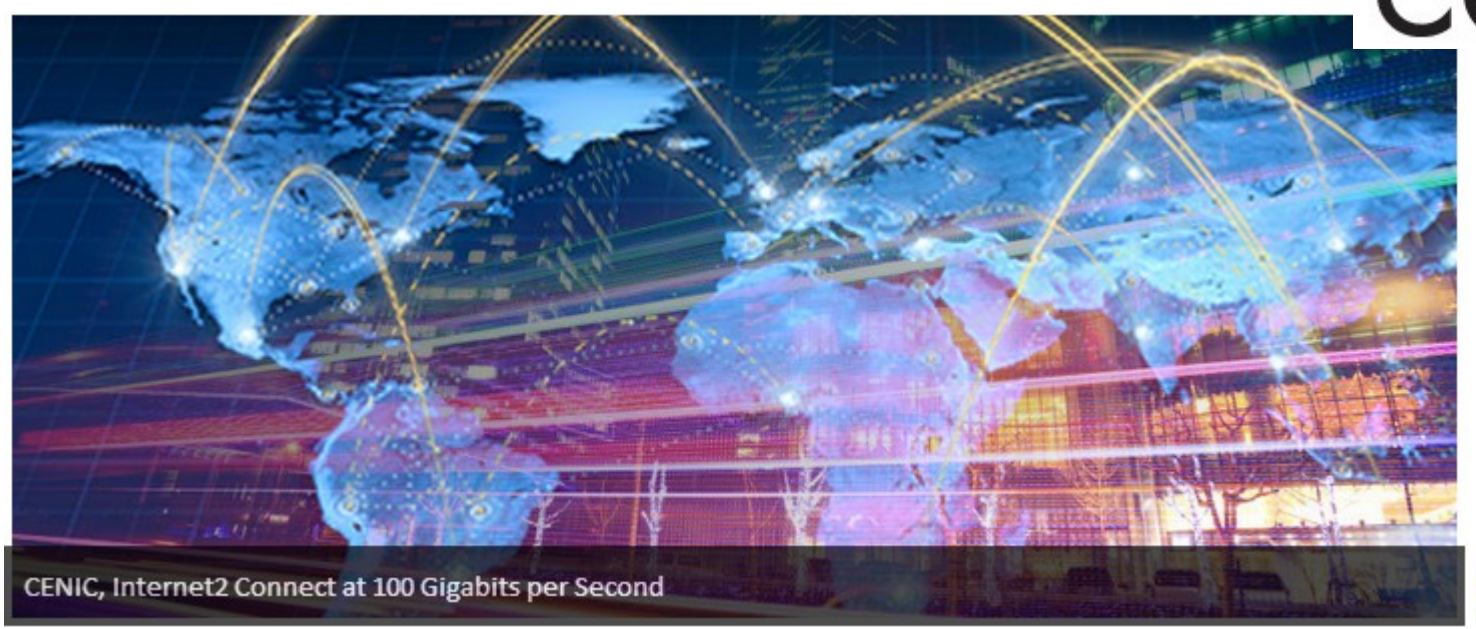
- 3,800+ miles of optical fiber
- **Members in all 58 counties connect via fiber-optic cable or leased circuits from telecom carriers**
- **Nearly *10,000* sites connect to CENIC**
- *10,000,000+* Californians use CENIC each day
- Governed by members on the segmental level

CENIC is Rapidly Moving to Connect at 100 Gbps



CENIC and ESnet Connected at 100 Gigabits per Second

cenic


CENIC, Internet2 Connect at 100 Gigabits per Second

it²

How Can a Campus Connect Its Researchers, Instruments, and Clusters at 10-100 Gbps?

- **Strategic Recommendation to the NSF #3: “**
  - **NSF should create a new program funding high-speed (currently 10 Gbps) connections from campuses to the nearest landing point for a national network backbone. The design of these connections must include support for dynamic network provisioning services and must be engineered to support rapid movement of large scientific data sets."**
  - **- pg. 6, NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging, Final Report, March 2011**
    - **www.nsf.gov/od/oci/taskforces/TaskForceReport_CampusBridging.pdf**

- **Led to Office of Cyberinfrastructure RFP March 1, 2012**
- **NSF’s Campus Cyberinfrastructure – Network Infrastructure & Engineering (CC-NIE) Program**
  - **1st Area: Data Driven Networking Infrastructure for the Campus and Researcher**
  - **2nd Area: Network Integration and Applied Innovation**



Examples of CC-NIE Winning Proposals In California

- **UC Davis**
  - Develop Infrastructure for Managing/Transfer/Analysis of Big Data
    - LSST (30TB/day), GENOME, and More Including Social Sciences
  - Provide Data to Campus Research Groups that Perform Network-Related Research (Security & Performance)
  - Create a Software Defined Network (SDN) – Use OpenFlow
  - Upgrade Intra-Campus and CENIC Connections
- **San Diego State University**
  - Implementing a Science DMZ through CENIC
  - Balancing Performance and Security Needs
    - Operational Network Use: security > performance
    - Research Network Use: performance > security
- **Stanford University**
  - Develop SDN-Based Private Cloud
  - Connect to Internet2 100G Innovation Platform
  - Campus-wide Sliceable/VIrtualized SDN Backbone (10-15 switches)
    - SDN control and management

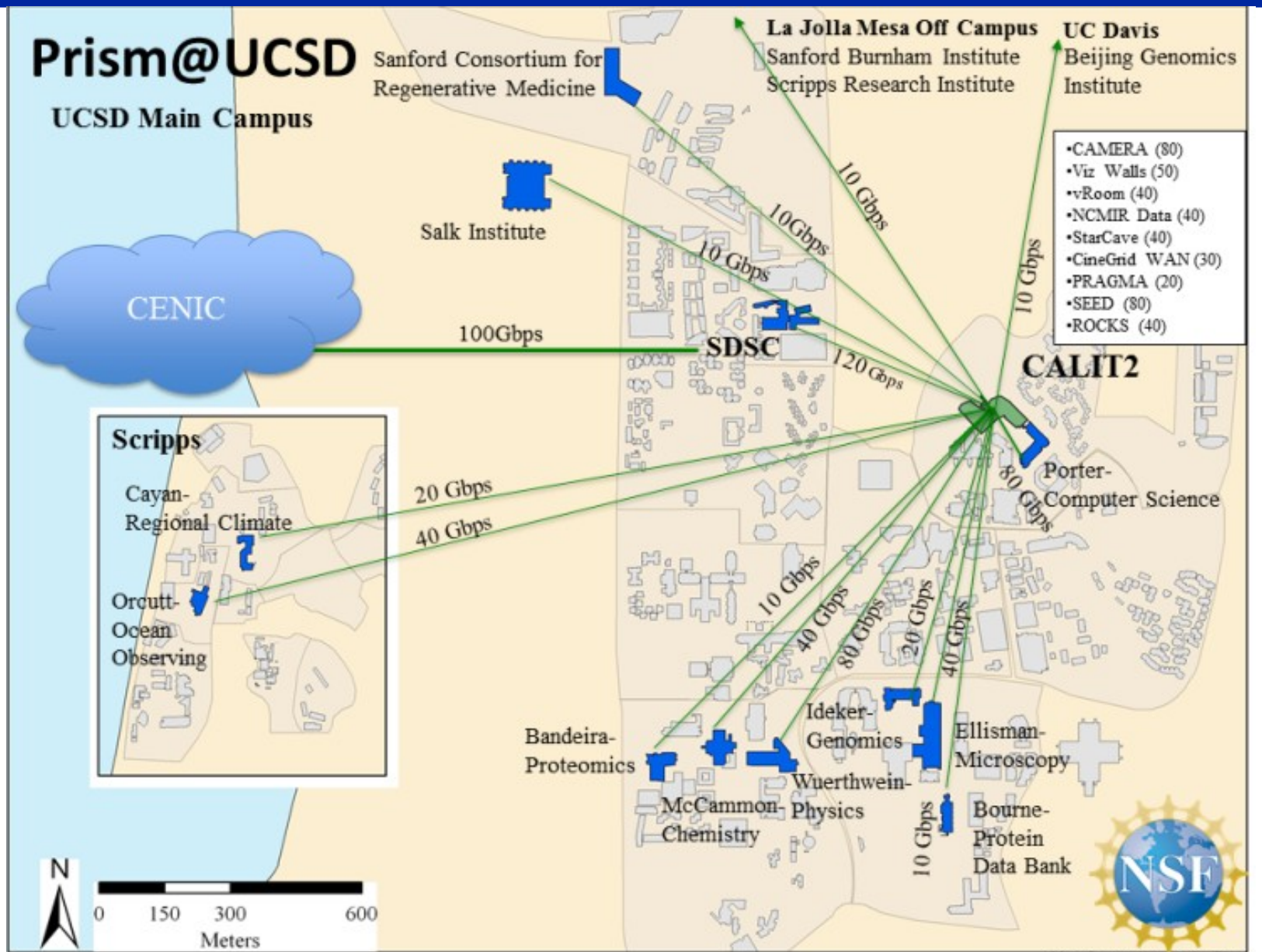
Also USC, Caltech, and UCSD

cenic

Source: Louis Fox, CENIC CEO

NSF

Creating a Big Data Freeway System: Use Optical Fiber with 1000x Shared Internet Speeds



NSF CC-NIE Has Awarded Prism@UCSD Optical Switch
Phil Papadopoulos, SDSC, Calit2, PI

Many Disciplines Beginning to Need Dedicated High Bandwidth on Campus

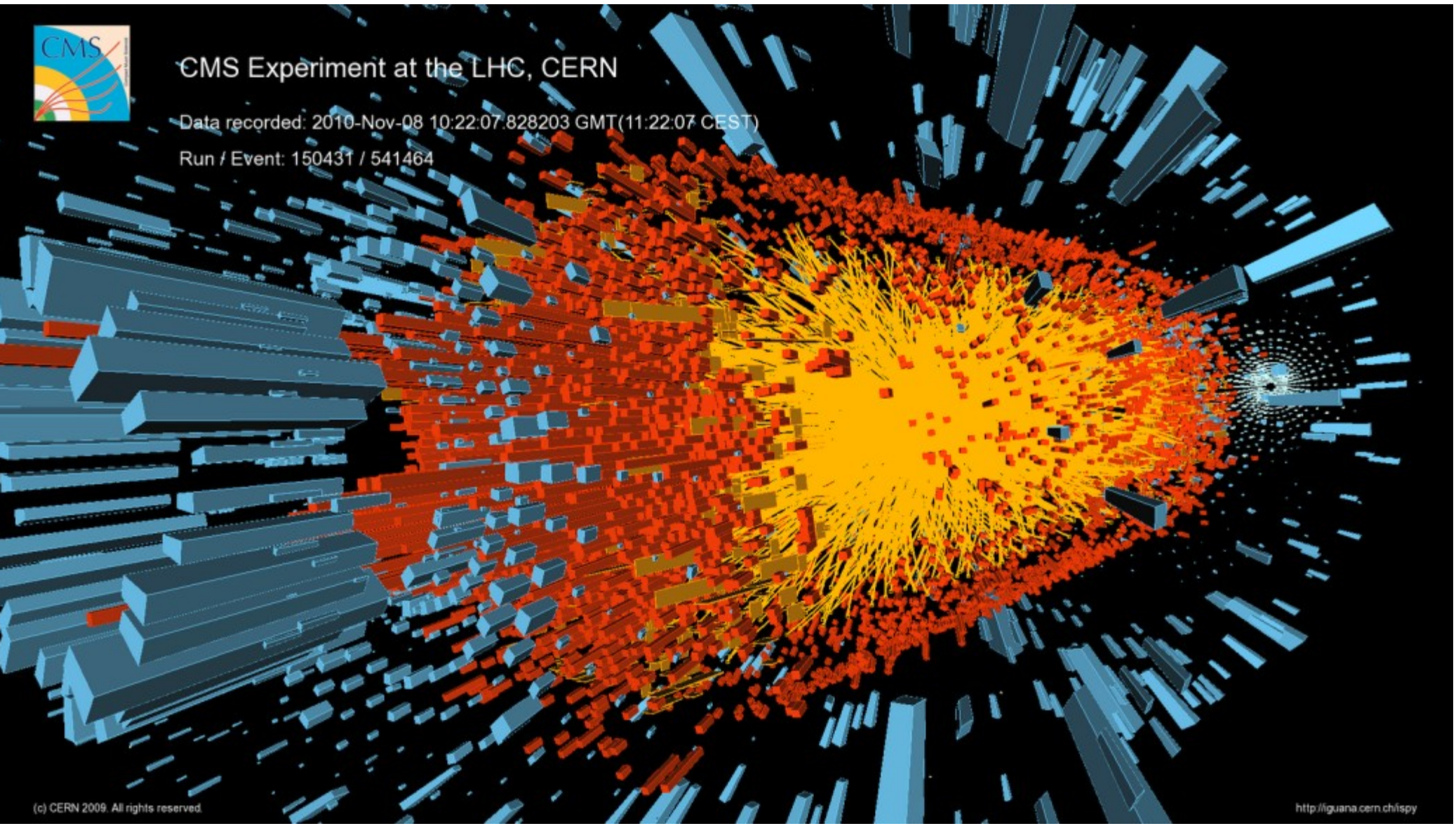
How to Utilize a CENIC 100G Campus Connection

- **Remote Analysis of Large Data Sets**
  - **Particle Physics**
- **Connection to Remote Campus Compute & Storage Clusters**
  - **Microscopy and Next Gen Sequencers**
- **Providing Remote Access to Campus Data Repositories**
  - **Protein Data Bank and Mass Spectrometry**
- **Enabling Remote Collaborations**
  - **National and International**

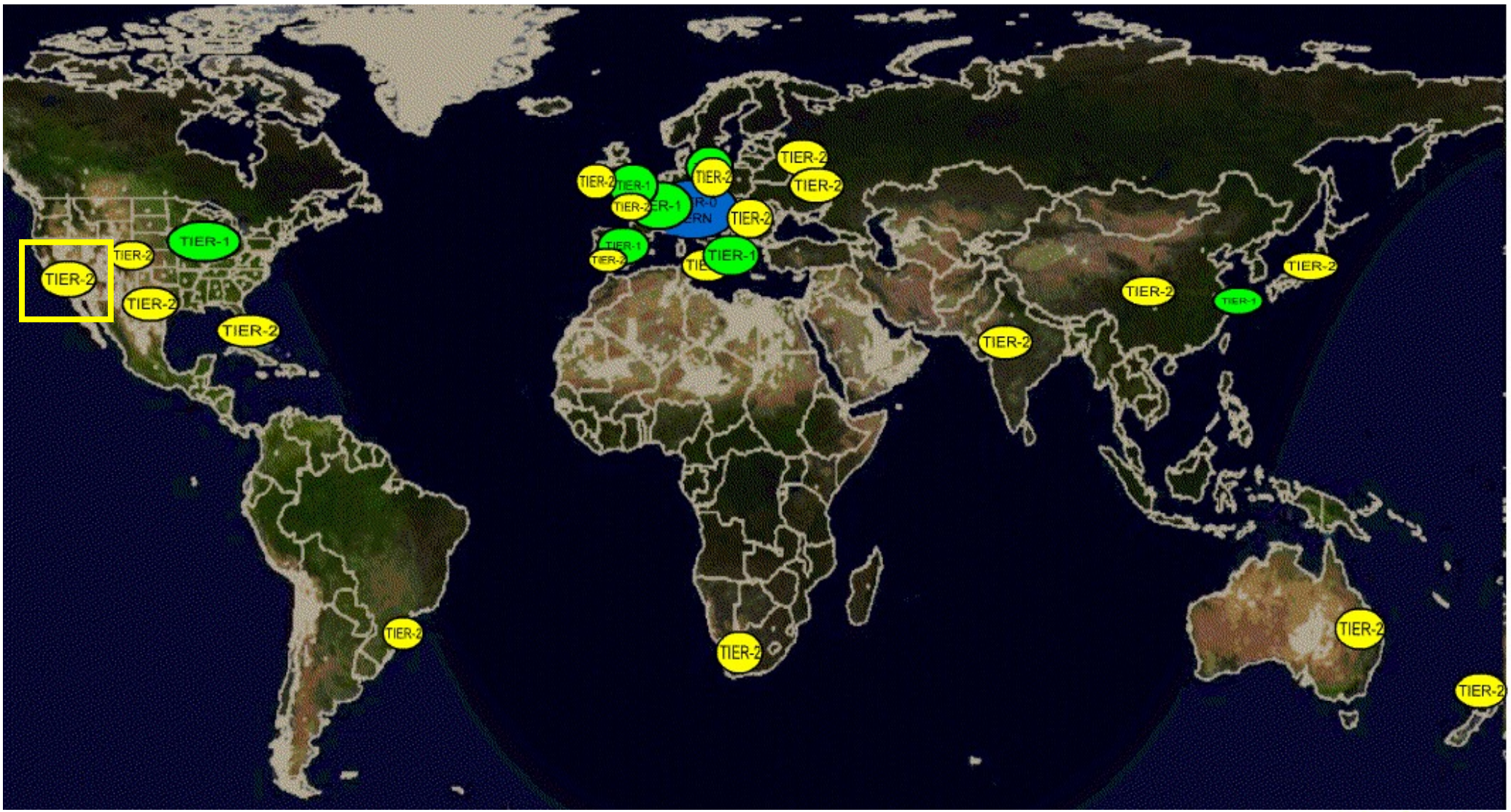
UC San Diego

it²

CERN's CMS Experiment Generates Massive Amounts of Data



UCSD is a Tier-2 LHC Data Center: CMS Flow into UCSD Physics Dept. Peaks at 2.4 Gbps



Source: Frank Wuerthwein, Physics UCSD

Planning for climate change in California

substantial shifts on top of already high climate variability

UCSD Campus Climate Researchers Need to Download Results from Remote Supercomputer Simulations to Make Regional Climate Change Forecasts

Dan Cayan

USGS Water Resources Discipline

Scripps Institution of Oceanography, UC San Diego

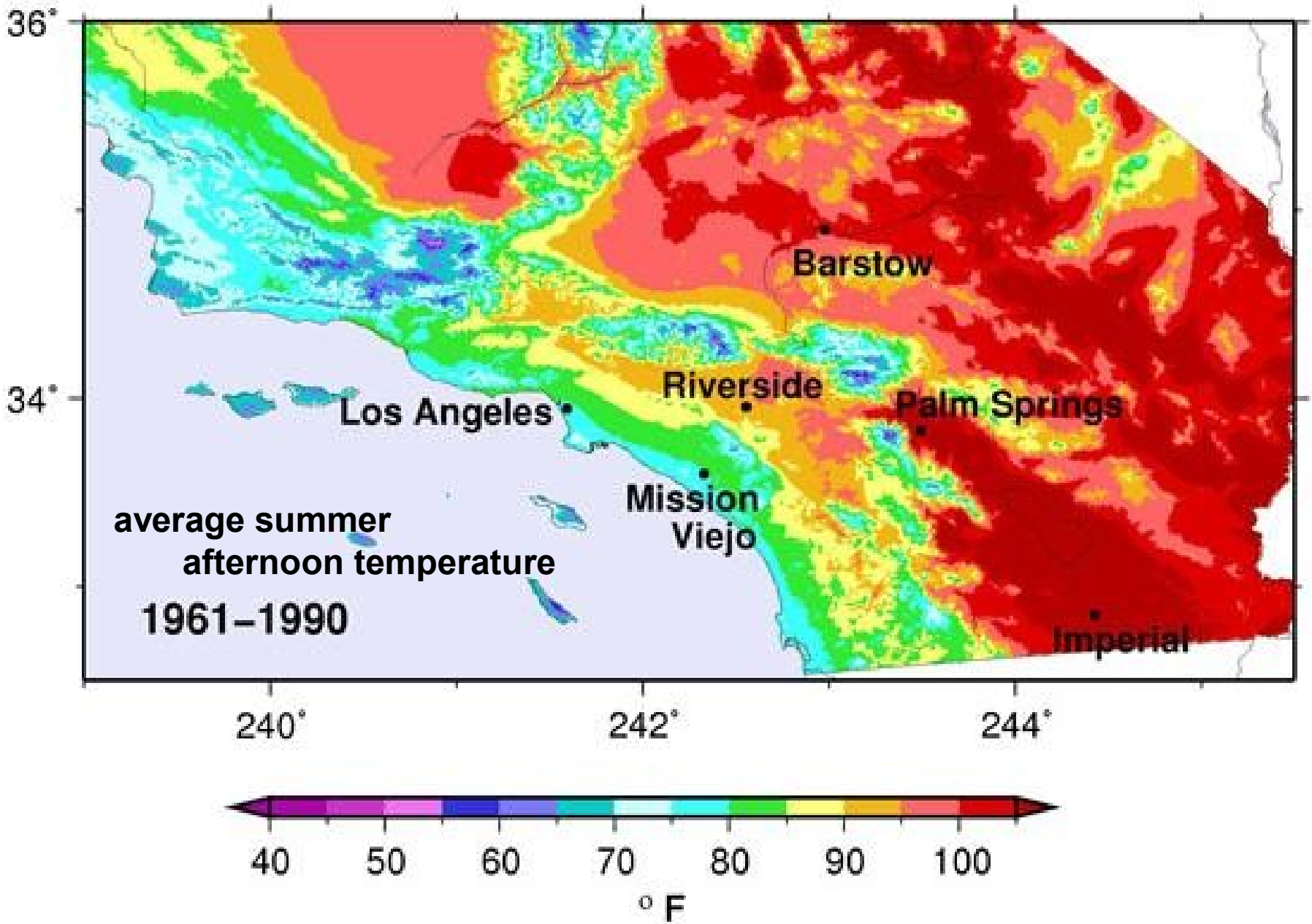
much support from Mary Tyree, Mike Dettinger, Guido Franco and other colleagues

Sponsors:

California Energy Commission

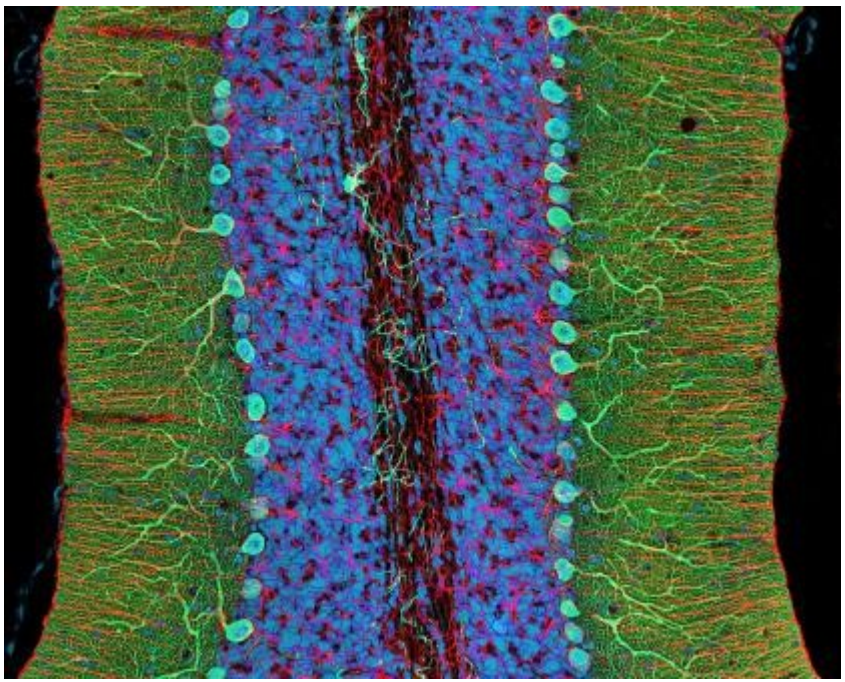
NOAA RISA program

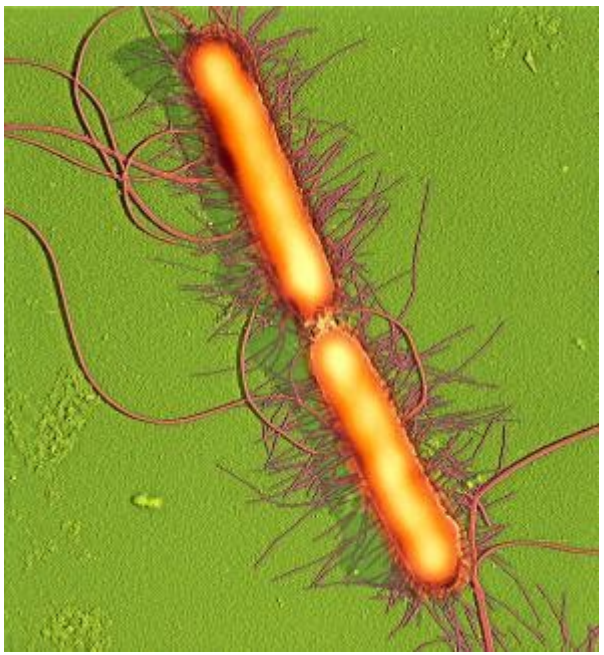
California DWR, DOE, NSF



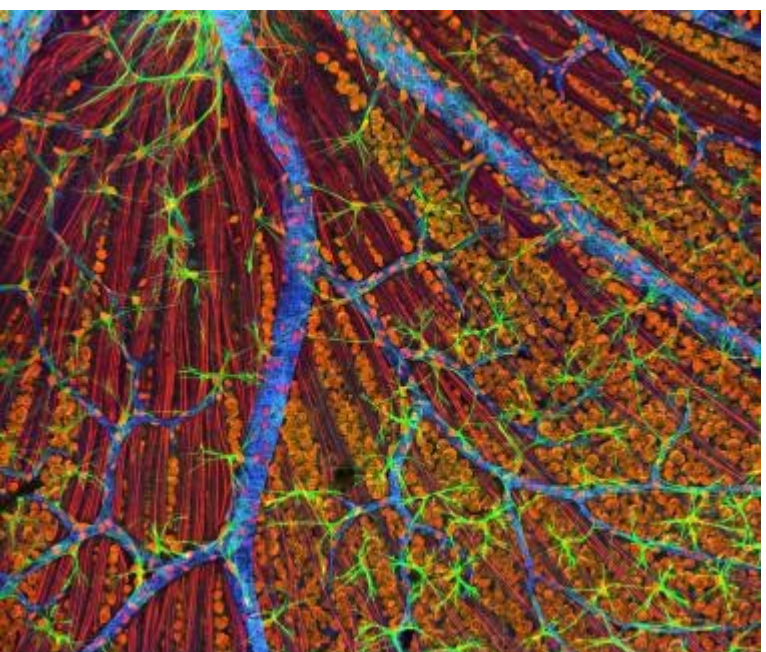
GFDL A2 1km downscaled to 1km
Hugo Hidalgo Tapash Das Mike Dettinger

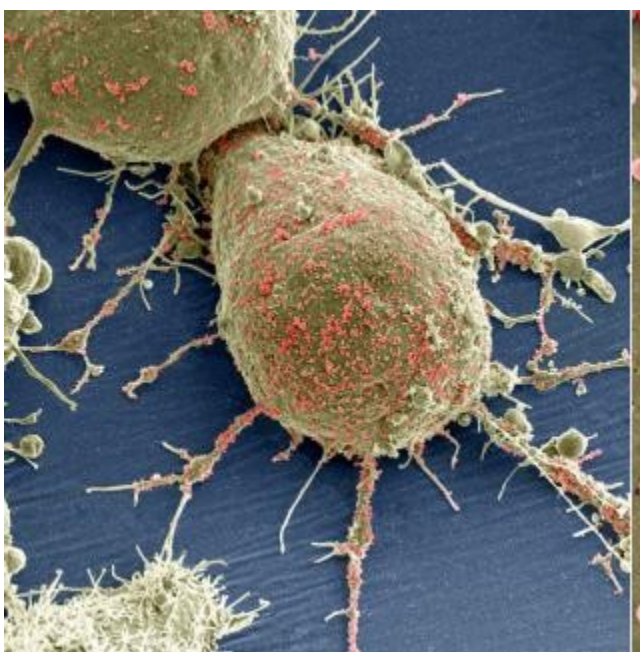
Ultra High Resolution Microscopy Images Created at the National Center for Microscopy Imaging

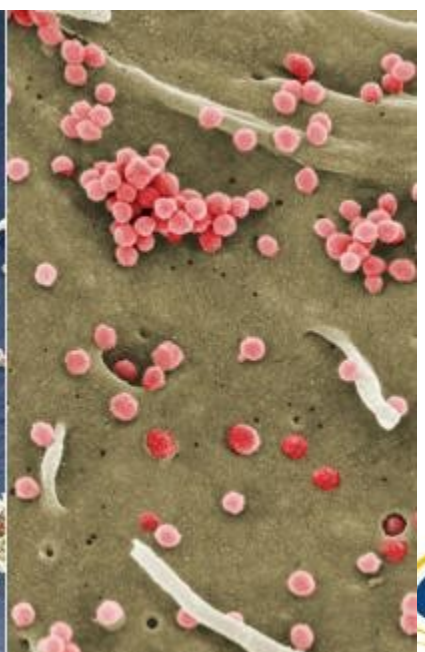




NCMIR







it²

NIH National Center for Microscopy & Imaging Research
Integrated Infrastructure of Shared Resources

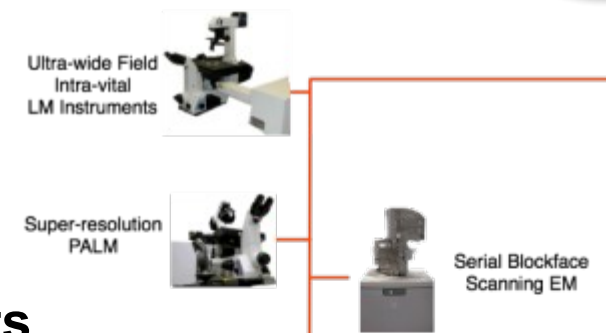
NCMIR



Shared Infrastructure



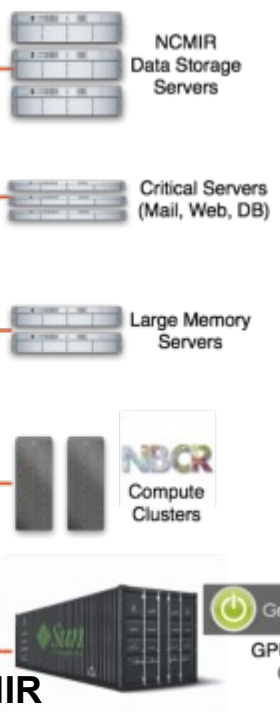
Scientific Instruments



End User Workstations



Local SOM Infrastructure

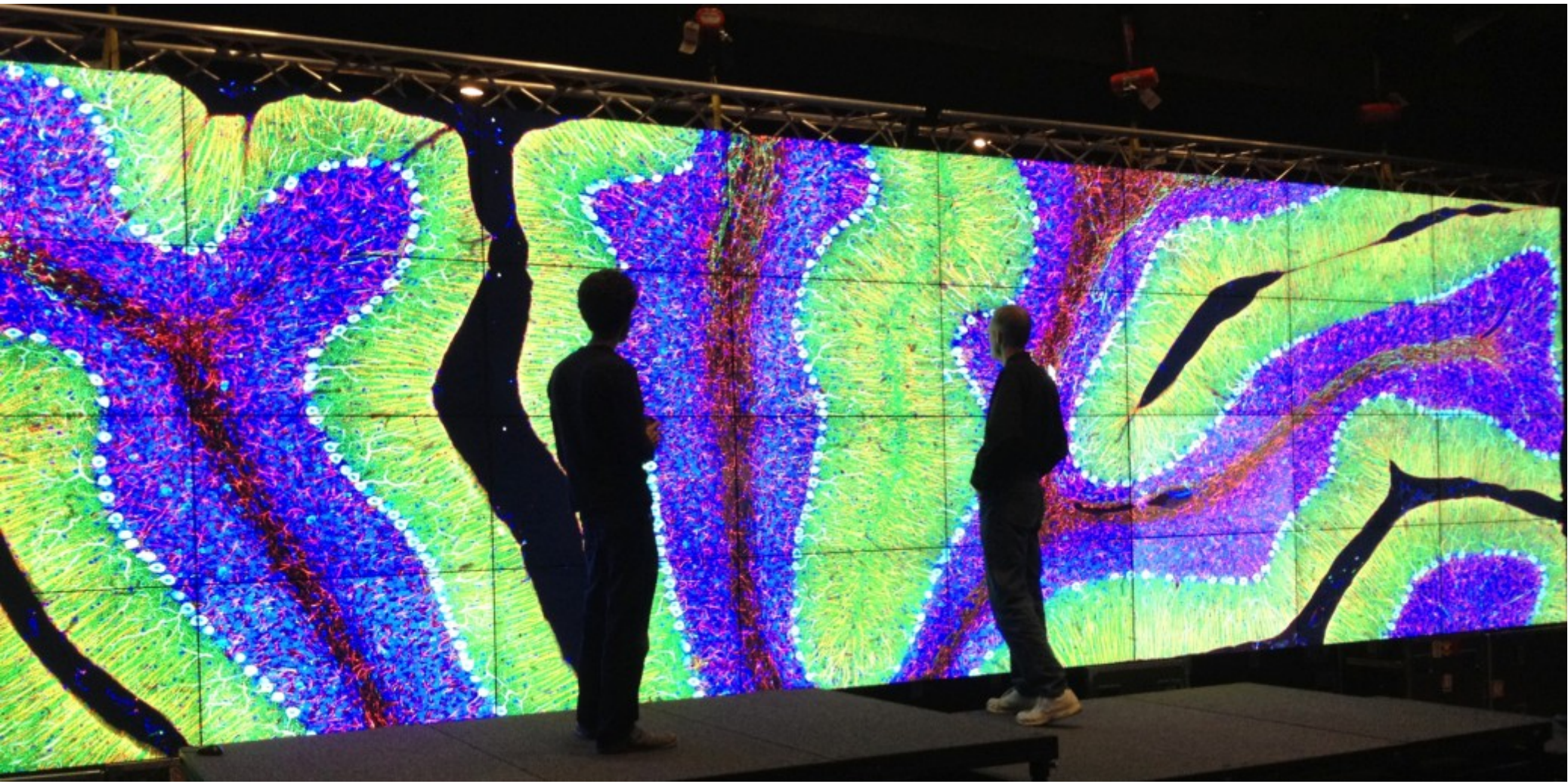


SDSC

Source: Steve Peltier, Mark Ellisman, NCMIR

it²

Using Calit2's VROOM to Explore Confocal Light Microscope Collages of Rat Brains



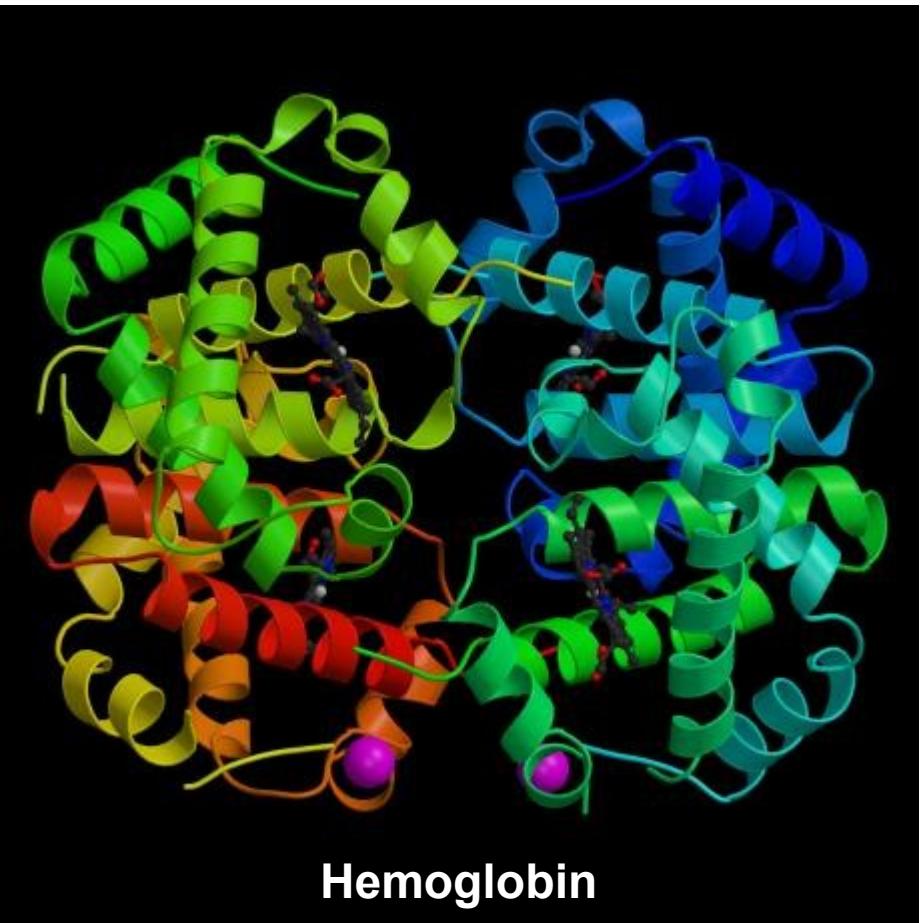


NCMIR

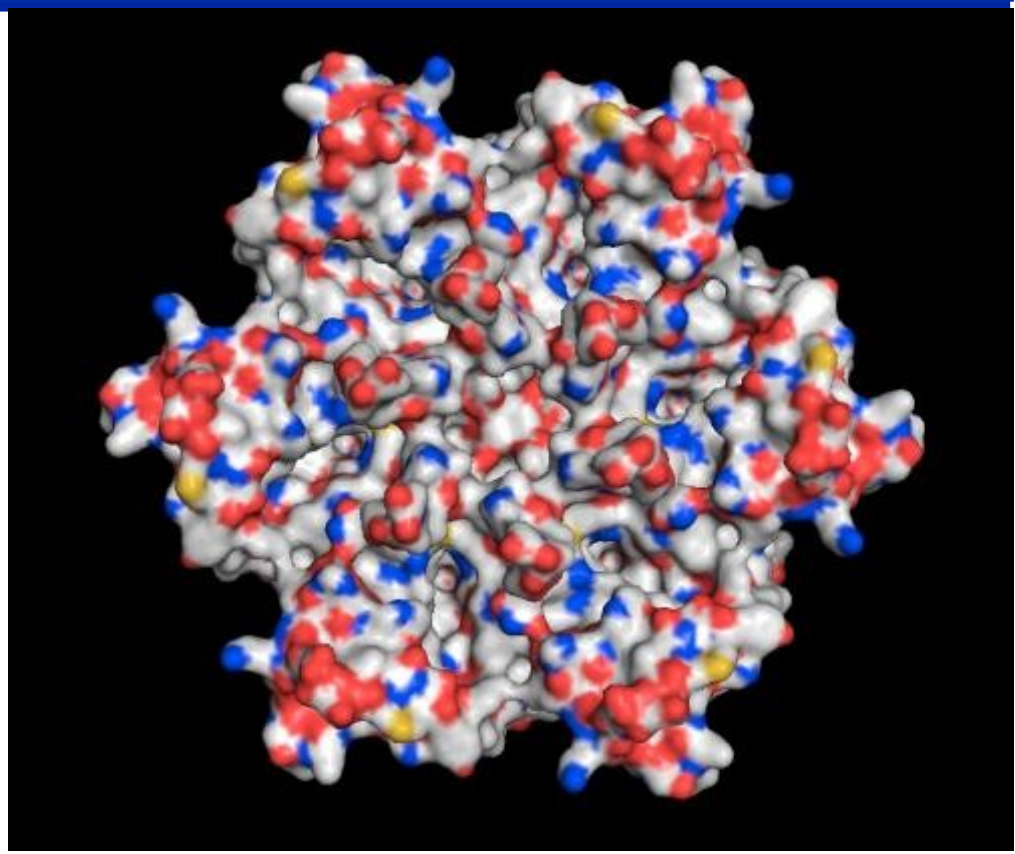
it²

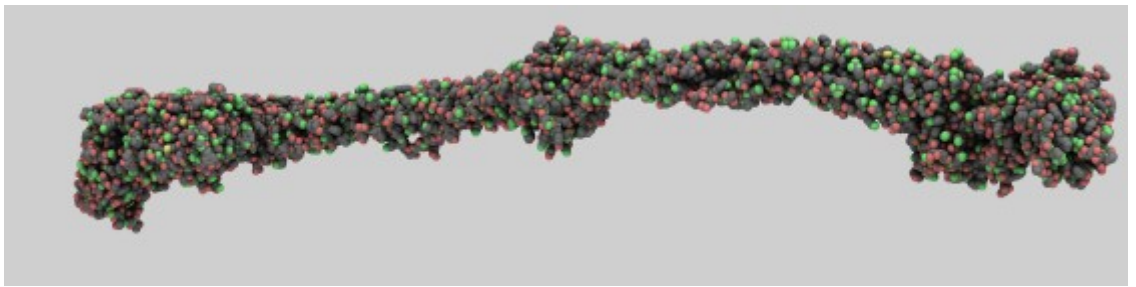
Protein Data Bank (PDB) Needs Bandwidth to Connect Resources and Users

- Archive of experimentally determined 3D structures of proteins, nucleic acids, complex assemblies
- One of the largest scientific resources in life sciences



Hemoglobin



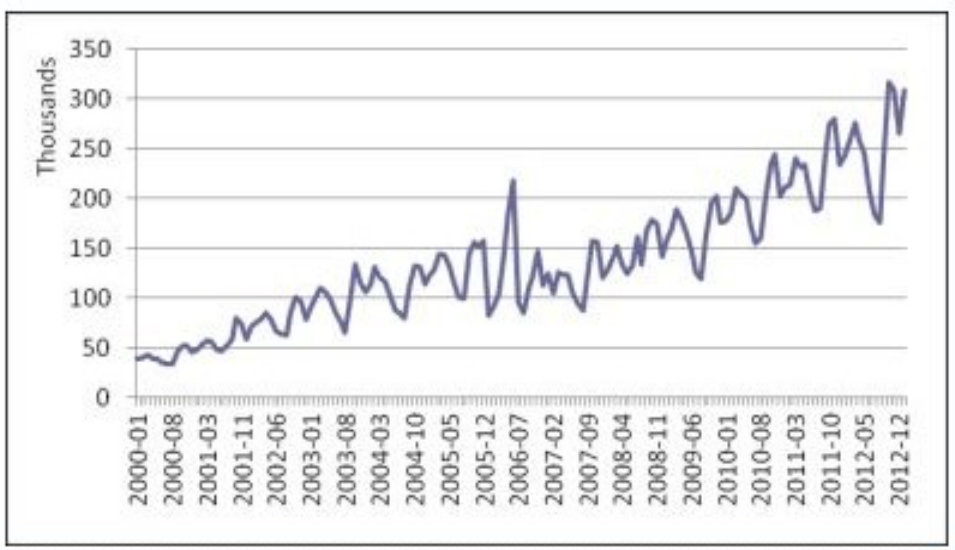


Source: Phil Bourne and Andreas Prlić, PDB

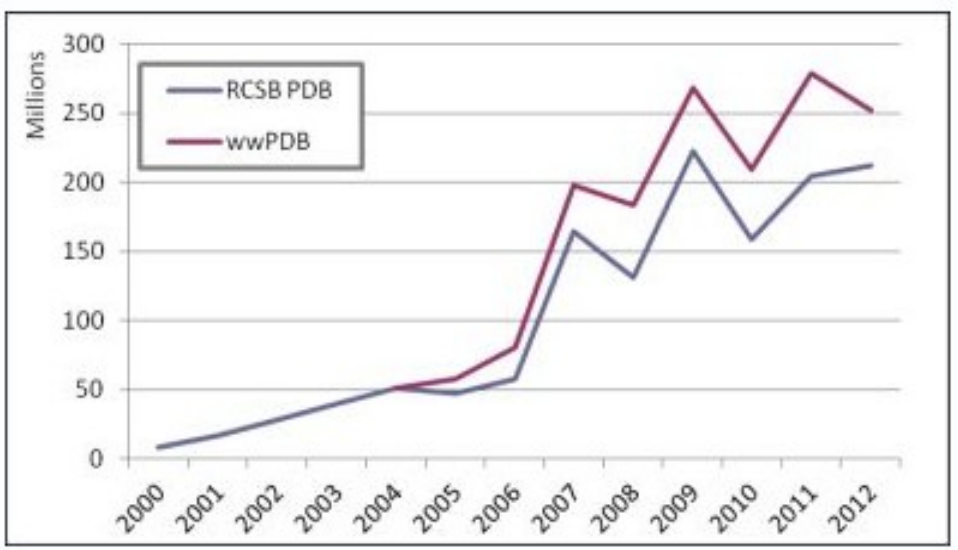
PDB Usage Is Growing Over Time

- More than 300,000 Unique Visitors per Month
- Up to 300 Concurrent Users
- ~10 Structures are Downloaded per Second 7/24/365
- Increasingly Popular Web Services Traffic

RCSB PDB Website Unique Users by Month

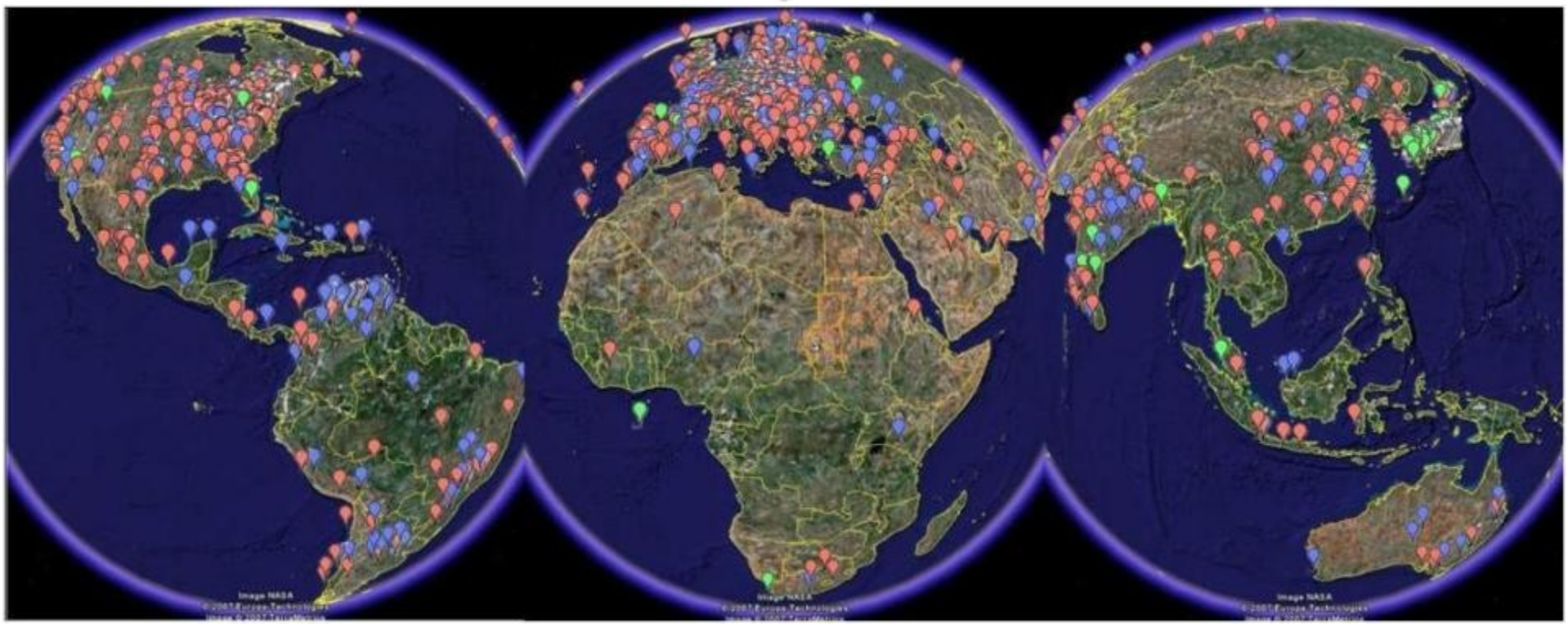


FTP Downloads by Year



Source: Phil Bourne and Andreas Prlić, PDB

2010 FTP Traffic



RCSB PDB
159 million
entry downloads

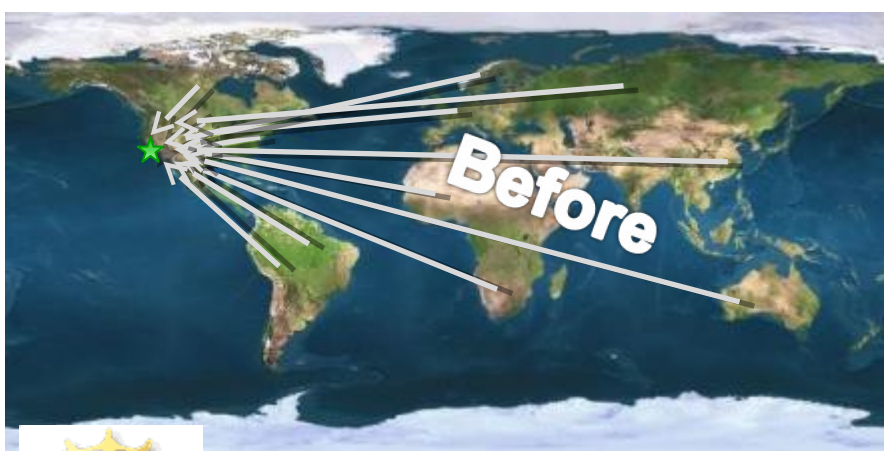
PDBe
34 million
entry downloads

PDBj
16 million
entry download

Source: Phil Bourne and Andreas Prlić, PDB

PDB Plans to Establish Global Load Balancing

- **Why is it Important?**
  - **Enables PDB to Better Serve Its Users by Providing Increased Reliability and Quicker Results**
- **How Will it be Done?**
  - **By More Evenly Allocating PDB Resources at Rutgers and UCSD**
  - **By Directing Users to the Closest Site**
- **Need High Bandwidth Between Rutgers & UCSD Facilities**





Source: Phil Bourne and Andreas Prlić, PDB

Tele-Collaboration for Audio Post-Production
Realtime Picture & Sound Editing Synchronized Over IP





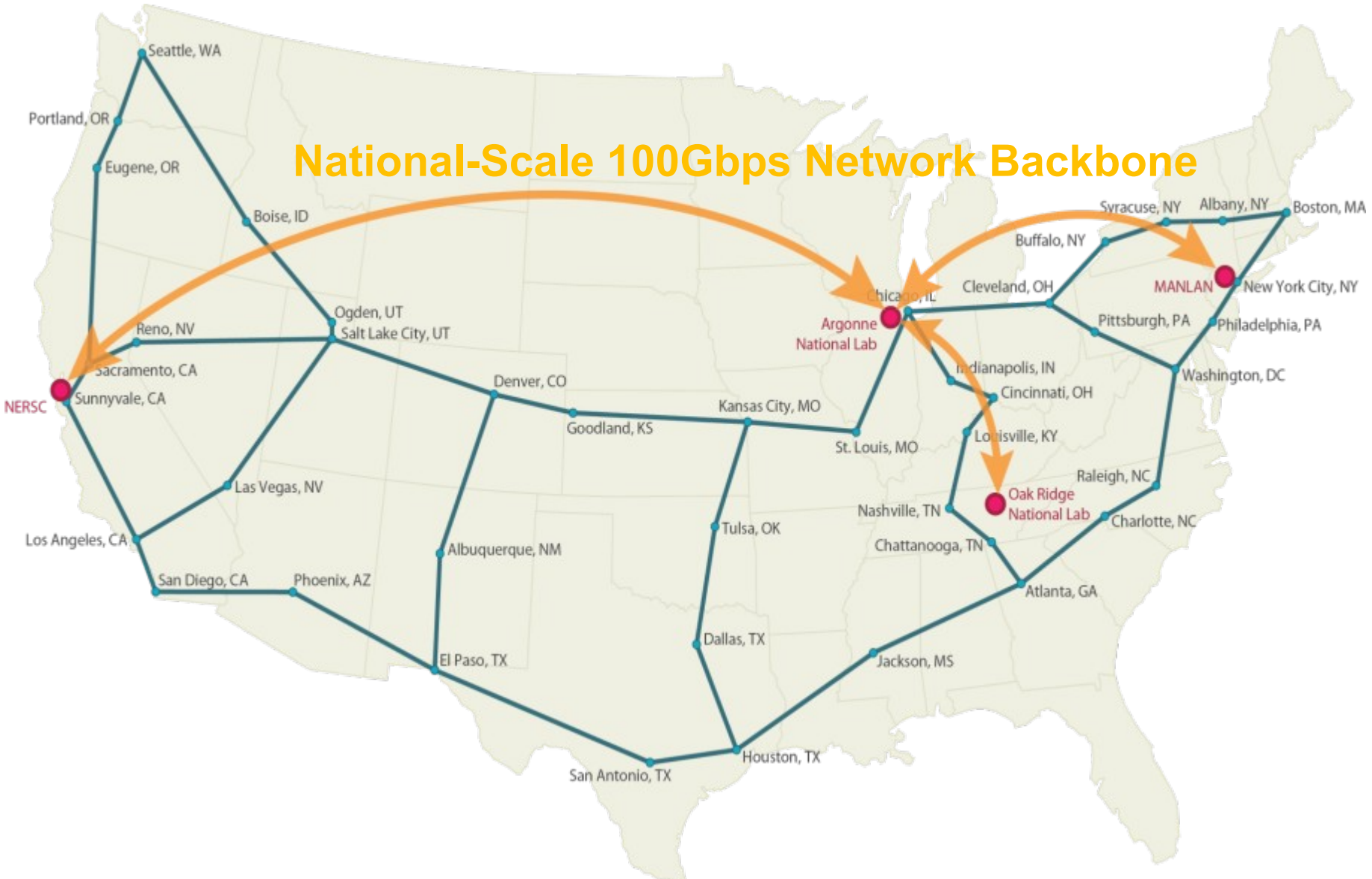
Skywalker Sound@Marin **Calit2@San Diego**

Collaboration Between EVL's CAVE2 and Calit2's VROOM Over 10Gb Wavelength



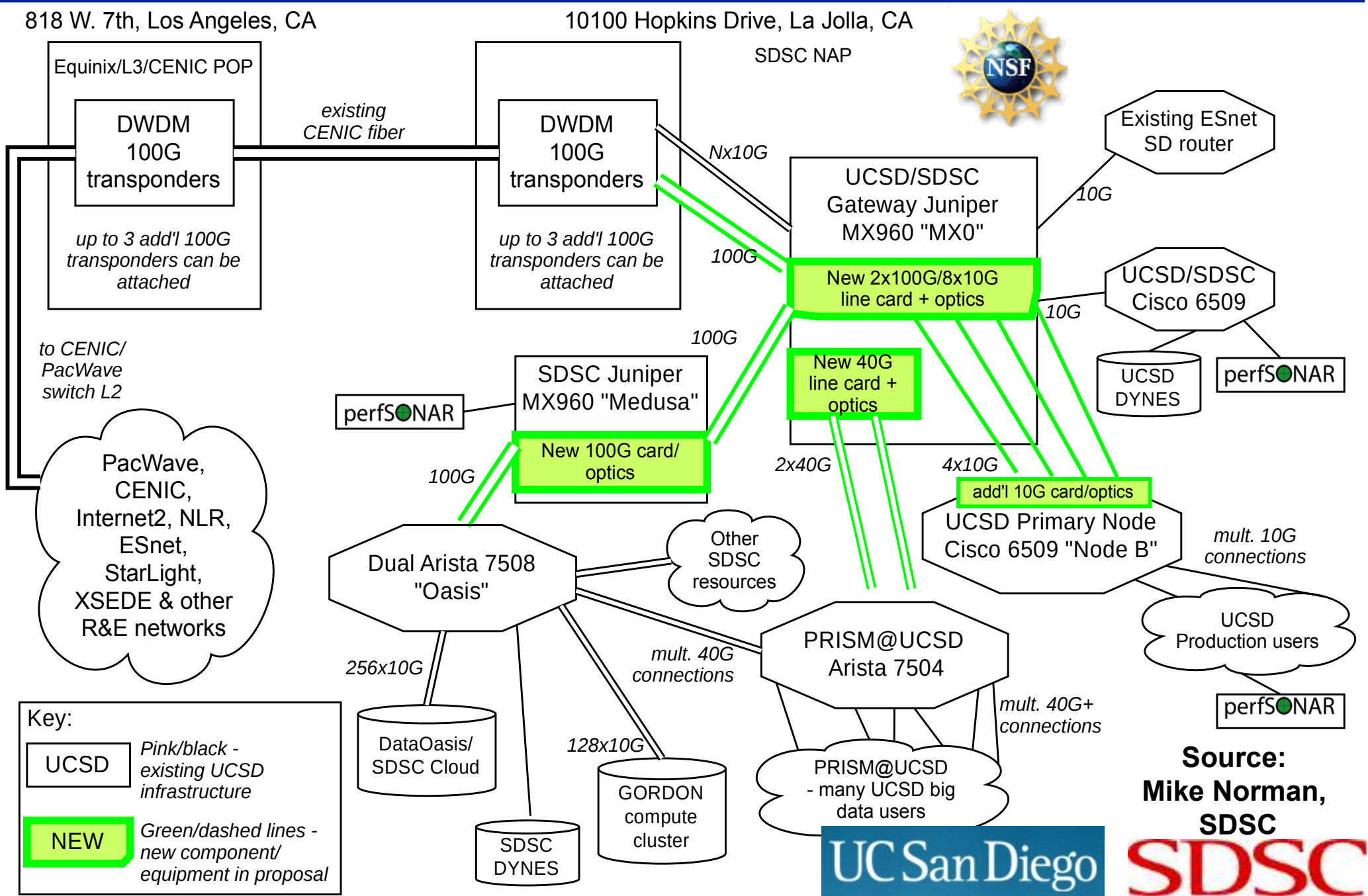
Source: NTT Sponsored ON*VECTOR Workshop at Calit2 March 6, 2013

Partnering Opportunities with DOE: ARRA Stimulus Investment for DOE Esnet 100Gbps

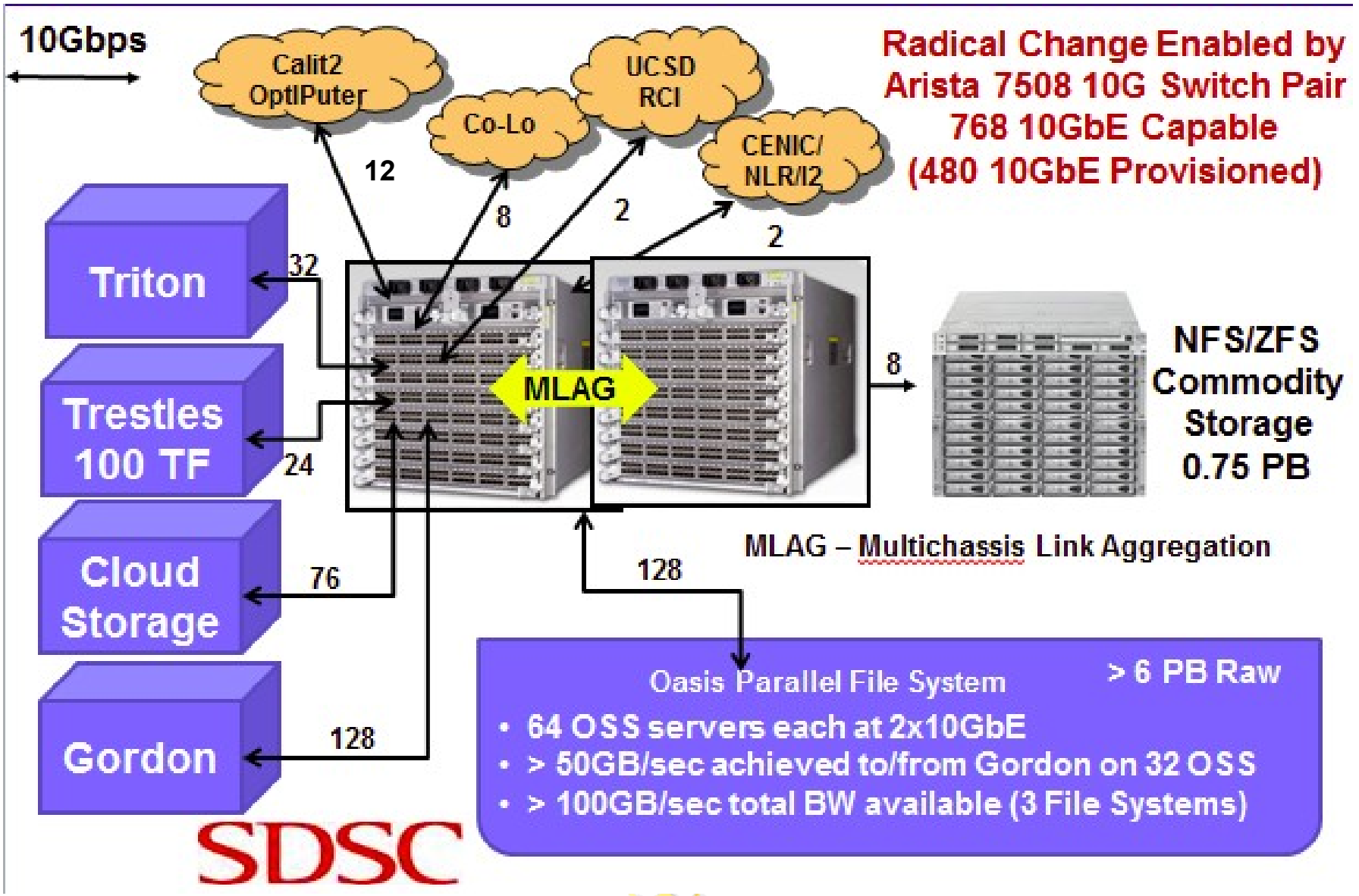


Source: Presentation to ESnet Policy Board

100G Addition CENIC to UCSD--Configurable, High-speed, Extensible Research Bandwidth (CHERuB)



Arista Enables SDSC's Massively Parallel 10G Switched Data Analysis Resource



We Used SDSC's Gordon Data-Intensive Supercomputer to Analyze a Wide Range of Gut Microbiomes

- ~180,000 Core-Hrs on Gordon
  - KEGG function annotation: 90,000 hrs
  - Mapping: 36,000 hrs
    - Used 16 Cores/Node and up to 50 nodes
  - Duplicates removal: 18,000 hrs
  - Assembly: 18,000 hrs
  - Other: 18,000 hrs
- Gordon RAM Required
  - 64GB RAM for Reference DB
  - 192GB RAM for Assembly
- Gordon Disk Required
  - Ultra-Fast Disk Holds Ref DB for All Nodes
  - 8TB for All Subjects



SDSC

Enabled by
a Grant of Time
on Gordon from SDSC
Director Mike Norman

NSF NIH HMP

J. Craig Venter INSTITUTE

CRBS it²

SDSC's Triton Shared Computing Cluster (TSCC)

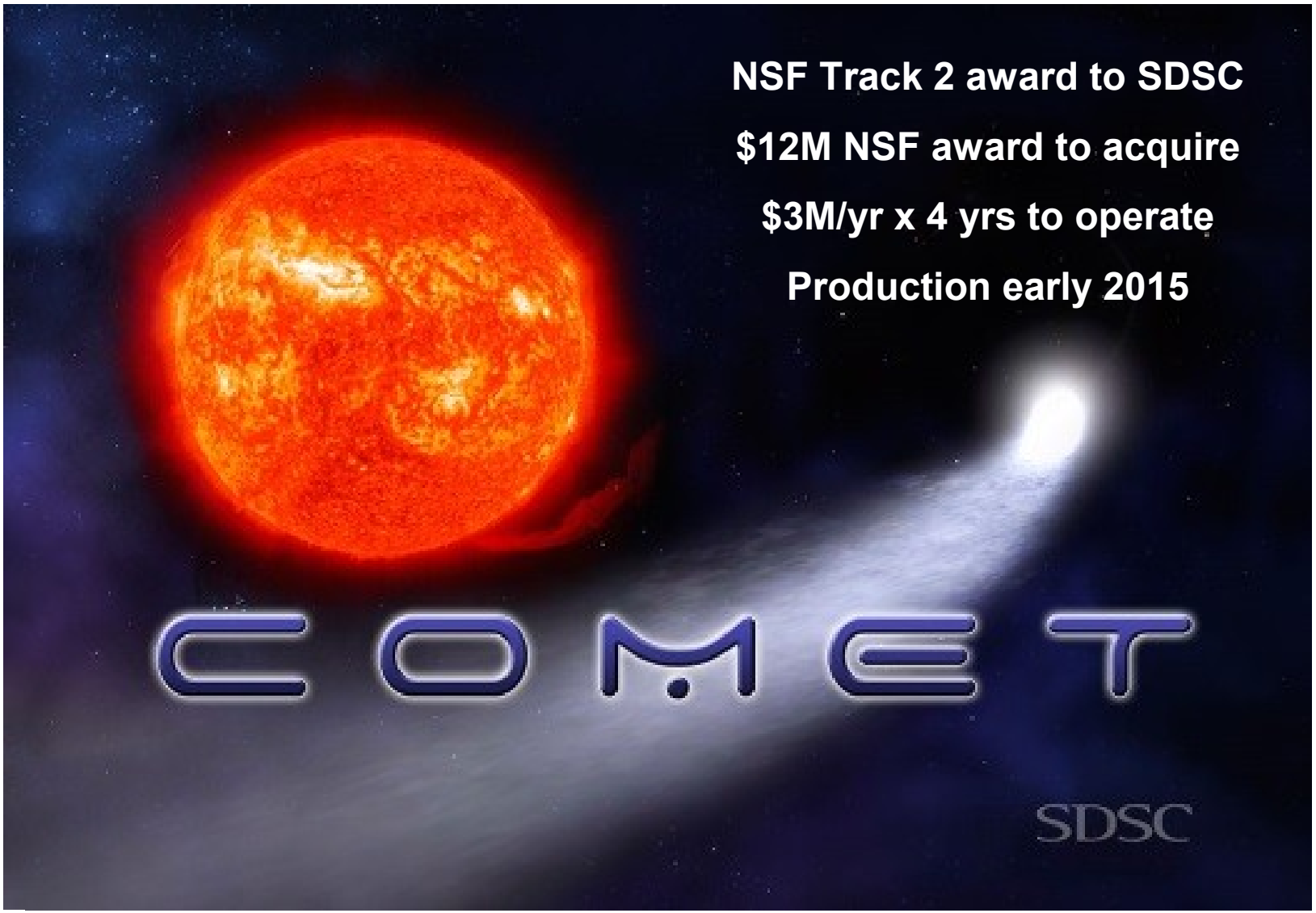
- High Performance Research Computing Facility
  Offered for UC researchers (Including from UC Riverside)
  - Faculty Using Startup Package Funds to Purchase Computing and Storage Time at SDSC
- Hybrid Business Model:
  - **"Condo" – PIs Purchase Nodes;**
    - **RCI Subsidizes Operating Fees**
  - **"Hotel" – Pay-as-you-go Computing Time**
- Launched June 2013 –
  - Seeing Strong Interest, Good/Growing Adoption

UCSD

UC San Diego
Research Cyberinfrastructure

SDSC

Comet is a ~2 PF System Architected for the "Long Tail of Science"



UCSD

SDSC



High Performance Wireless Research and Education Network

http://hpwren.ucsd.edu/

National Science Foundation awards 0087344, 0426879 and 0944131

HPWREN UCSD SDSC SAN DIEGO SUPERCOMPUTER CENTER it² SAN DIEGO STATE UNIVERSITY NATIONAL SCIENCE FOUNDATION















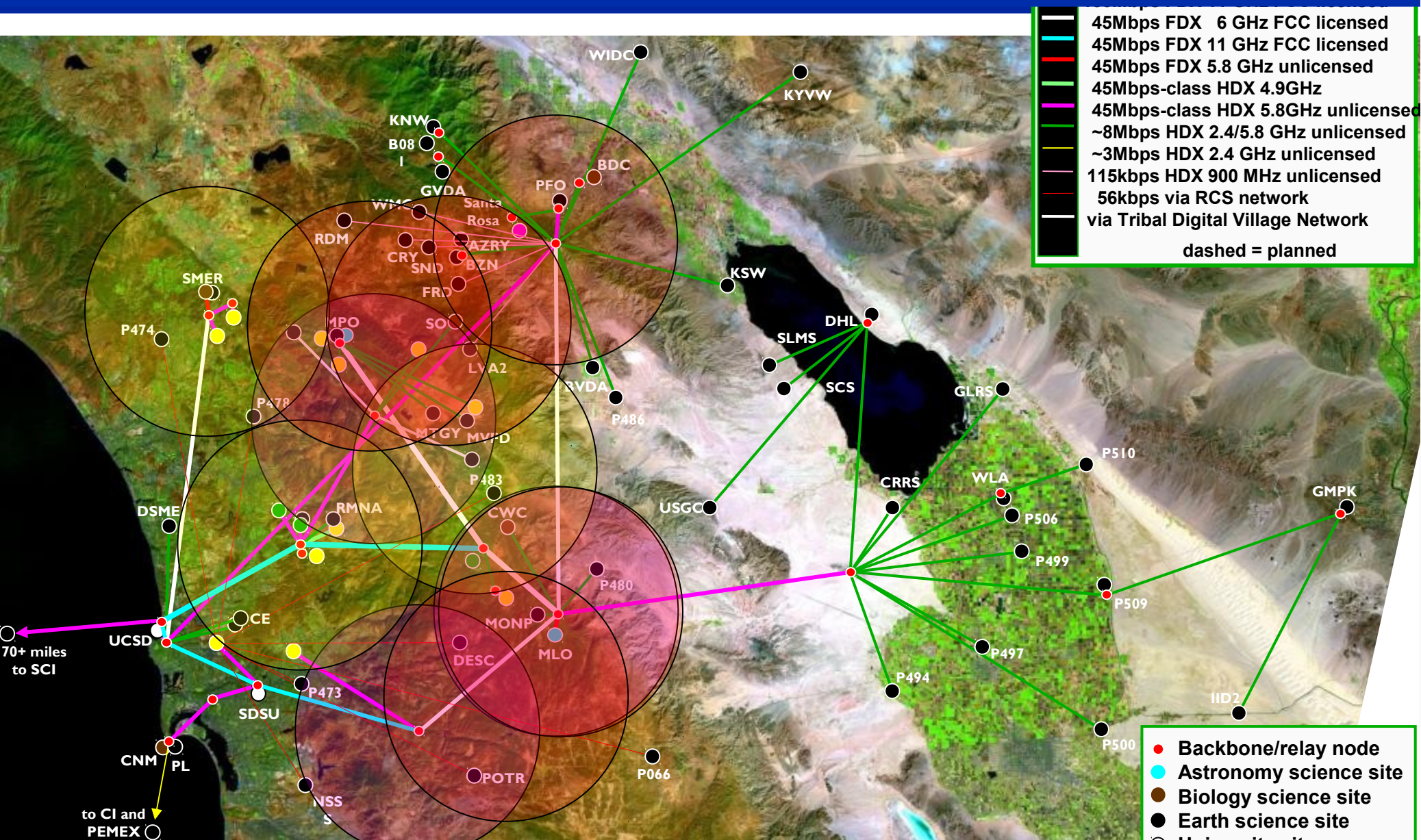
Outreach



HPWREN

Source: Hans Werner Braun, HPWREN PI

HPWREN Topology, 360 Degree Cameras



Red circles: HPWREN supplied cameras

Yellow circles: SD County supplied cameras

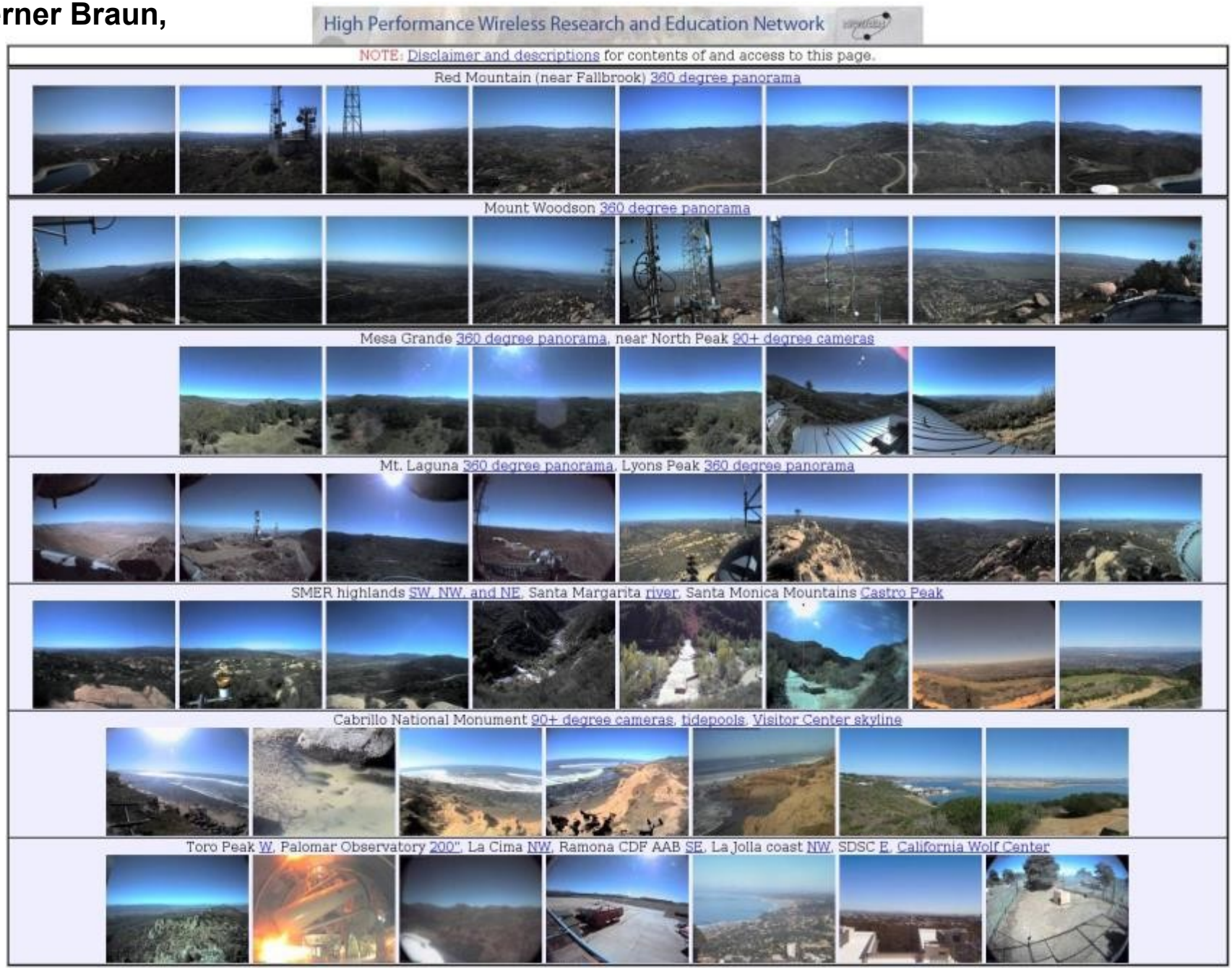
approximately 50 miles:

Source: Hans Werner Braun, HPWREN PI

Note: locations are approximate

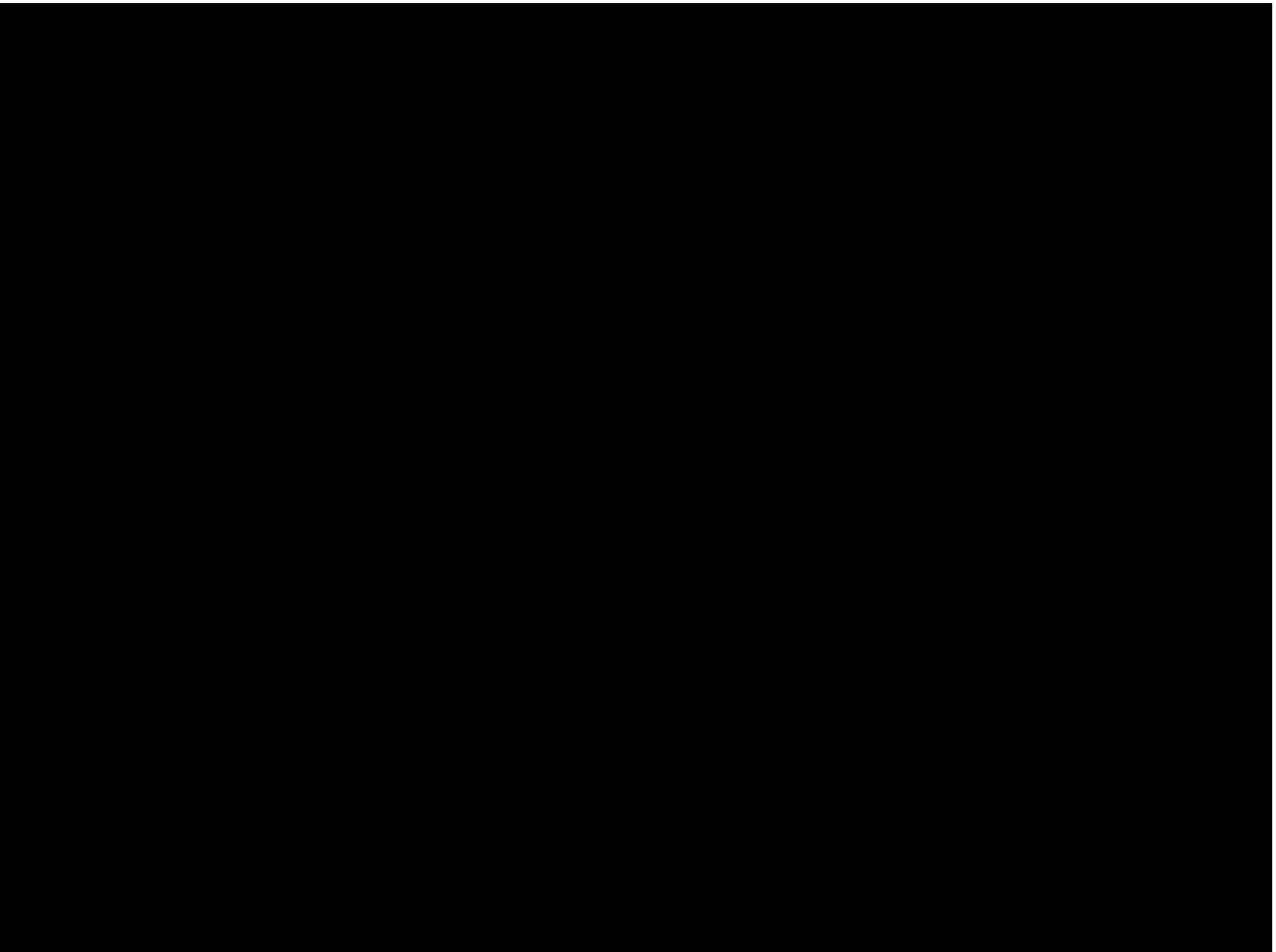
Various Real-Time Network Cameras for Environmental Observations

Source: Hans Werner Braun, HPWREN PI



Time-Lapse Video of Mt. Laguna Chariot Wildfire From HPWREN Camera (July 8, 2013)

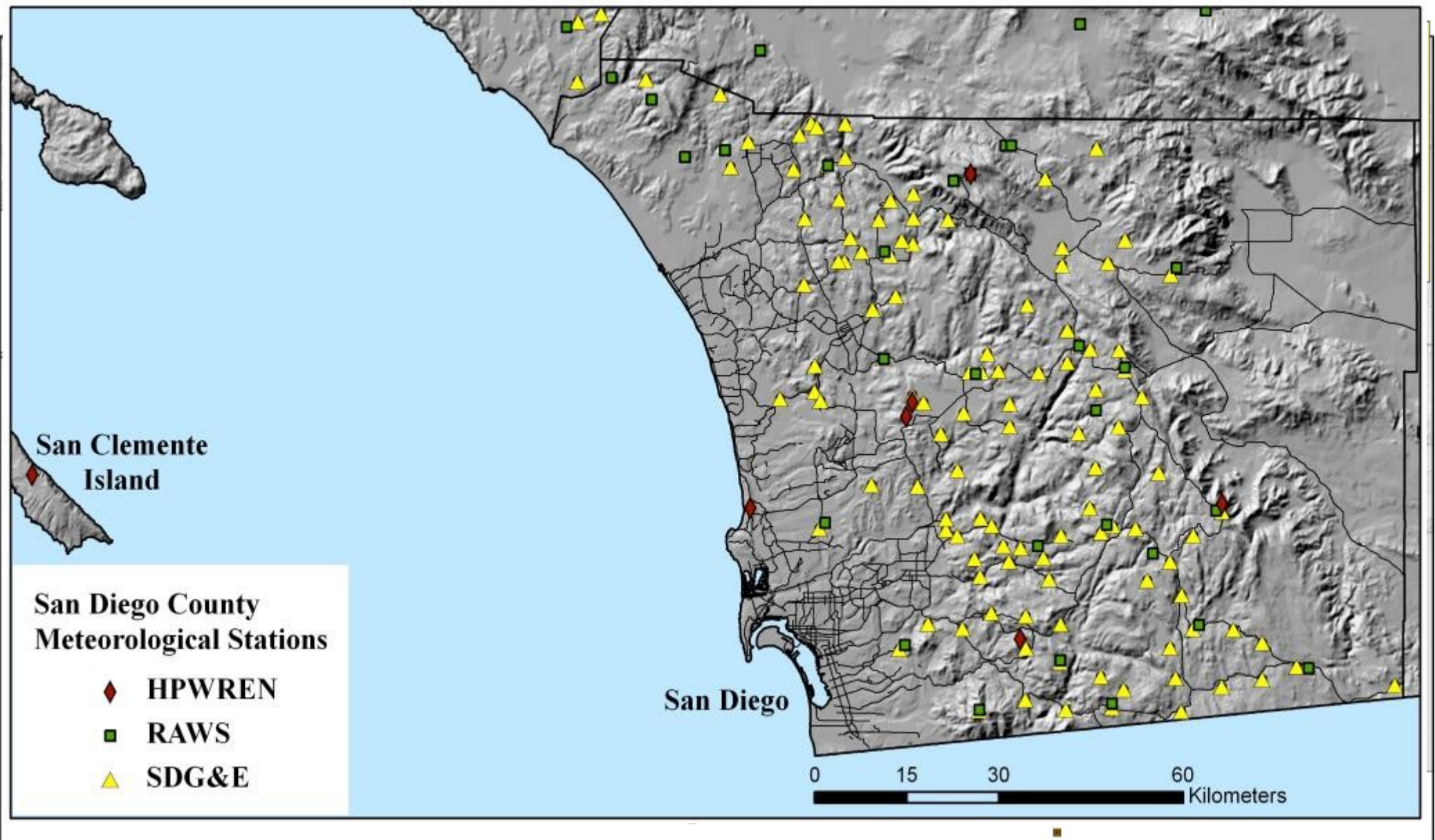
Source: Hans Werner Braun, HPWREN PI



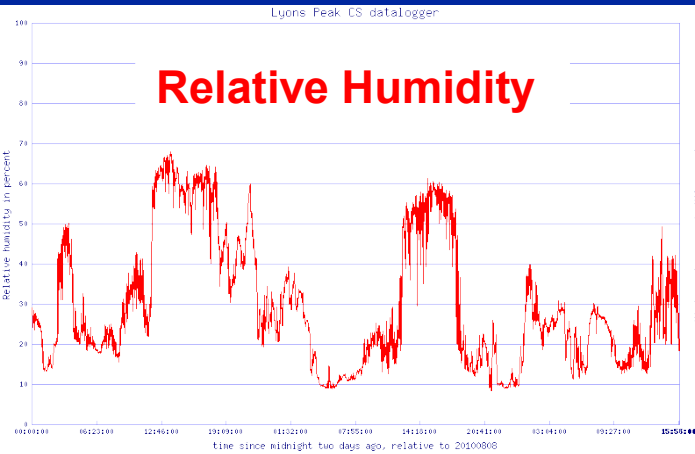
Similar Video of Mountain Fire in Riverside

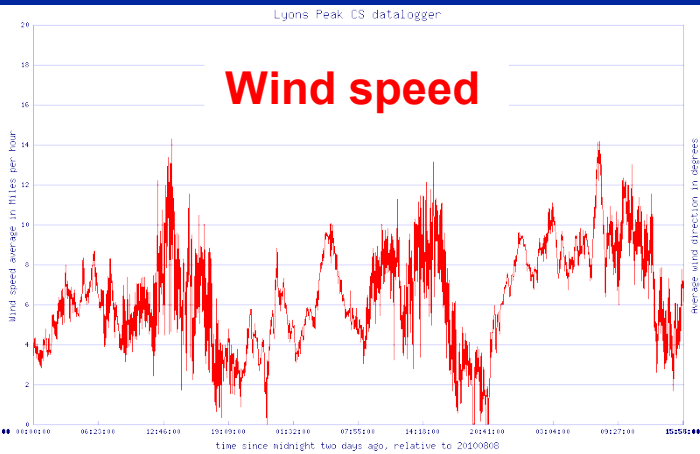
SoCal Weather Stations:
Note the High Density in San Diego County

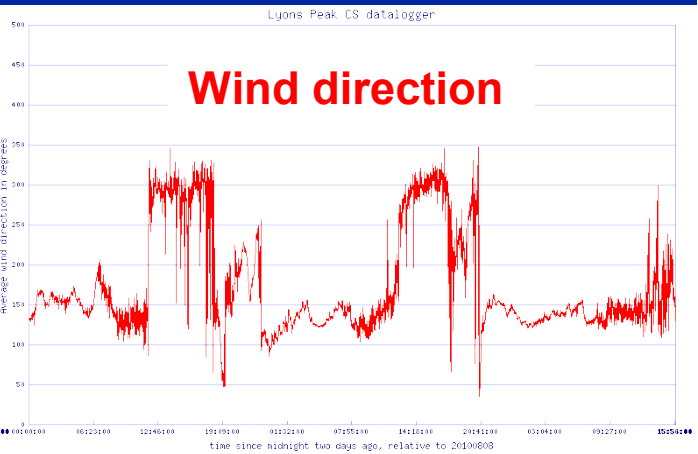
Source: Jessica Block, Calit2



Real-Time Data-Based Alerts to First Responders



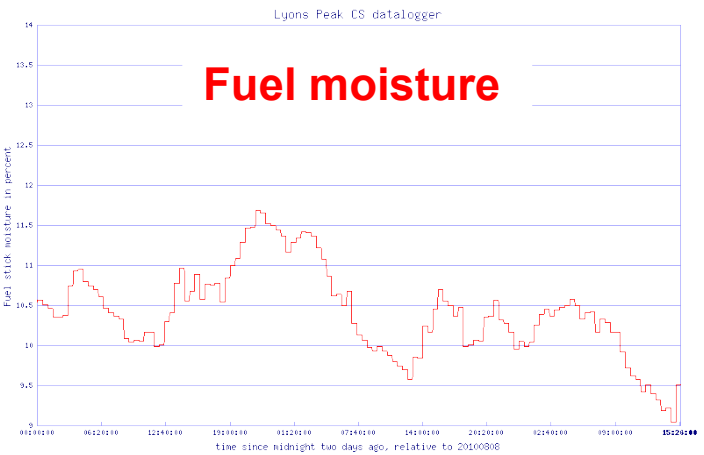




Trigger real-time computer-generated alerts, if:

condition "A" **AND** condition "B" **AND** condition "C"
OR condition "D"

exists, in which case several San Diego emergency officers are being paged or emailed during such alert conditions, based on HPWREN data parameterization by a CDF Division Chief. This system has been in operation since 2004.



Date: Wed, 4 Aug 2010 09:31:05 -0700
Source: Hans Werner Braun, HPWREN PI
Subject: URGENT weather sensor alert

LP: RH=26.1 WD=135.2 WS=1.9 FM=6.8 AT=80.7 at 20100804.093100
More details at http://hpwren.ucsd.edu/Sensors/

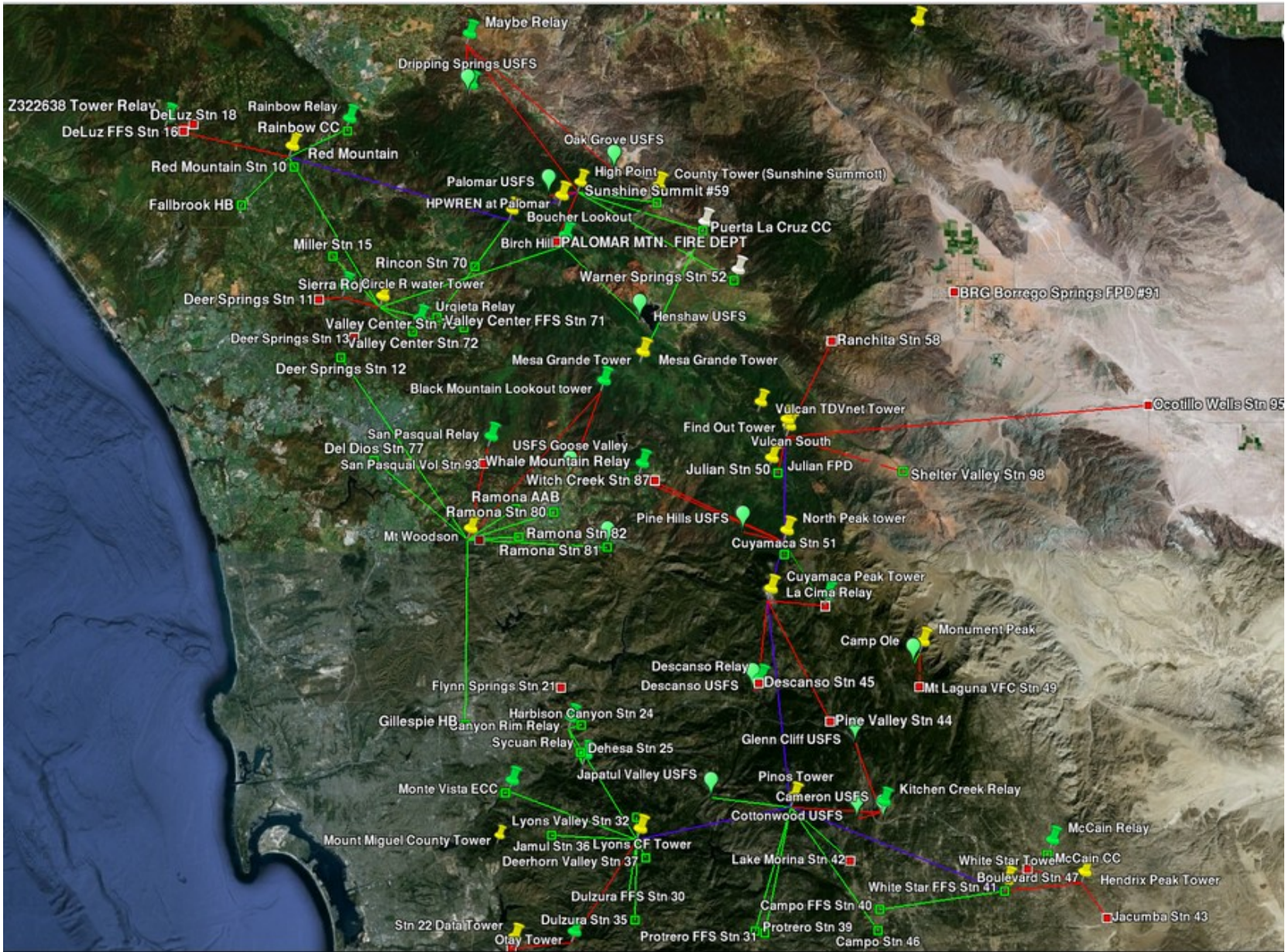
San Diego Wildfire First Responders Meeting at Calit2 Aug 25, 2010



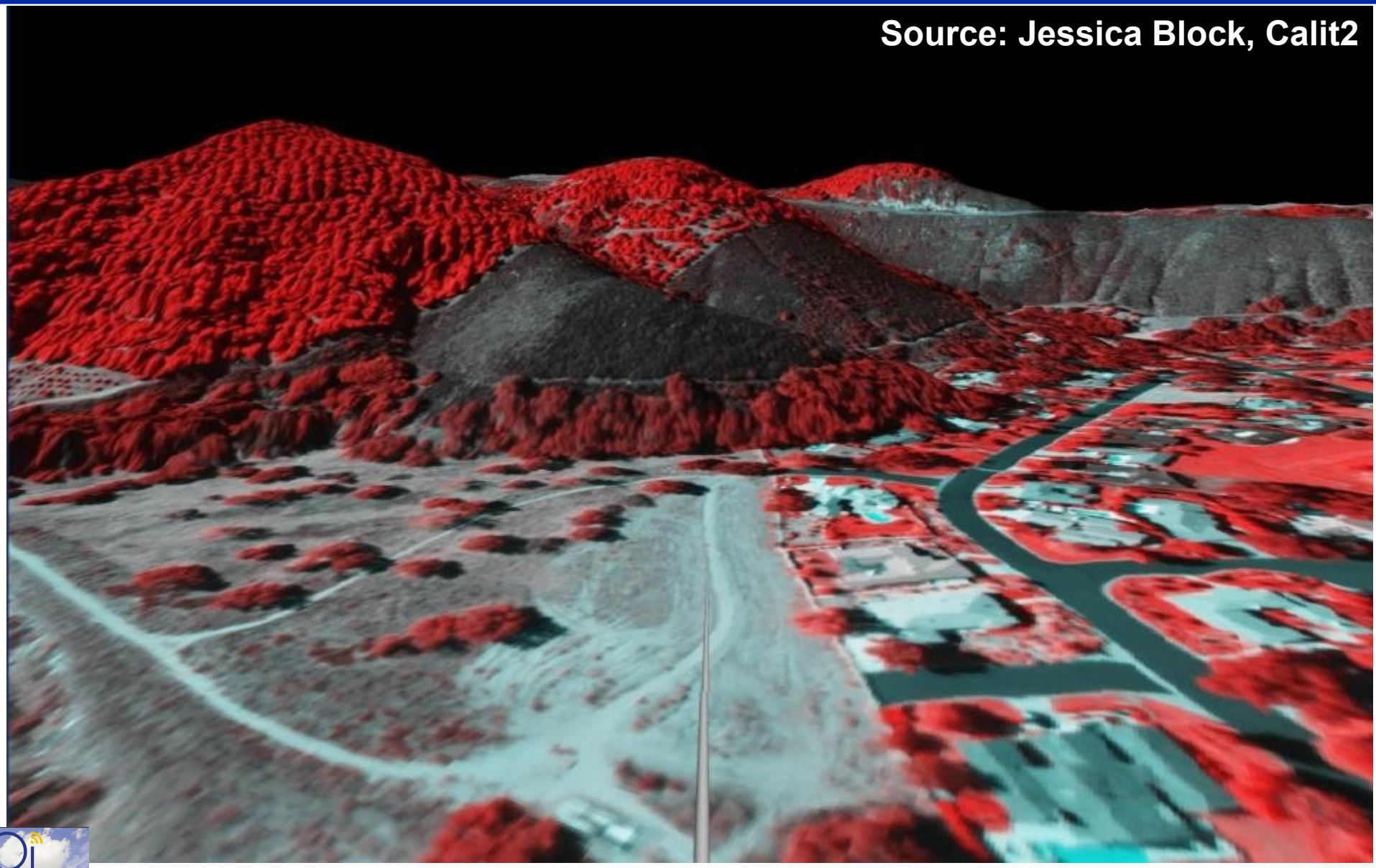
SDSC's Hans-Werner Braun Explains His High Performance Wireless Research and Education Network

Area Situational Awareness for Public Safety Network (ASAPnet) Extends HPWREN to Connect Fire Stations

Connecting 60 backcountry fire stations as the region nears the peak of its fire season.
Aug. 14, 2013 www.calit2.net/newsroom/release.php?id=2210



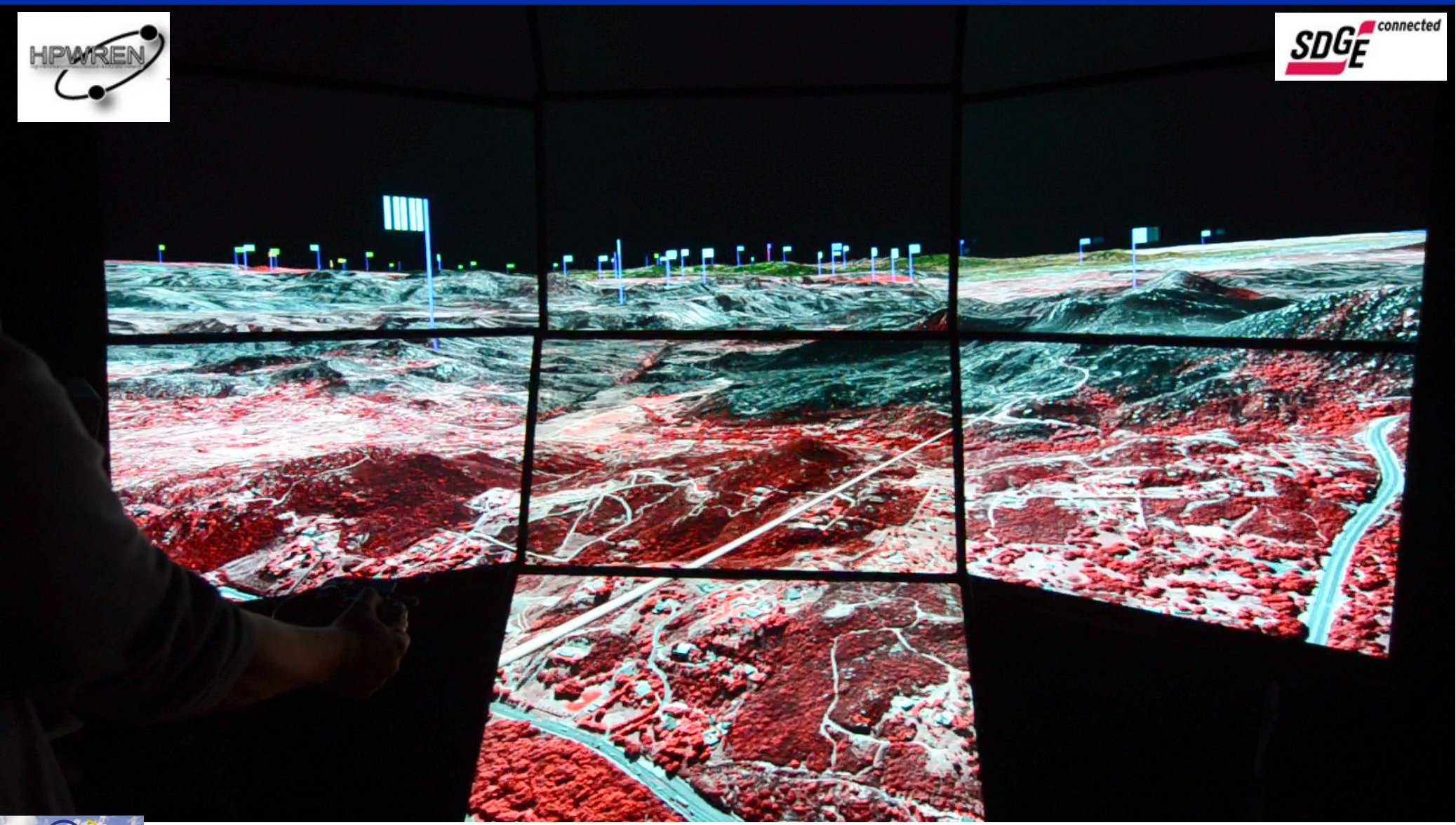
Creating a Digital "Mirror World": Interactive Virtual Reality of San Diego County





0.5 meter image resolution.
2meter resolution elevation

All Meteorological Stations Are Represented in Realtime: Wind Direction, Velocity, and Temperature



Source: Jessica Block, Calit2

Using Calit2's Qualcomm Institute NexCAVE for CAL FIRE Research and Planning





Source: Jessica Block, Calit2

A Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience Cyberinfrastructure for Wildfires (WiFire)

NSF Has Just Awarded the WiFire Grant – Ilkay Altintas SDSC PI

Development of end-to-end "cyberinfrastructure" for "analysis of large dimensional heterogeneous real-time sensor data"



System integration of

- real-time sensor networks,
- satellite imagery,
- near-real time data management tools,
- wildfire simulation tools
- connectivity to emergency command centers before during and after a firestorm

The Department of FIRE PROTECTION ENGINEERING

UCSD Jacobs MECHANICAL and AEROSPACE ENGINEERING

SDSC