“Building an Information Infrastructure to Support Microbial Metagenomic Sciences"

Presentation for the Microbe Project Interagency Team

[www.microbeproject.gov]

UCSD

La Jolla, CA

January 14, 2006

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

OptIPuter

it²

Calit2 Brings Computer Scientists and Engineers Together with Biomedical Researchers

- **Some Areas of Concentration:**
  - Metagenomics
  - Genomic Analysis of Organisms
  - Evolution of Genomes
  - Cancer Genomics
  - Human Genomic Variation and Disease
  - Mitochondrial Evolution
  - Proteomics
  - Computational Biology
  - Information Theory and Biological Systems





1200 Researchers in Two Buildings

OptIPuter

it²

CAMERA: Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

National LambdaRail
Direct Connect
Computation and Storage Complex

Funded by: Gordon and Betty Moore Foundation

Joint Partnership of:

OptIPuter

PI Larry Smarr

Building On Centers of Excellence and Achievement









Powered by Rocks



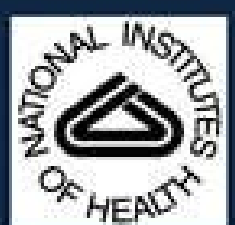






The Scripps Genome Center

Supported by:









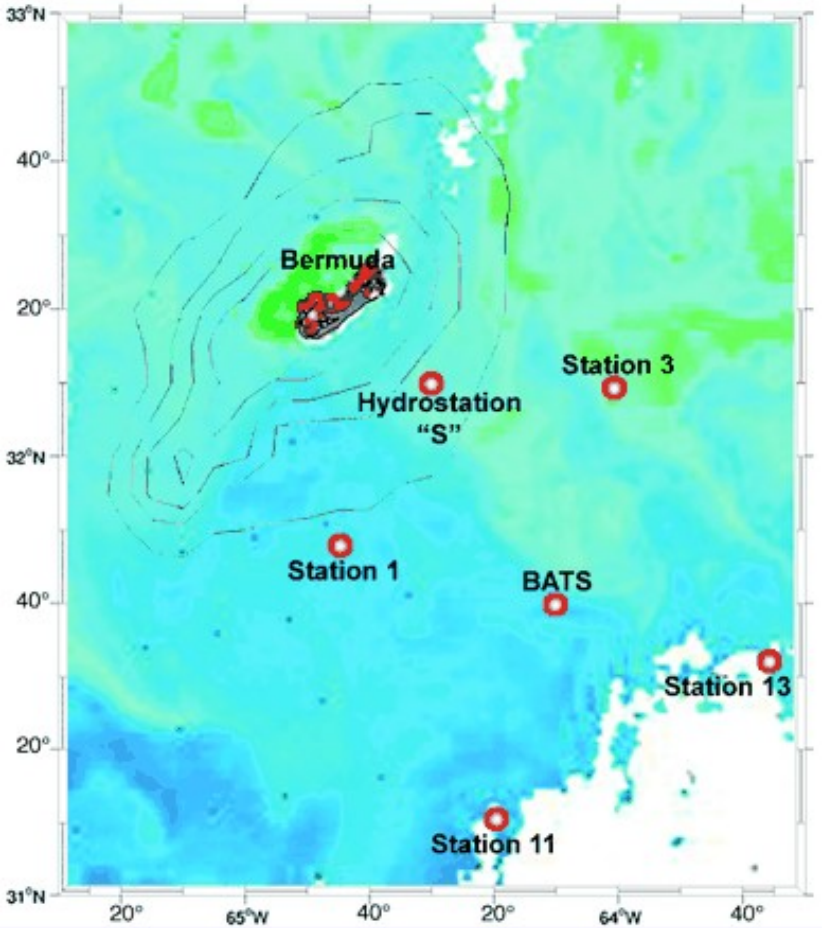


OptIPuter

Announcing Tuesday January 17, 2006

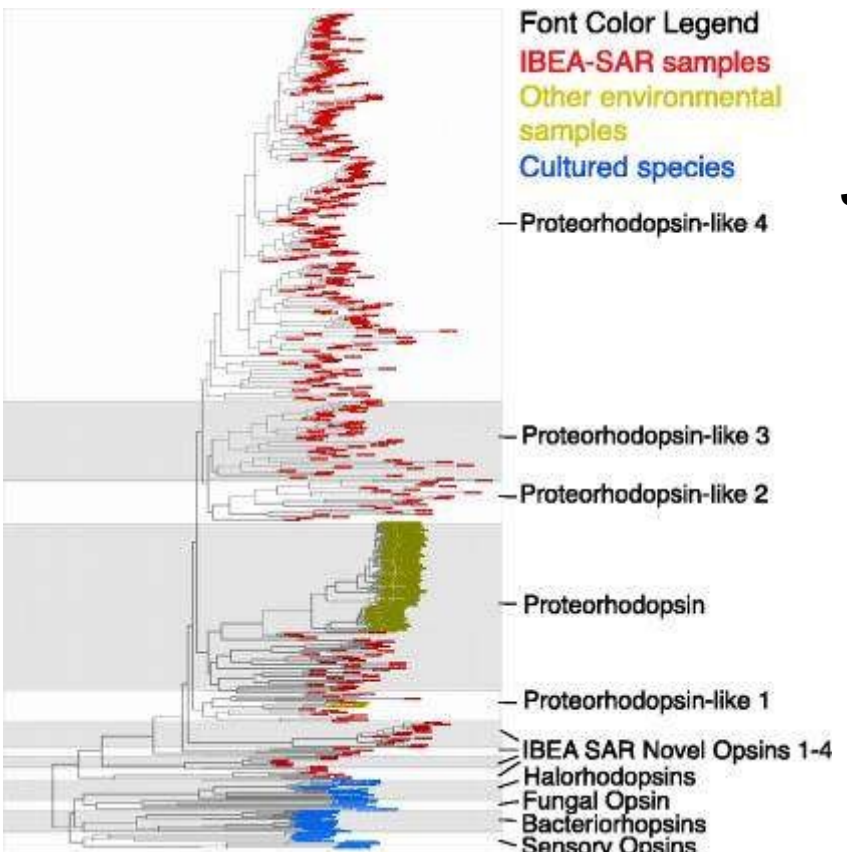
it²

The Sargasso Sea Experiment
The Power of Environmental Metagenomics



MODIS-Aqua satellite image of ocean chlorophyll in the Sargasso Sea grid about the BATS site from 22 February 2003

- Yielded a Total of Over 1 billion Base Pairs of Non-Redundant Sequence
- Displayed the Gene Content, Diversity, & Relative Abundance of the Organisms
- Sequences from at Least 1800 Genomic Species, including 148 Previously Unknown
- Identified over 1.2 Million Unknown Genes



J. Craig Venter, et al.
Science
2 April 2004:
Vol. 304.
pp. 66 - 74

OptIPuter

it²

Marine Genome Sequencing Project
Measuring the Genetic Diversity of Ocean Microbes



CAMERA will include
All Sorcerer II Metagenomic Data

MARINE MICROBIOLOGY INITIATIVE

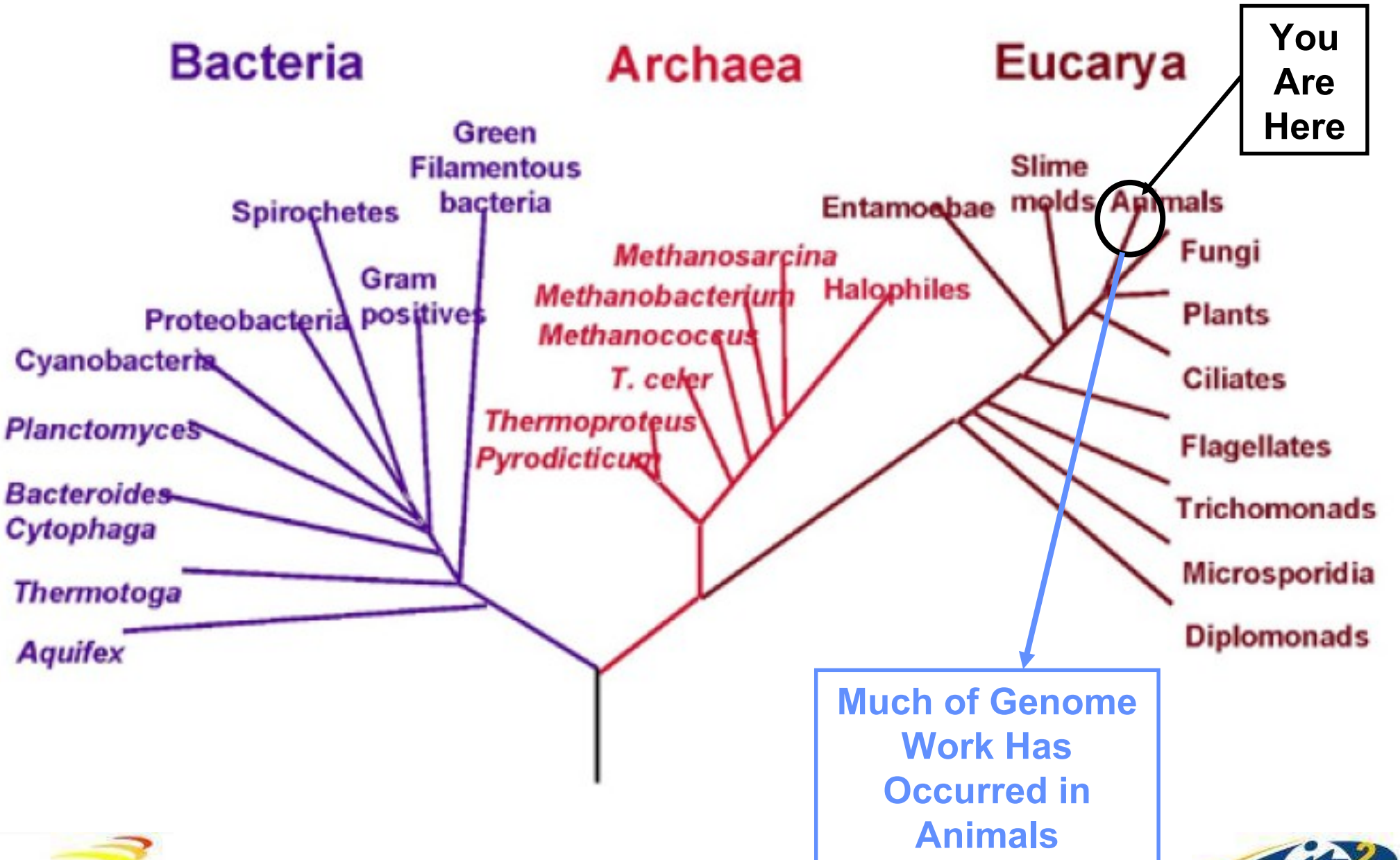
Gordon and Betty MOORE FOUNDATION

OptIPuter

J. Craig Venter INSTITUTE

it²

Evolution is the Principle of Biological Systems: Most of Evolutionary Time Was in the Microbial World



Source: Carl Woese, et al

Major New Science Challenge: Understanding the Transition from Collective to Species Evolution

"Bacteria naturally reside in communities, in ecosystems. It is hard to find a bacterial niche that does not comprise hundreds or thousands of different species, all interacting in intricate delicate ways, to make a fascinatingly complex and stable whole."

"In an era of rampant horizontal gene transfer, organismal evolution would be basically collective. It is the community of organisms that evolves, not the various individual organismal types."

"This shift from a primitive genetic free-for-all to modern organisms must by all account have been one of the most profound happenings in the whole of evolutionary history."

--Carl Woese , *Evolving Biological Organization* in Microbial Phylogeny and Evolution, ed. Jan Sapp (2005)

OptIPuter

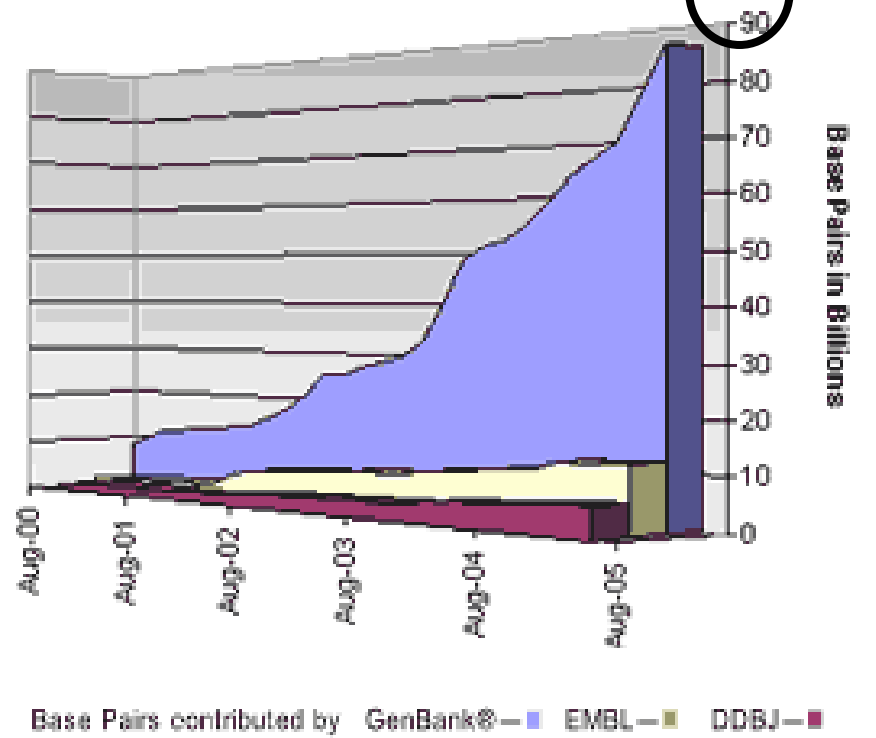
it²

Genomic Data Is Growing Rapidly, But Metagenomics Will Vastly Increase The Scale…

100 Billion Bases!

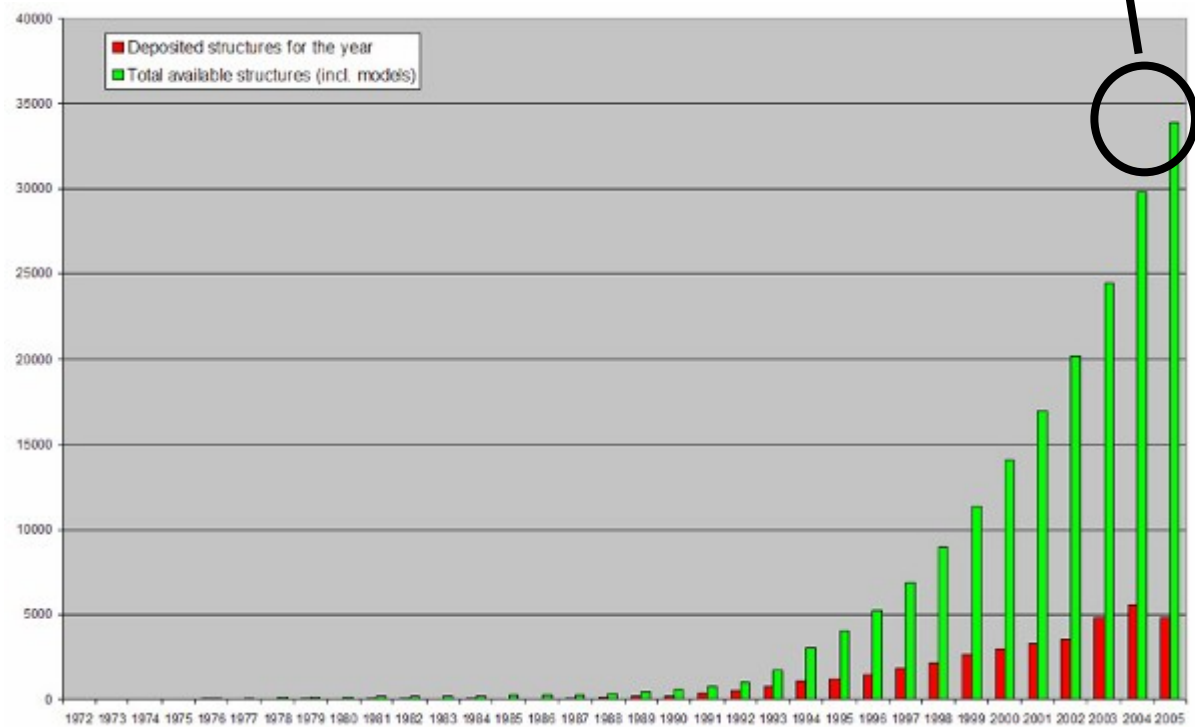
35,000 Structures

GenBank



www.ncbi.nlm.nih.gov/Genbank

Protein Data Bank



www.rcsb.org/pdb/holdings.html

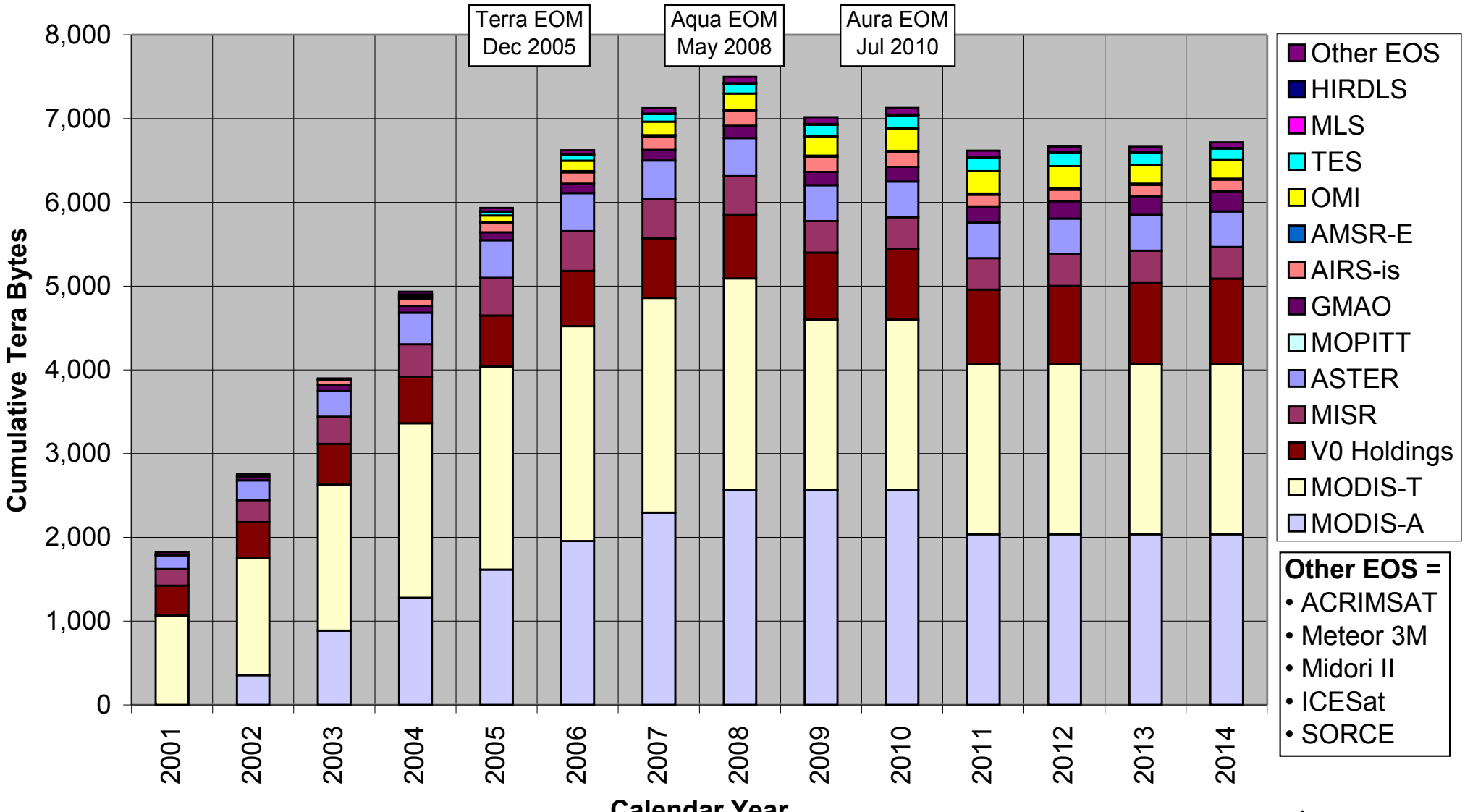
OptIPuter

Total Data < 1TB

it²

Metagenomics Will Couple to Earth Observations Which Add Several TBs/Day

Cumulative Archive Holdings by Instruments/Missions

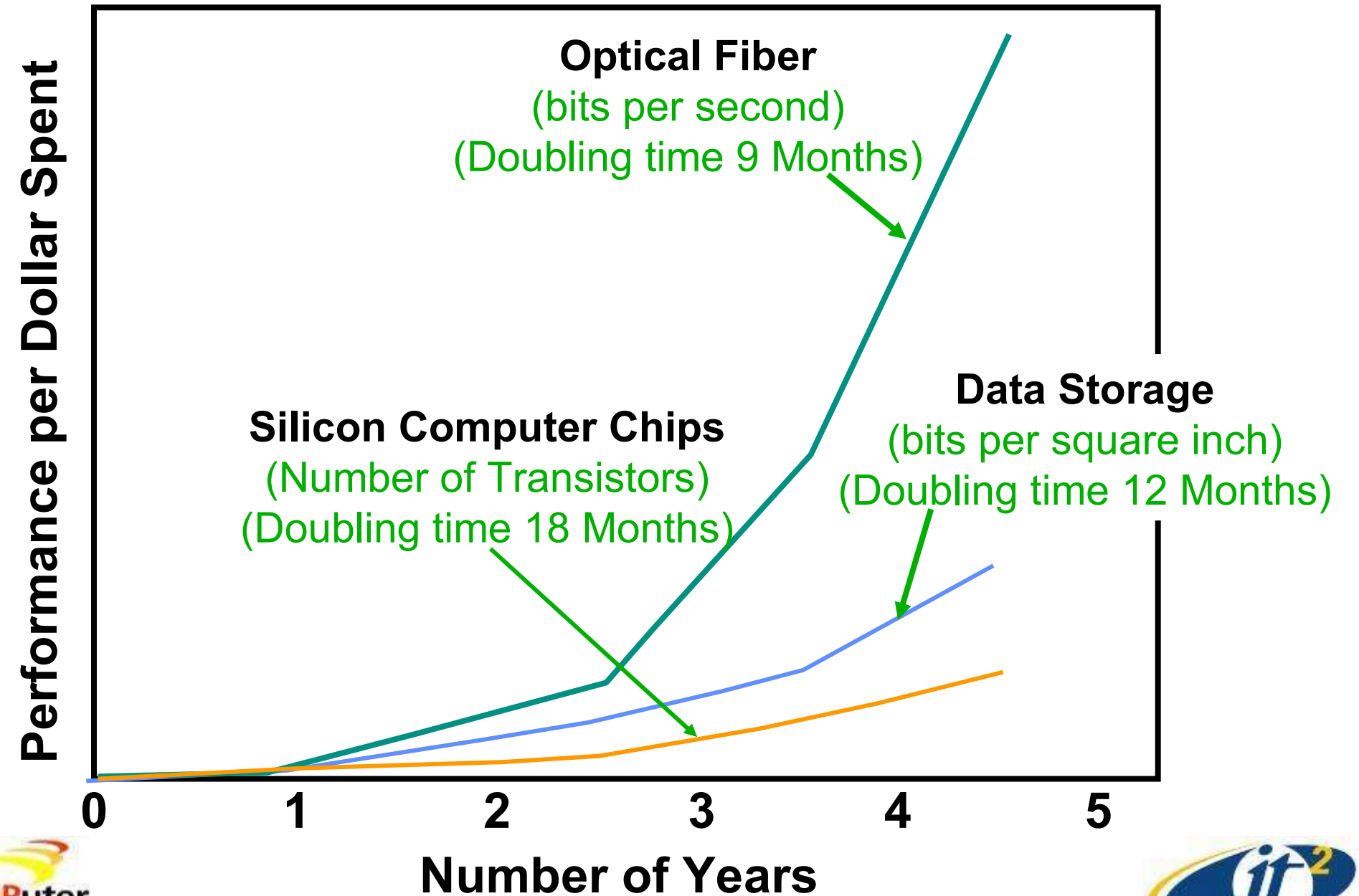


file name: archive holdings_122204.xls
tab: all instr bar

NOTE: Data remains in the archive pending transition to LTA

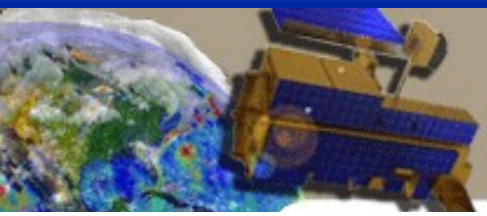
Source: Glenn Iona, EOSDIS Element Evolution Technical Working Group January 6-7, 2005

Optical Networks Are Becoming the 21st Century Cyberinfrastructure Driver



Scientific American, January 2001

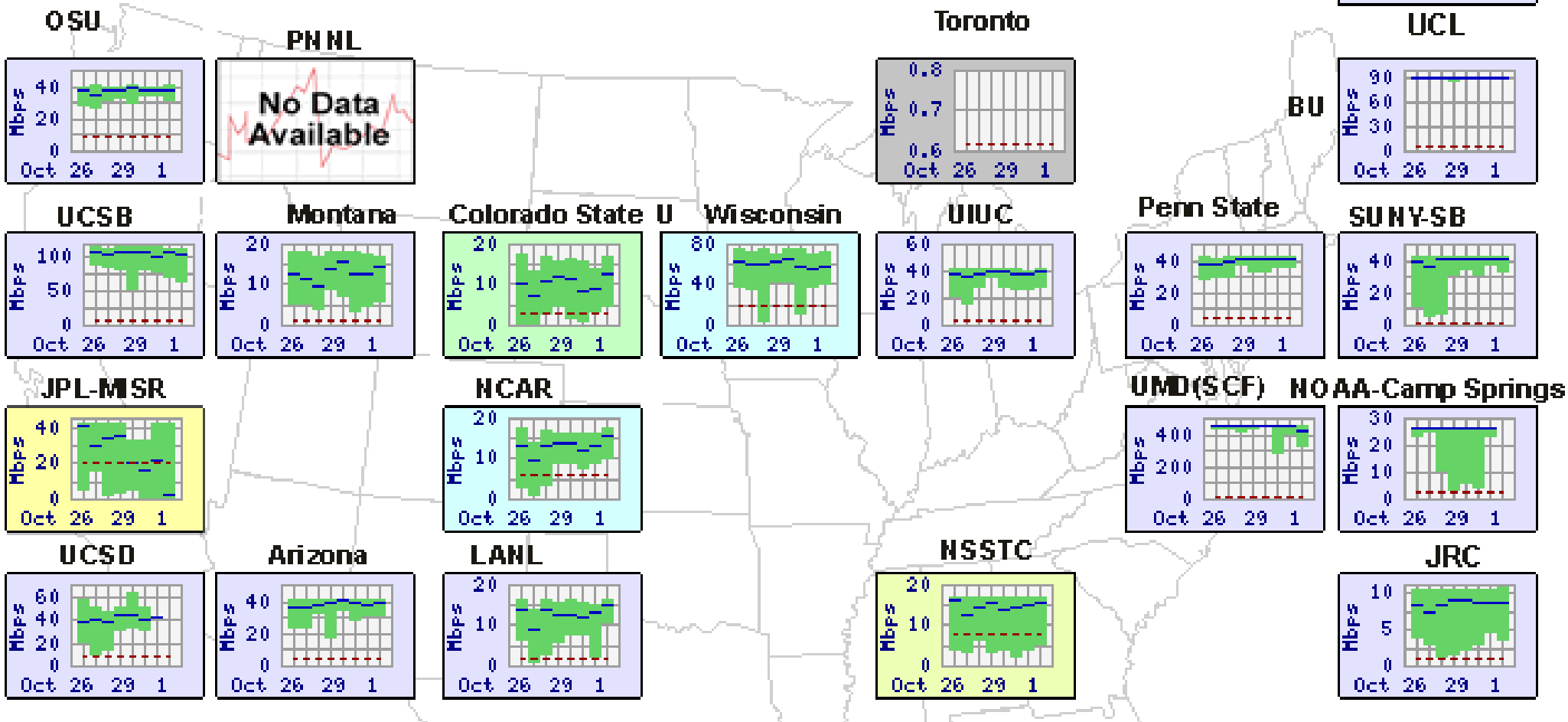
Challenge: Average Throughput of NASA Data Products to End User is < 50 Mbps



ENSIGHT Active Testing TERRA Destinations

Tested October 2005

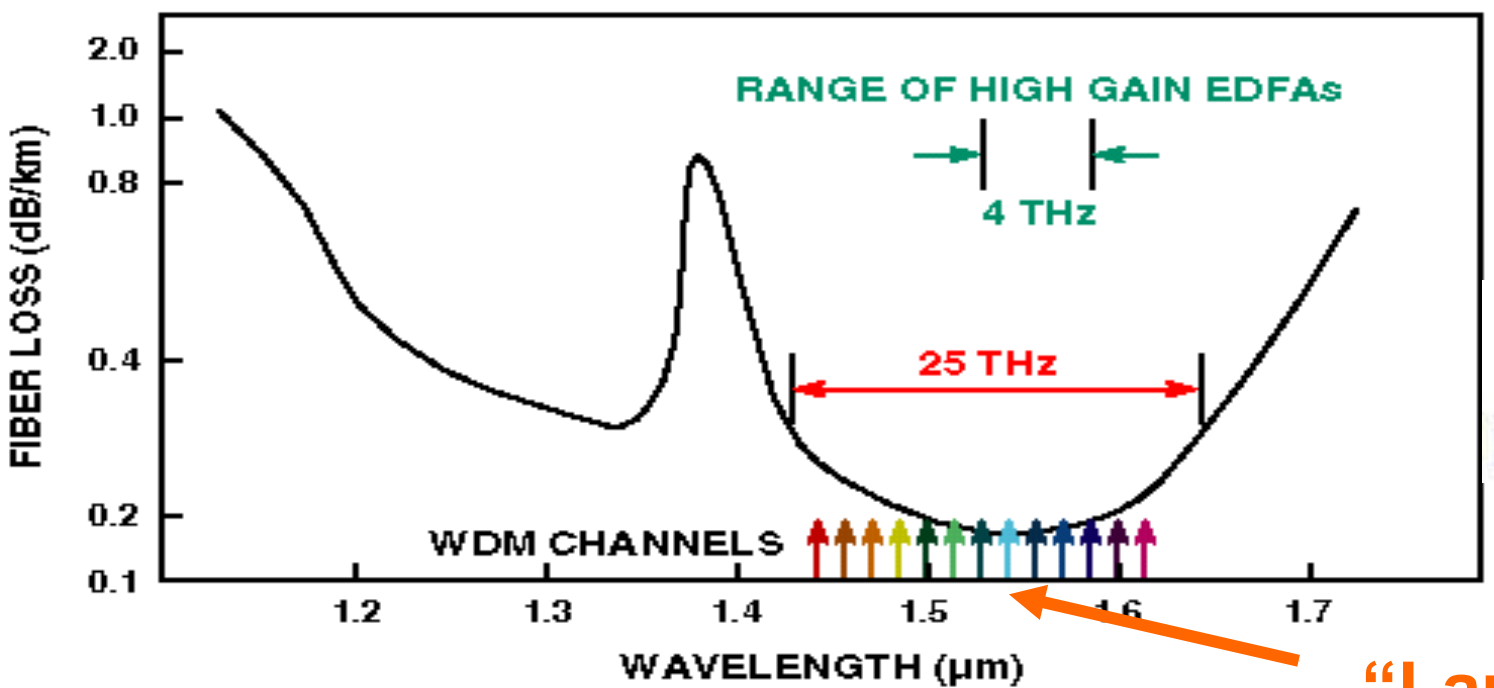
Internet2 Backbone is 10,000 Mbps!
Throughput is < 0.5% to End User



Solution: Individual 1 or 10Gbps Lightpaths -- "Lambdas on Demand"

WAVELENGTH DIVISION MULTIPLEXING (WDM)

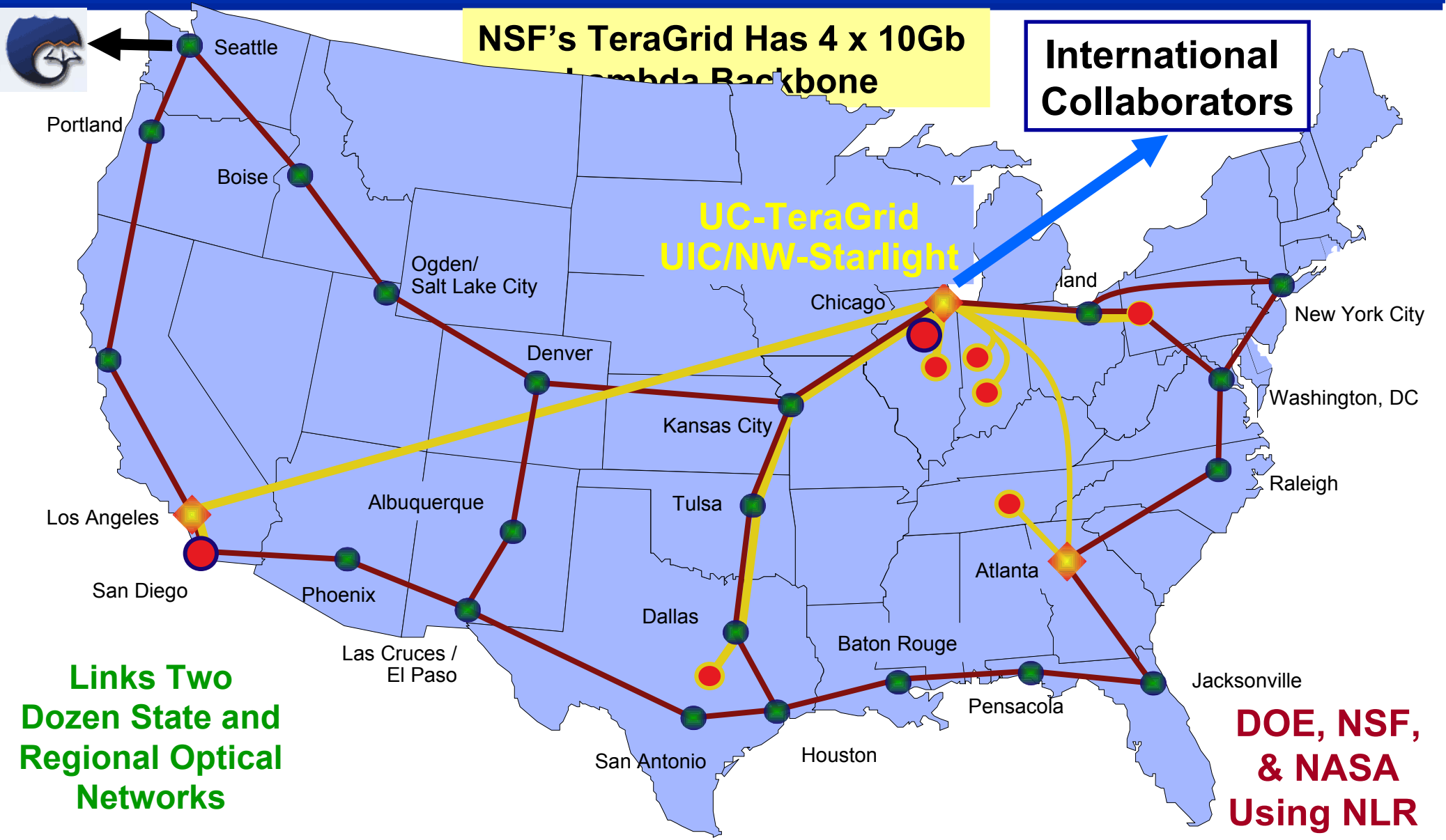
- EXPLOITS
  - ENORMOUS BANDWIDTH OF SILICA FIBER
  - HIGH-GAIN WIDEBAND OPTICAL AMPLIFIERS



"Lambdas"

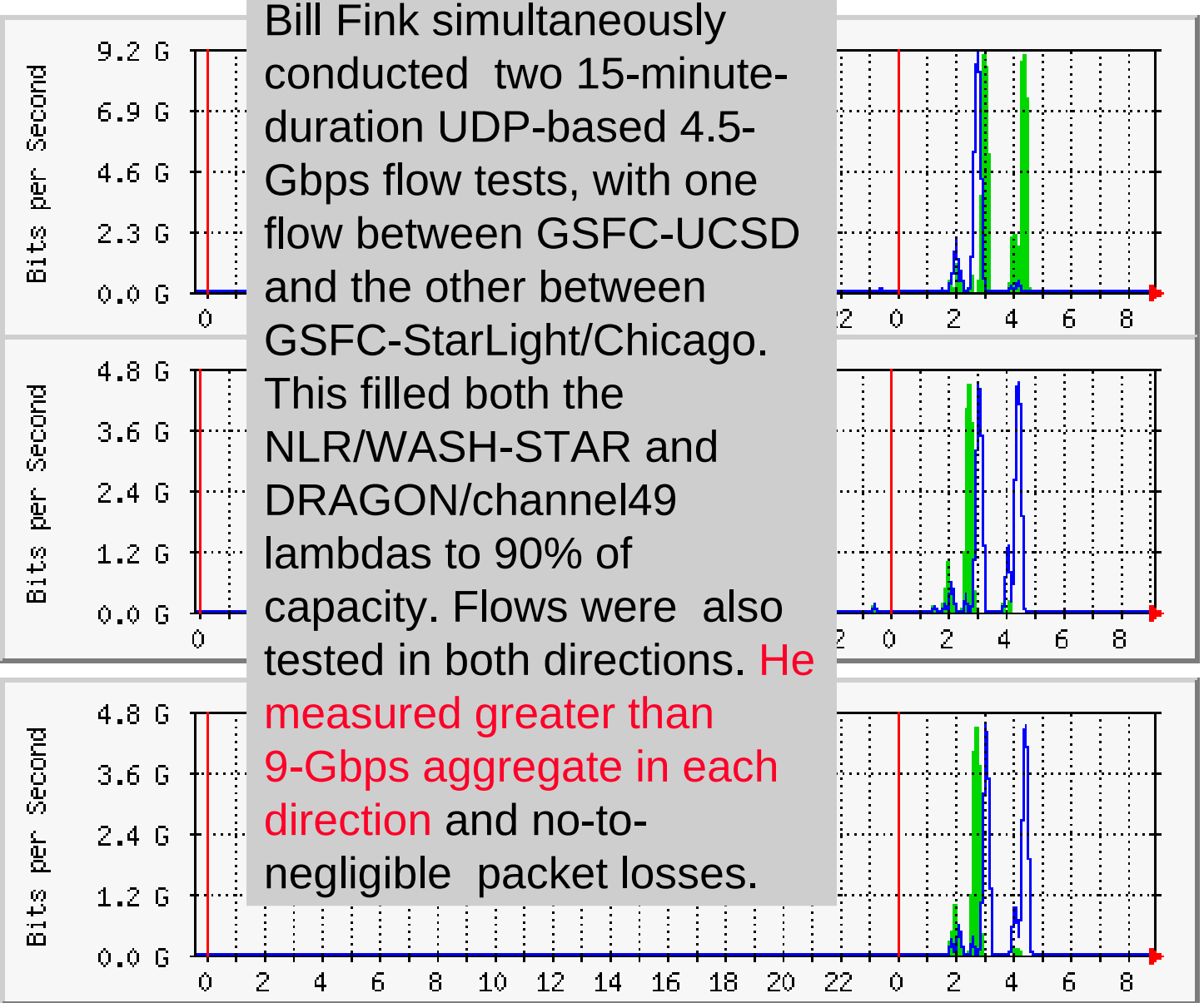
Source: Steve Wallach, Chiaro Networks

National Lambda Rail (NLR) and TeraGrid Provides Cyberinfrastructure Backbone for U.S. Researchers



Lambdas Give End Users Sustained ~ 10 Gbps Data Flow Rates

On August 5, 2005, GSFC's Bill Fink simultaneously conducted two 15-minute-duration UDP-based 4.5-Gbps flow tests, with one flow between GSFC-UCSD and the other between GSFC-StarLight/Chicago. This filled both the NLR/WASH-STAR and DRAGON/channel49 lambdas to 90% of capacity. Flows were also tested in both directions. He measured greater than 9-Gbps aggregate in each direction and no-to-negligible packet losses.



DRAGON 10Gig DWDM XFP
5 August 2005

200 Times Faster Than Standard Internet2!

chance1 10Gig (eth1 Intel Pro/10GbE)
5 August 2005

chance2 10Gig (eth1 Intel Pro/10GbE)
5 August 2005

Source: Pat Gary, NASA GSFC

Global Connections
Between University Research Centers at 10Gbps

Maxine Brown, Tom DeFanti, Co-Chairs

iGrid 2005

THE GLOBAL LAMBDA INTEGRATED FACILITY

www.igrid2005.org



the UC Discovery Grant

September 26-30, 2005
Calit2 @ University of California, San Diego
California Institute for Telecommunications and Information Technology



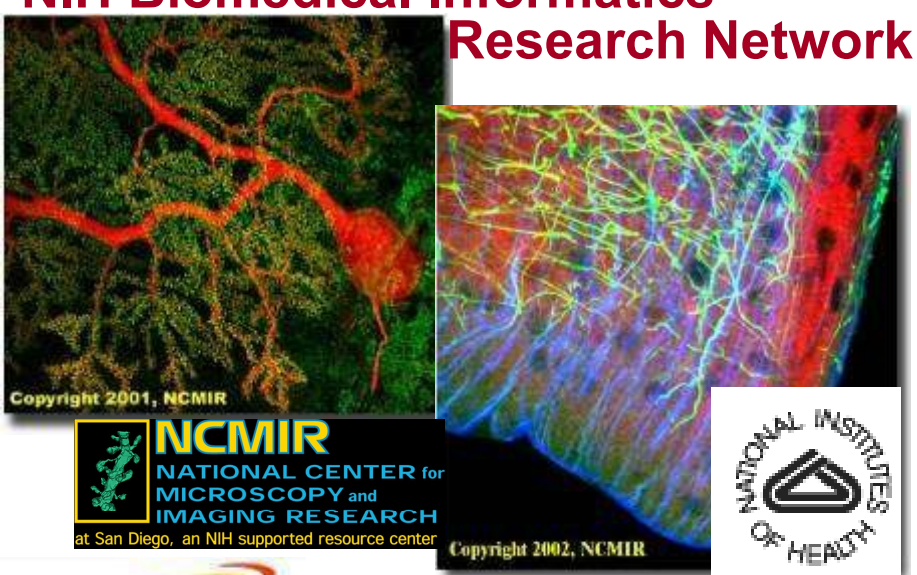
21 Countries Driving 50 Demonstrations
1 or 10Gbps to Calit2@UCSD Building
Sept 2005

The OptIPuter Project –

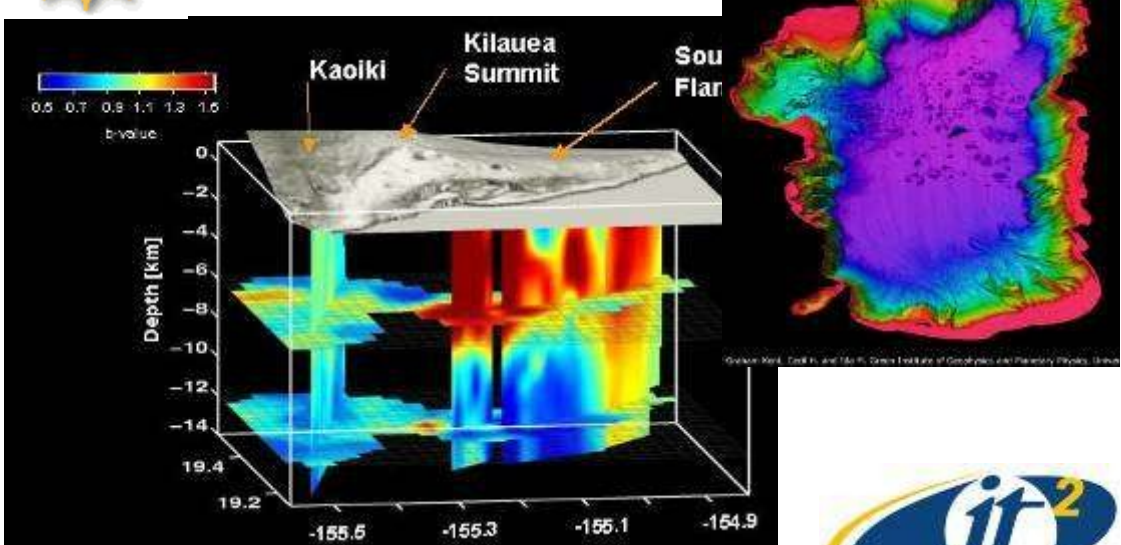
Creating a LambdaGrid "Web" for Gigabyte Data Objects

- **NSF Large Information Technology Research Proposal**
  - Calit2 (UCSD, UCI) and UIC Lead Campuses—Larry Smarr PI
  - Partnering Campuses: USC, SDSU, NW, TA&M, UvA, SARA, NASA
- **Industrial Partners**
  - IBM, Sun, Telcordia, Chiaro, Calient, Glimmerglass, Lucent
- **$13.5 Million Over Five Years**
- **Linking Global Scale Science Projects to User's Linux Clusters**

NIH Biomedical Informatics Research Network



NSF EarthScope and ORION



OptIPuter

SDSC SAN DIEGO SUPERCOMPUTER CENTER

it2

What is the OptIPuter?

- Applications Drivers → Interactive Analysis of Large Data Sets
- OptIPuter Nodes → Scalable PC Clusters with Graphics Cards
- IP over Lambda Connectivity →Predictable Backplane
- Open Source LambdaGrid Middleware→ Network is Reservable
- Data Retrieval and Mining → Lambda Attached Data Servers
- High Defn. Vis., Collab. SW → High Performance Collaboratory

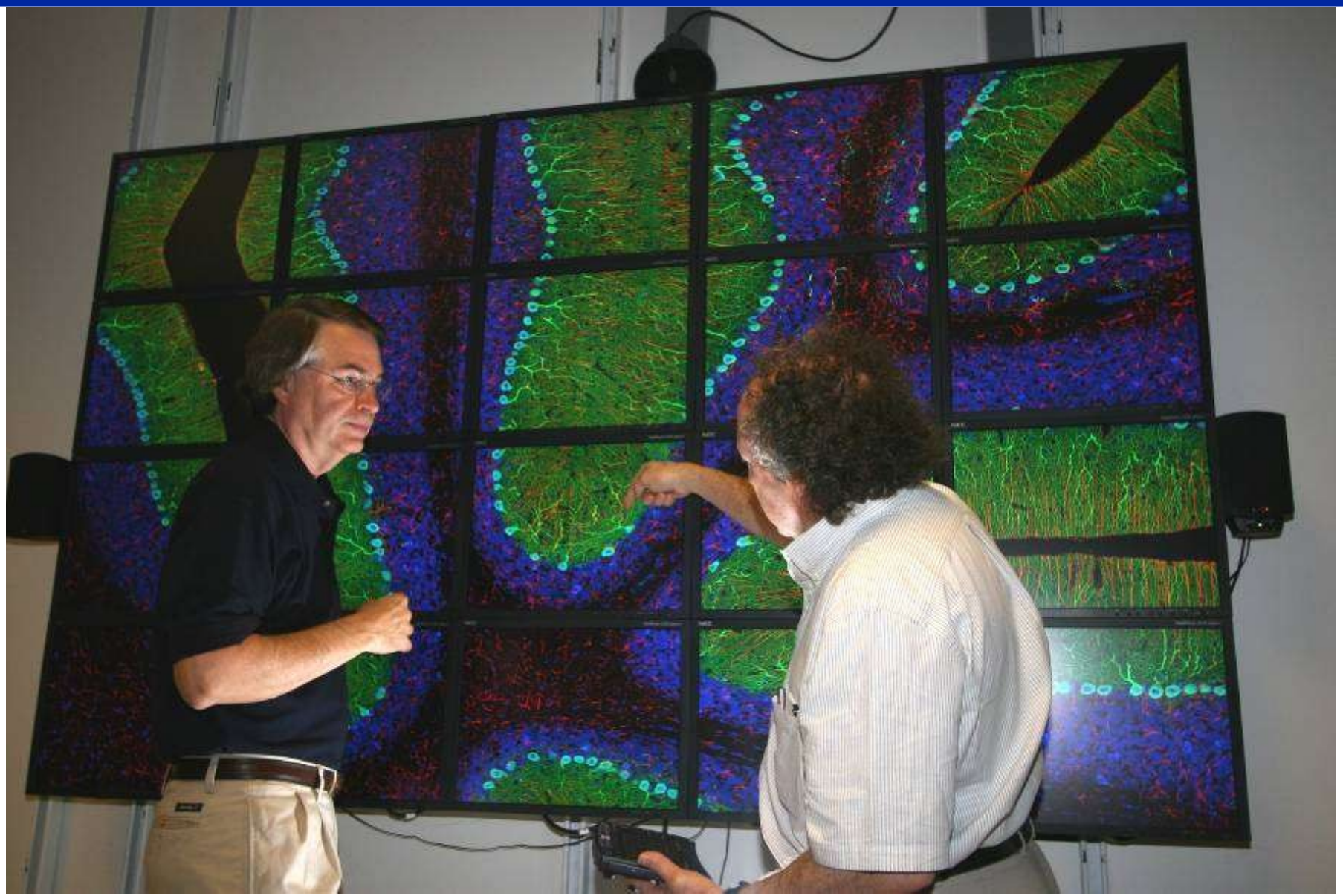
www.optiputer.net

OptIPuter

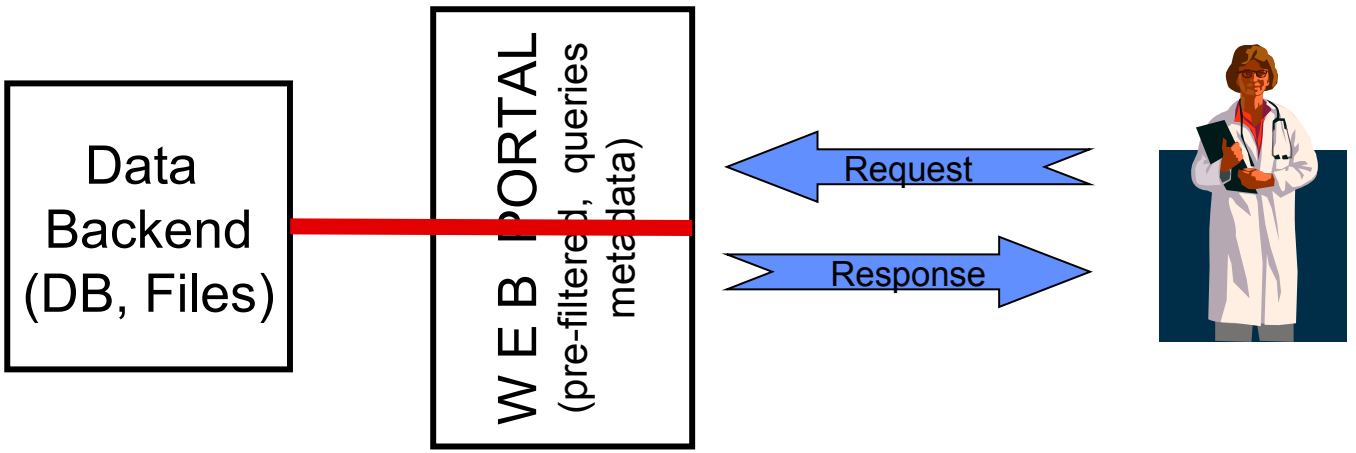
See Nov 2003 Communications of the ACM for Articles on OptIPuter Technologies

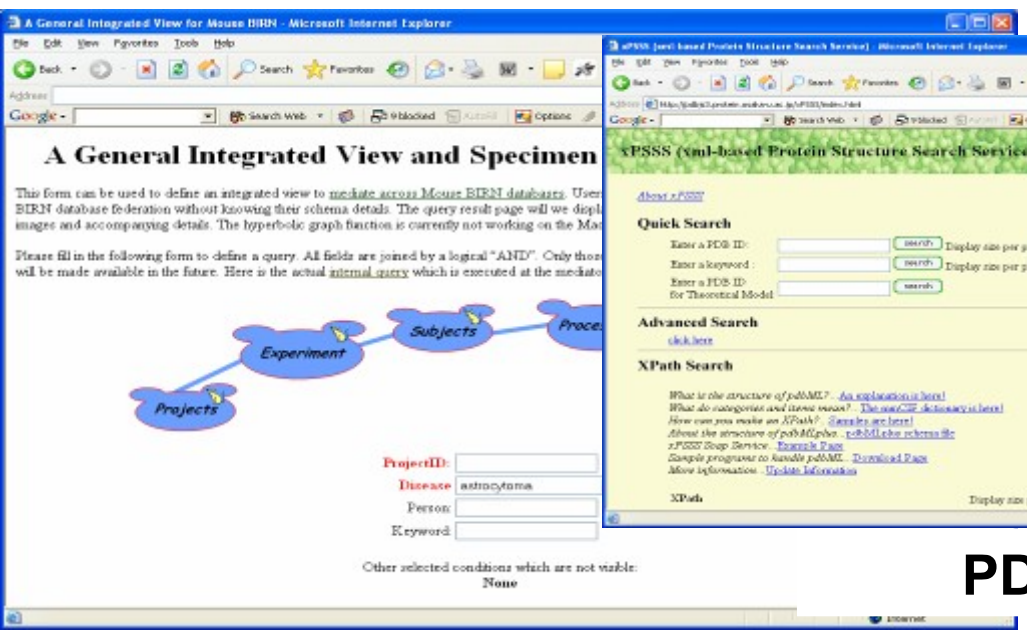
it²

End User Device: Tiled Wall Driven by OptIPuter Graphics Cluster



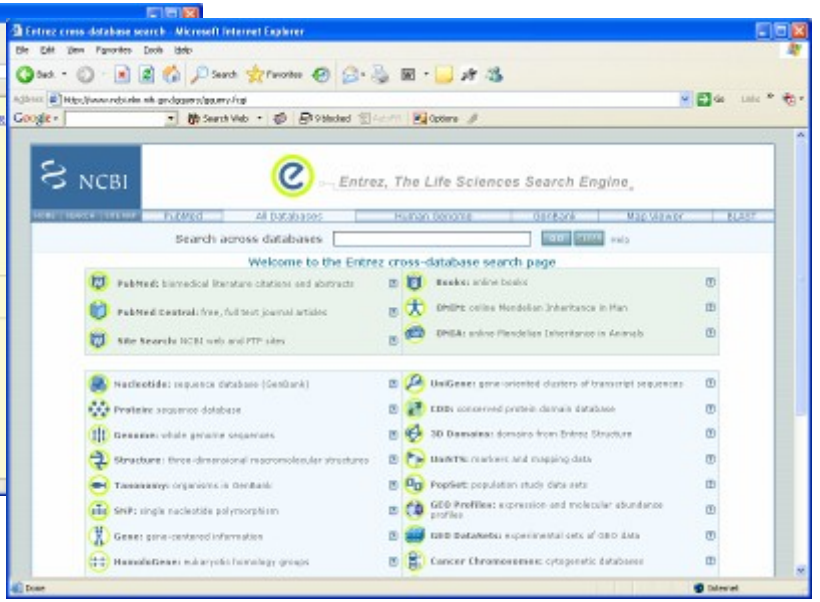
Calit2 Intends to Jump Beyond Traditional Web-Accessible Databases





BIRN

PDB



NCBI Genbank

+ many others

Source: Phil Papadopoulos, SDSC, Calit2

Calit2's Direct Access Core Architecture Will Create Next Generation Metagenomics Server

Sargasso Sea Data

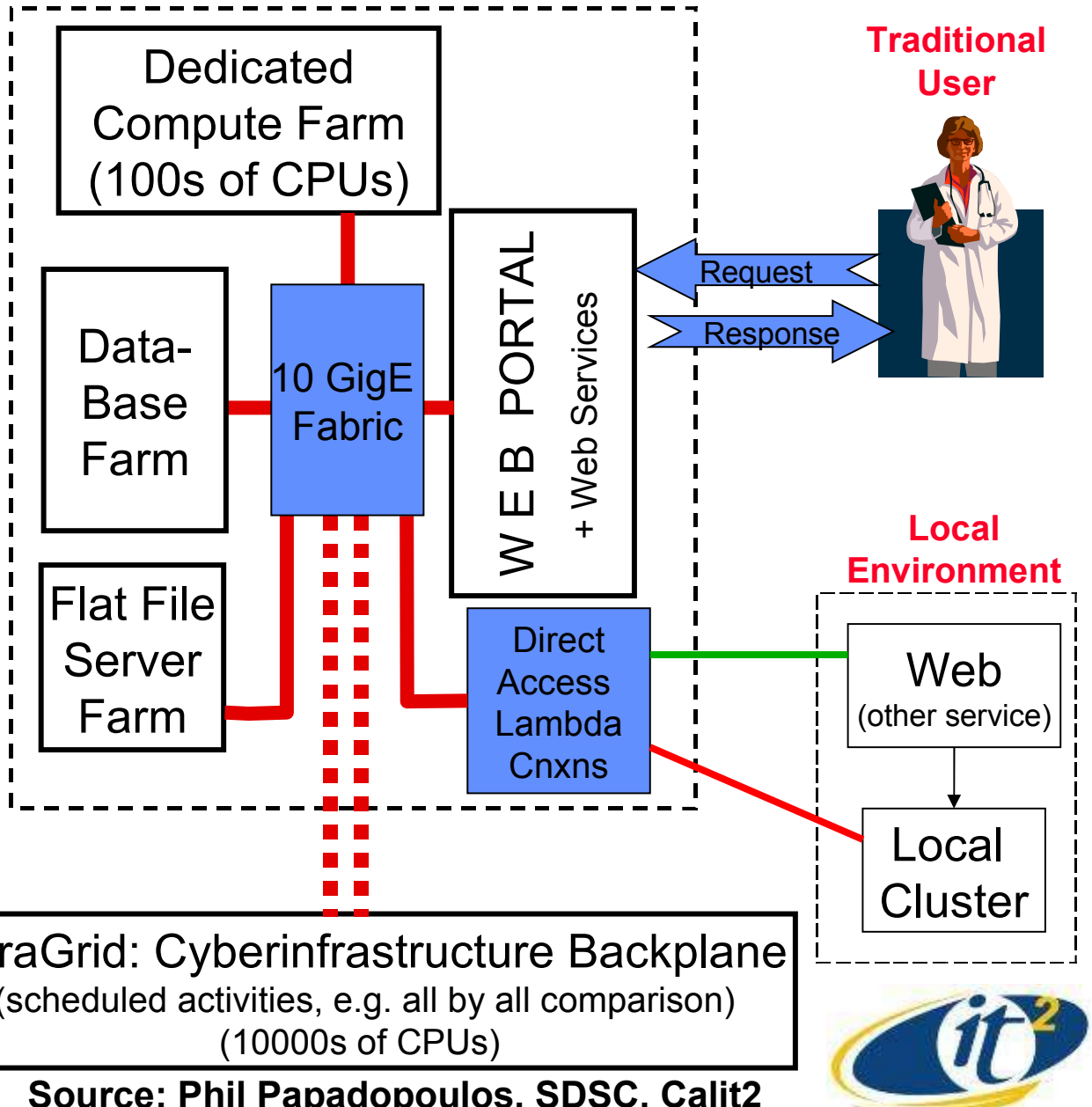
Sorcerer II Expedition (GOS)

JGI Community Sequencing Project

Moore Marine Microbial Project

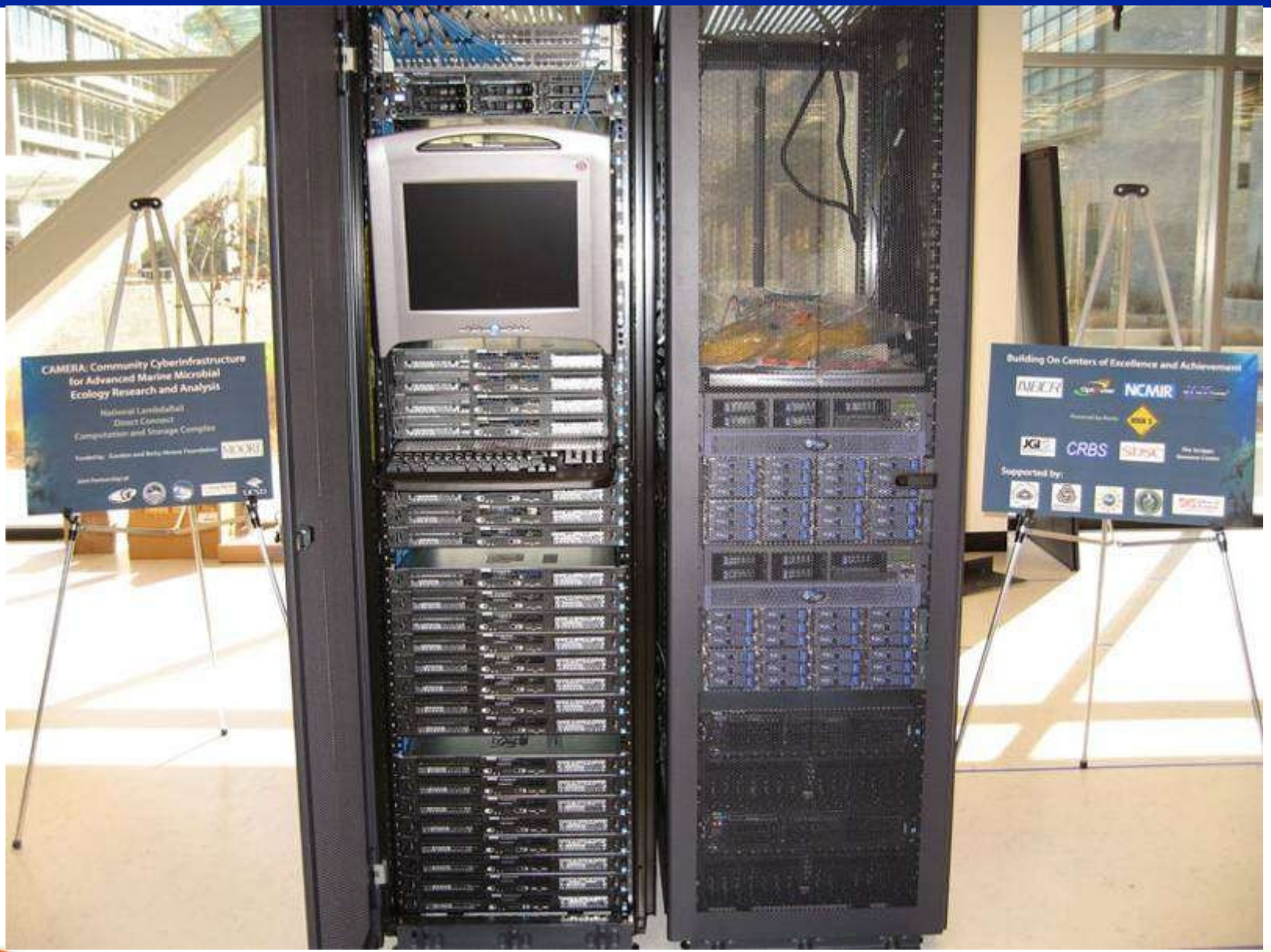
NASA Goddard Satellite Data

Community Microbial Metagenomics Data



Source: Phil Papadopoulos, SDSC, Calit2

First Implementation of the CAMERA Complex



Compute

Database & Storage

OptIPuter

it2

Analysis Data Sets, Data Services, Tools, and Workflows

- Assemblies of Metagenomic Data
  - e.g, GOS, JGI CSP
- Annotations
  - Genomic and Metagenomic Data
- "All-against-all" Alignments of ORFs
  - Updated Periodically
- Gene Clusters and Associated Data
  - Profiles, Multiple-Sequence Alignments,
  - HMMs, Phylogenies, Peptide Sequences
- Data Services
  - 'Raw' and Specialized Analysis Data
  - Rich Query Facilities
- Tools and Workflows
  - Navigate and Sift Raw and Analysis Data
  - Publish Workflows and Develop New Ones
  - Prioritize Features via Dialogue with Community

Source: Saul Kravitz
Director of Software Engineering
J. Craig Venter Institute

CAMERA Timeline

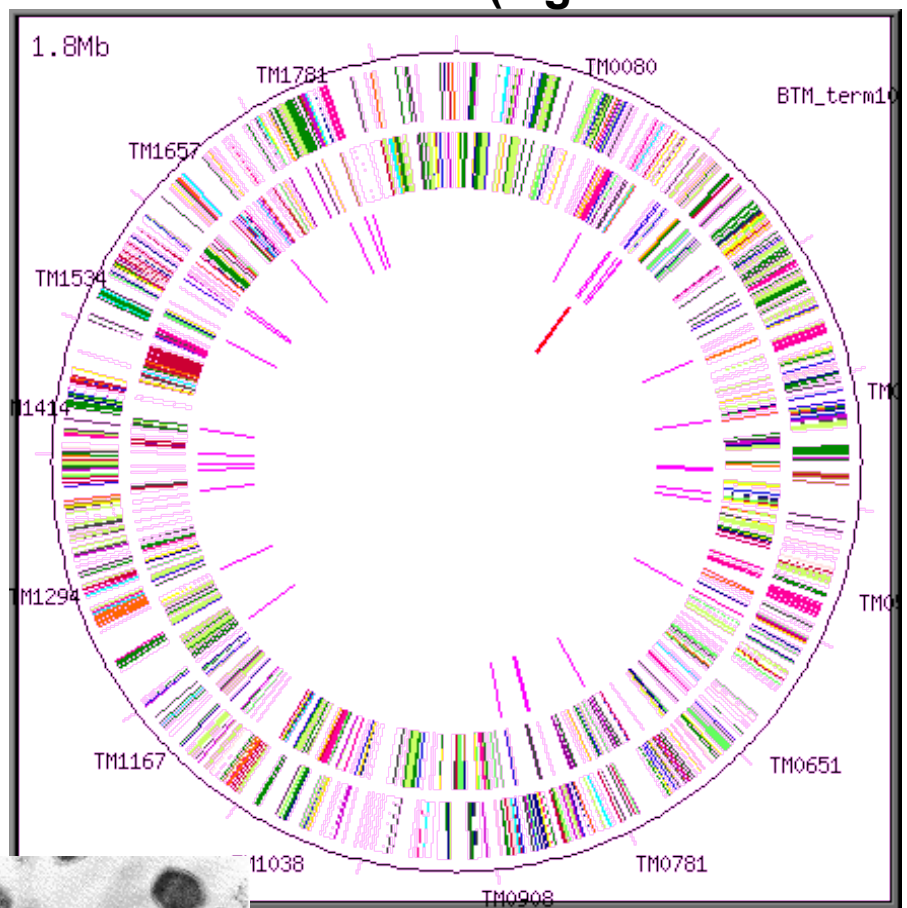
- **Release 1: Mid-2006**
  - Majority of GOS + Moore Microbe Genome Data
    - 6 Gbp Has Been Assembled
  - Initial Versions of Core Tools
    - BLAST, Reference Alignment Viewer
- **Release 2: Early-2007**
  - Additional Data
  - Additional/Improved Tools
  - Improved Usability
- **Subsequent**
  - Move Towards Semantic DB, Direct Access
  - Additional Tools & Data Based on Community Feedback

OptIPuter

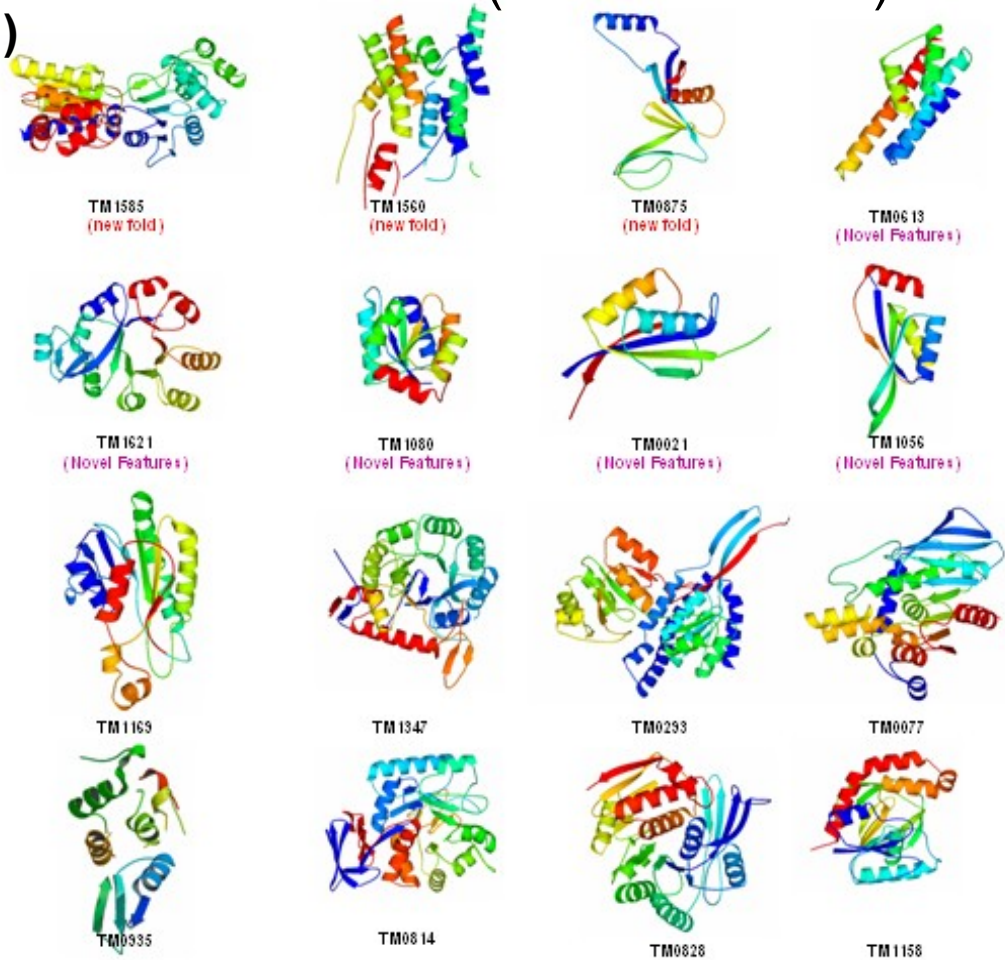
it²

The Bioinformatics Core of the Joint Center for Structural Genomics will be Housed in the Calit2@UCSD Building

Extremely Thermostable -- Useful for Many Industrial Processes (e.g. Chemical and Food)



173 Structures (122 from JCSG)





- Determining the Protein Structures of the Thermotoga Maritima Genome
- 122 T.M. Structures Solved by JCSG (75 Unique In The PDB)
- Direct Structural Coverage of 25% of the Expressed Soluble Proteins
- Probably Represents the Highest Structural Coverage of Any Organism

Source: John Wooley, UCSD

Providing Integrated Grid Software and Infrastructure for Multi-Scale BioModeling

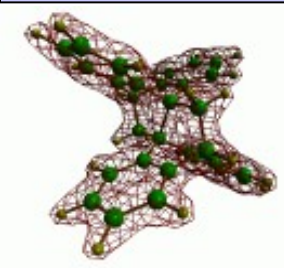
NBCR

National Biomedical Computation Resource

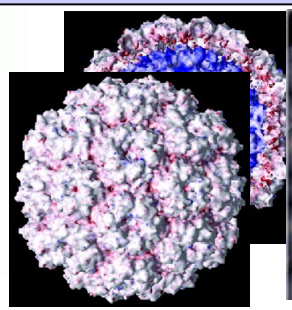
an NIH supported resource center

Located in Calit2@UCSD Building

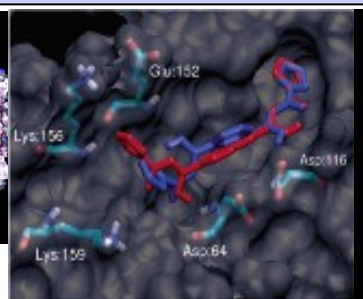
Grid and Cluster Computing Applications



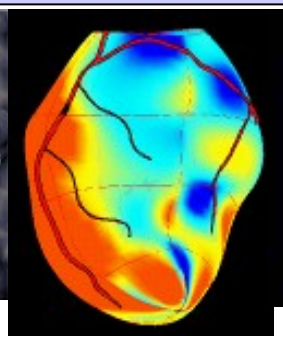
QMView
GAMESS



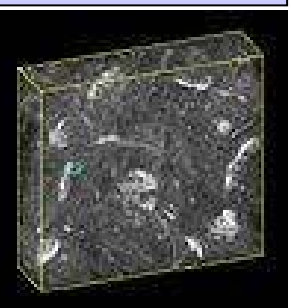
APBS



Autodock



Continuity



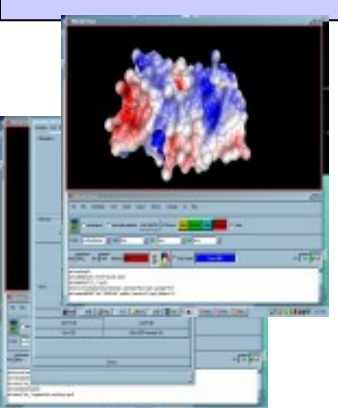
Gtomo2
TxBR

Infrastructure

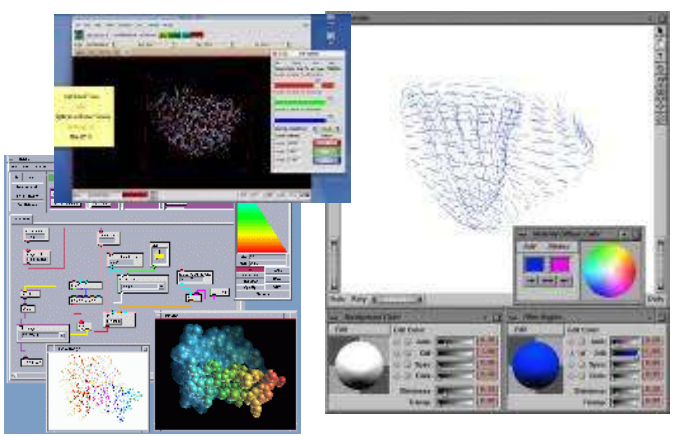


Rocks Grid of Clusters

Rich Clients



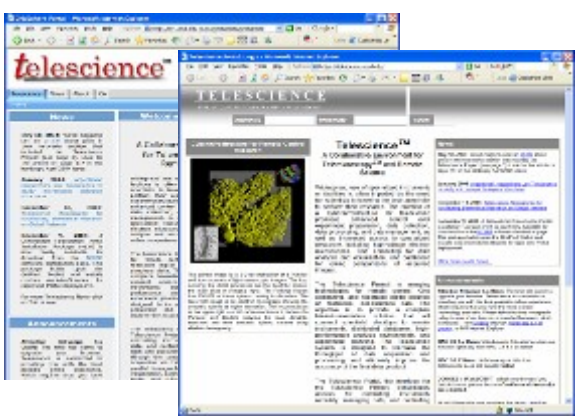
APBSCommand



PMV ADT
Vision

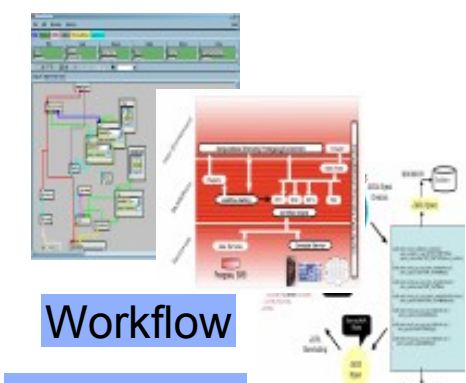
Continuity

Web Portal



Telescience Portal

Grid Middleware and Web Services



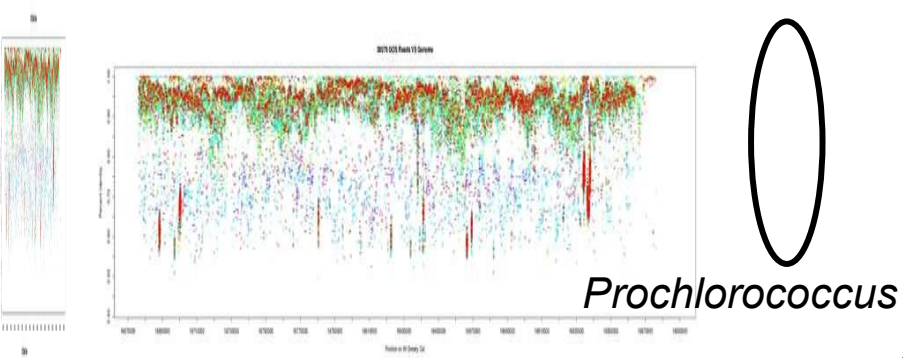
Workflow

Middleware

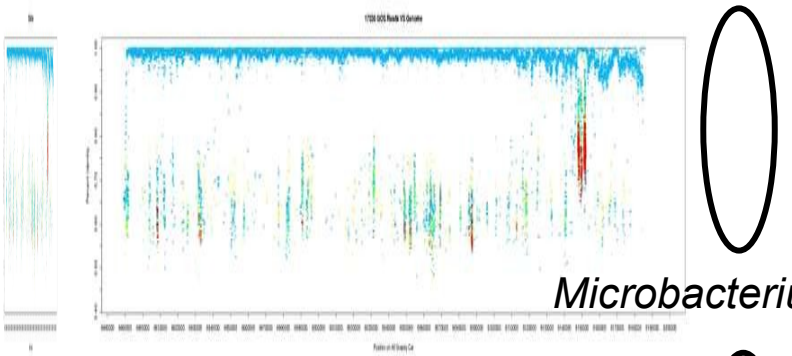
OptIPuter

it2

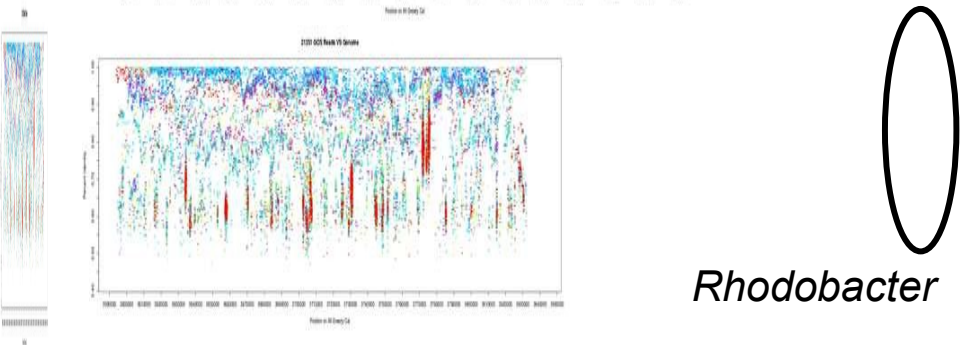
Metagenomics "Extreme Assembly" Requires Large Amount of Pixel Real Estate



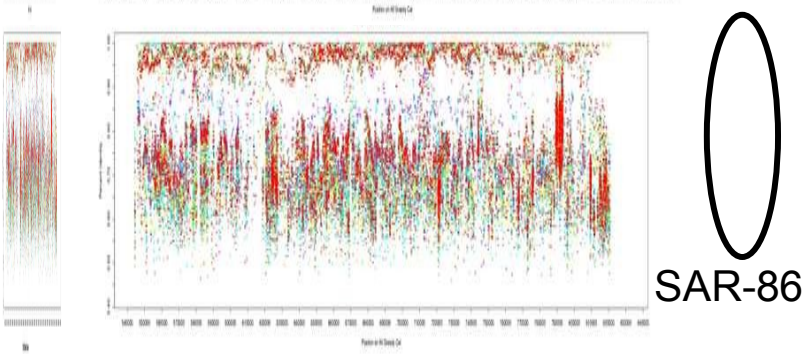
Prochlorococcus



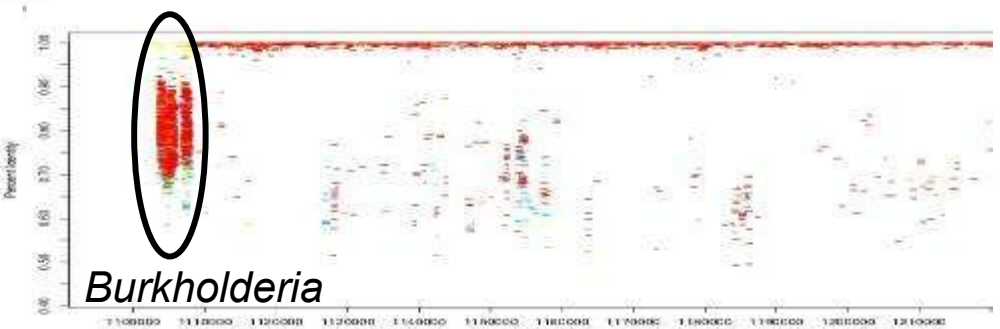
Microbacterium



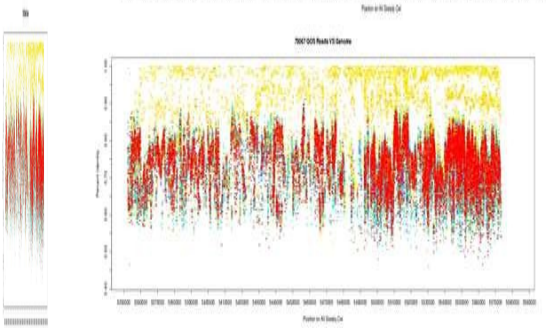
Rhodobacter



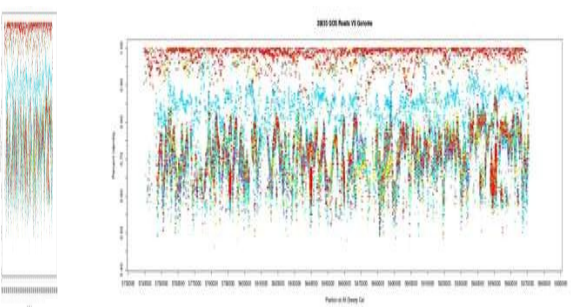
SAR-86



Burkholderia



unknown

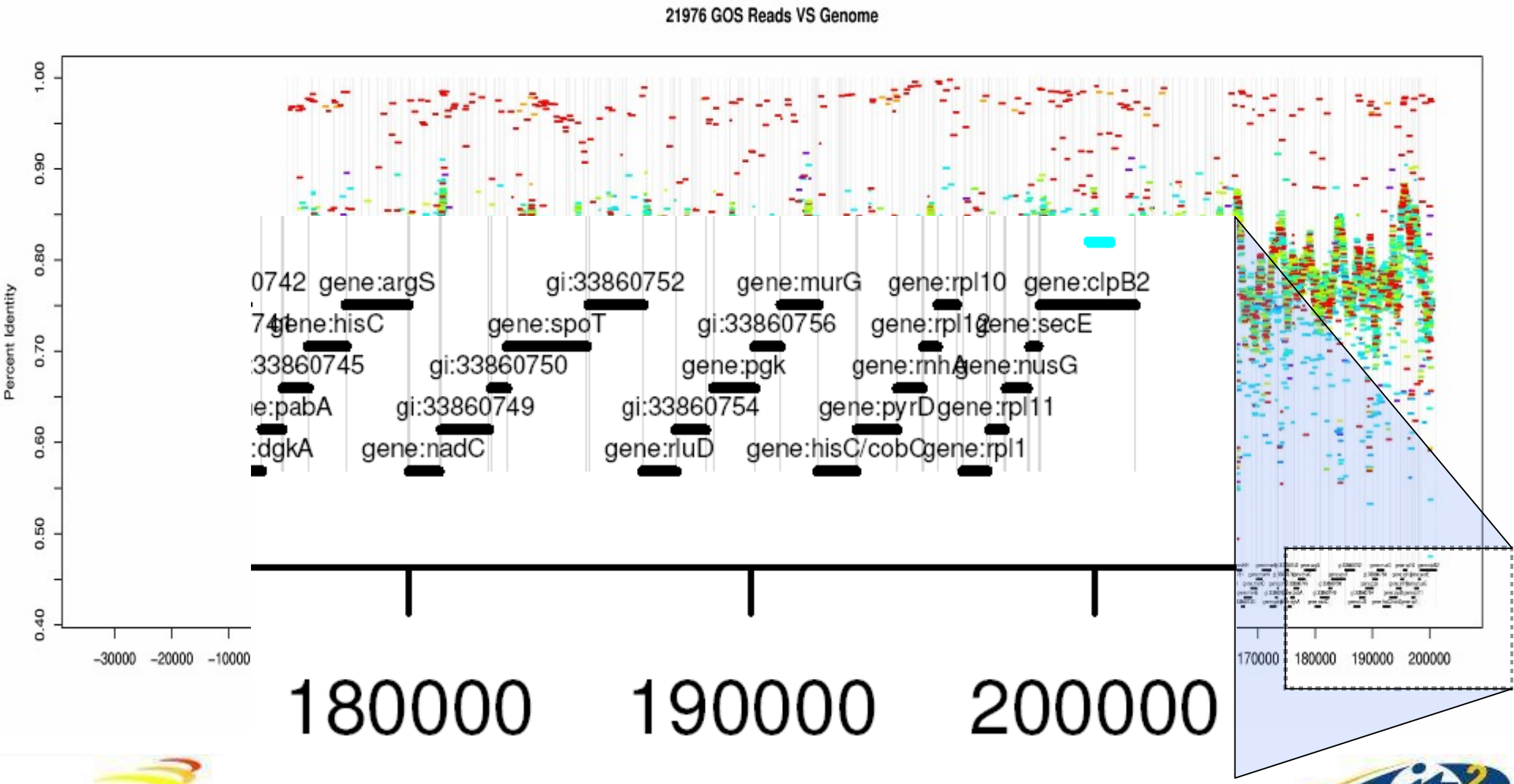


unknown

Source: Karin Remington
J. Craig Venter Institute

Metagenomics Requires a Global View of Data and the Ability to Zoom Into Detail Interactively

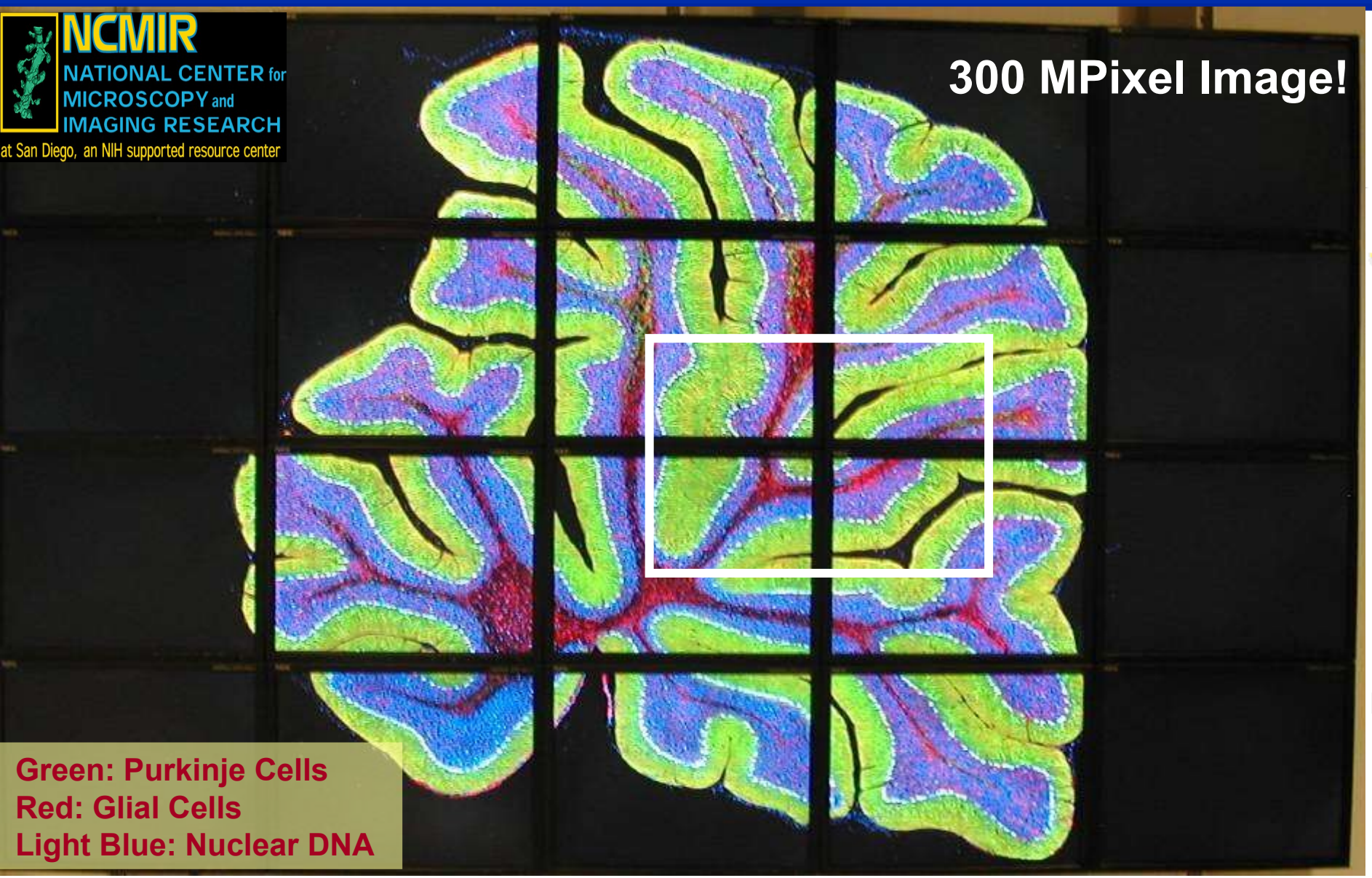
Overlay of Metagenomics Data onto Sequenced Reference Genomes (This Image: Prochloroccocus marinus MED4)



Source: Karin Remington
J. Craig Venter Institute

The OptIPuter – Creating High Resolution Portals Over Dedicated Optical Channels to Global Science Data

NCMIR
NATIONAL CENTER for MICROSCOPY and IMAGING RESEARCH
at San Diego, an NIH supported resource center



NSF

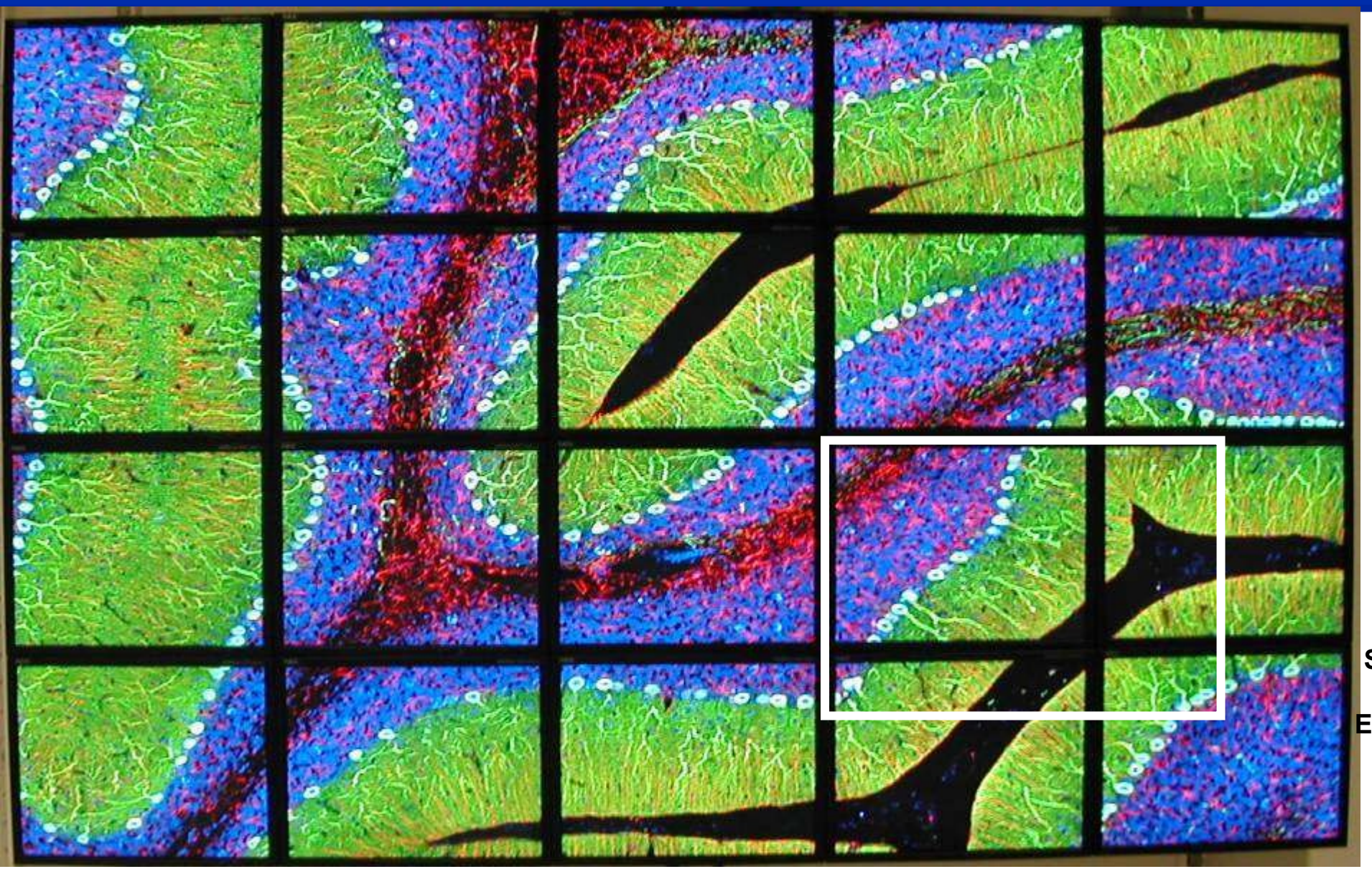
Source: Mark Ellisman, David Lee, Jason Leigh

OptIPuter

Calit2 (UCSD, UCI) and UIC Lead Campuses—Larry Smarr PI
Partners: SDSC, USC, SDSU, NW, TA&M, UvA, SARA, KISTI, AIST

it²

Scalable Displays Allow Both Global Content and Fine Detail



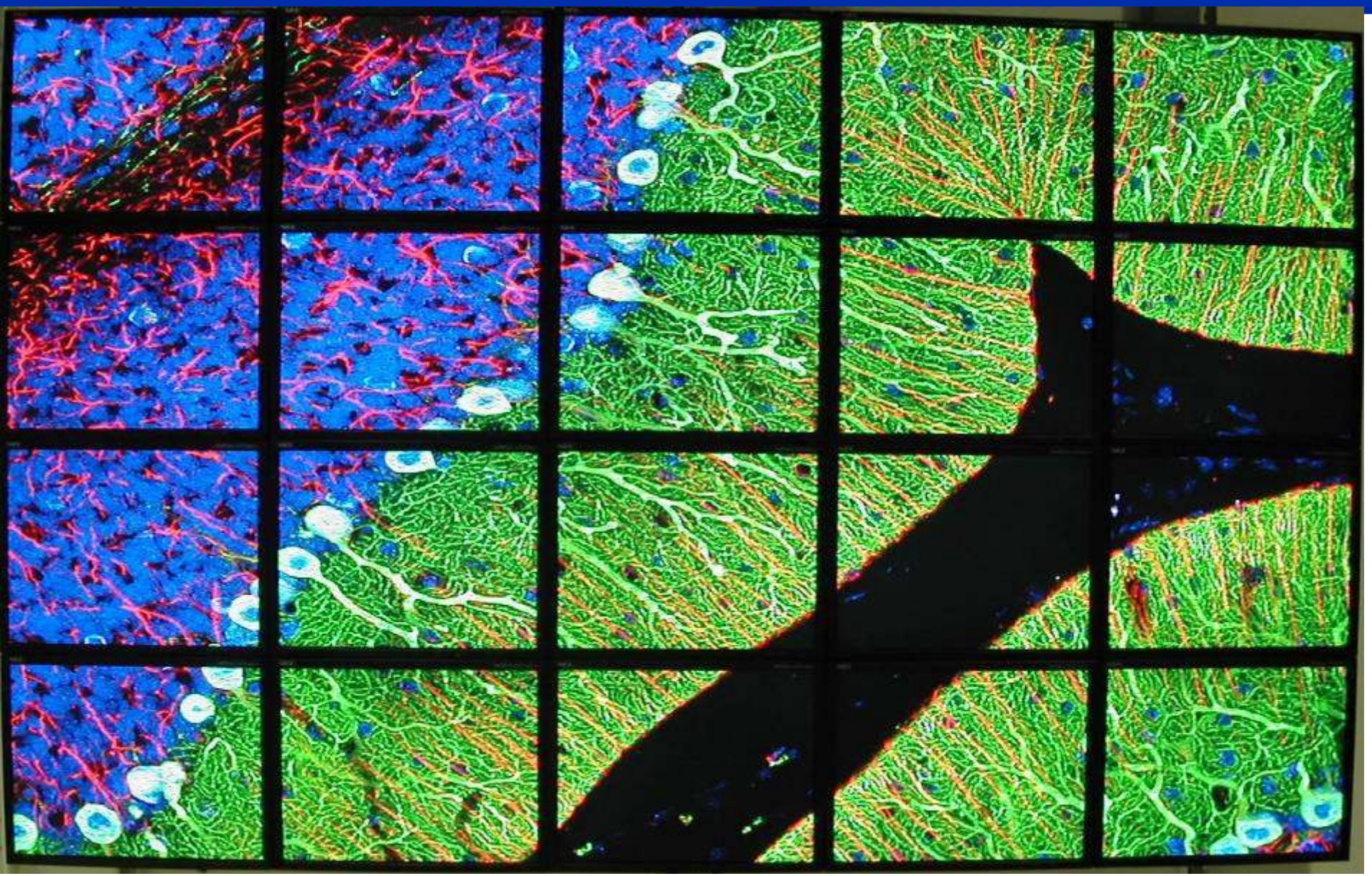
Source: Mark Ellisman, David Lee, Jason Leigh

OptIPuter

30 MPixel SunScreen Display Driven by a 20-node Sun Opteron Visualization Cluster

it²

Allows for Interactive Zooming from Cerebellum to Individual Neurons



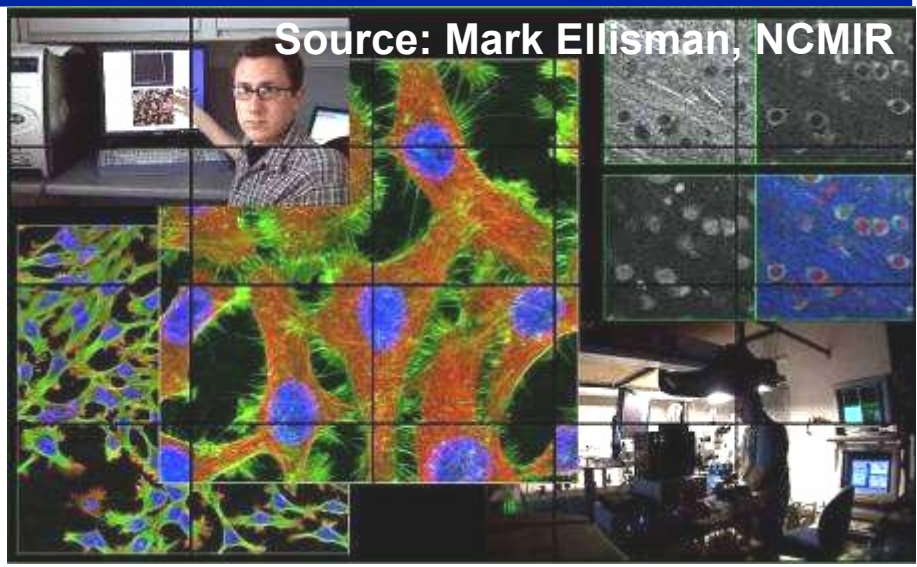
OptIPuter

Source: Mark Ellisman, David Lee, Jason Leigh

it²

The OptIPuter Enabled Collaboratory: Remote Researchers Jointly Exploring Complex Data





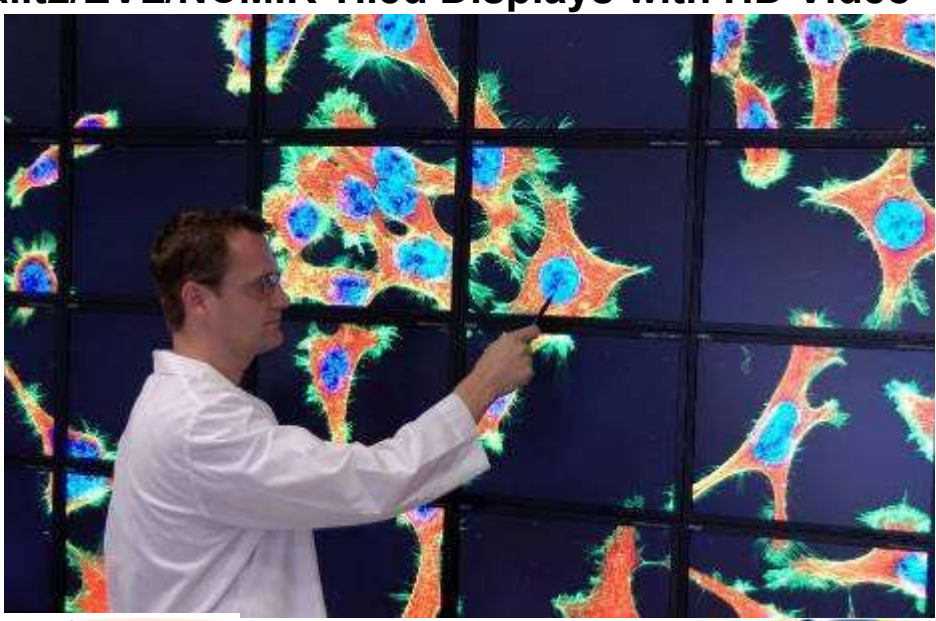
Source: Mark Ellisman, NCMIR

Calit2/EVL/NCMIR Tiled Displays with HD Video

New Home of SDSC/Calit2 Synthesis Center



Source: Chaitan Baru, SDSC



OptIPuter

SDSC SAN DIEGO SUPERCOMPUTER CENTER

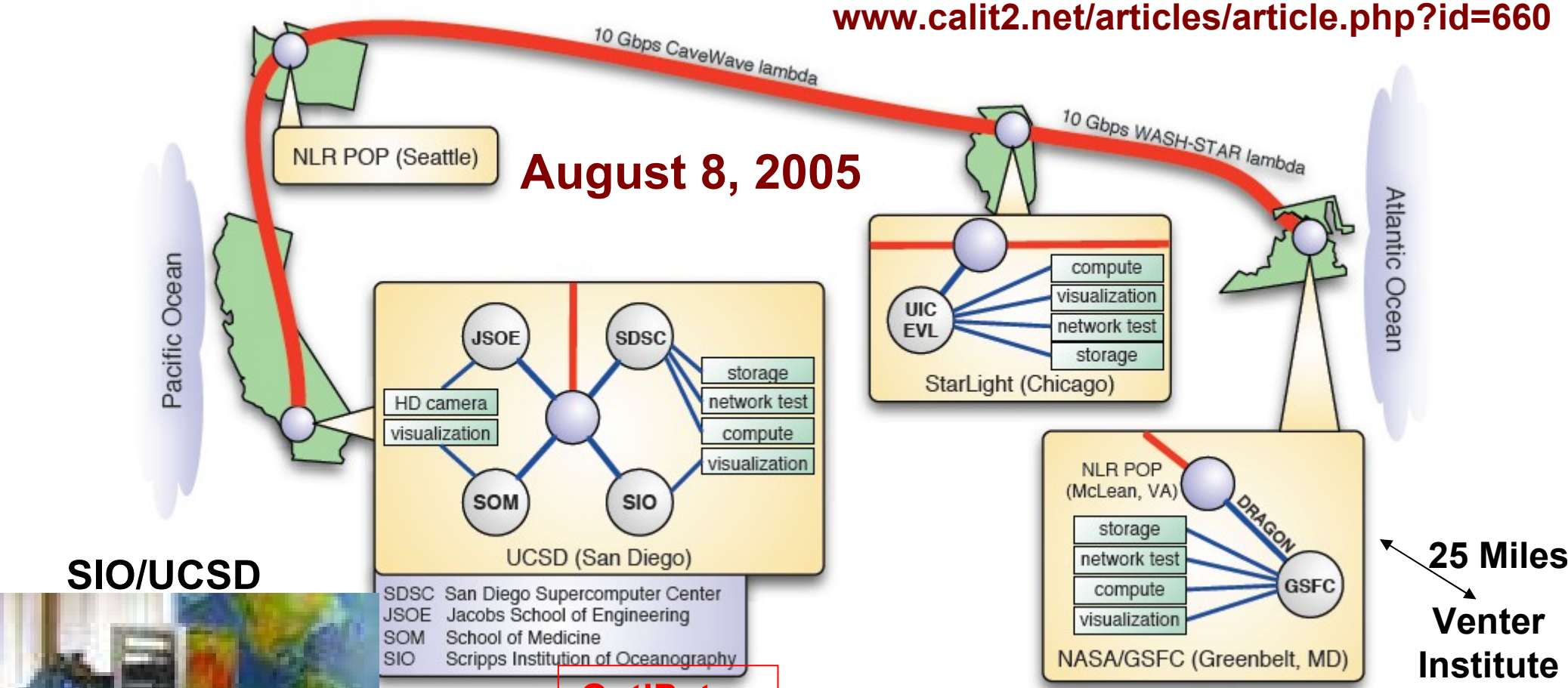
evl electronic visualization laboratory

NCMIR NATIONAL CENTER for MICROSCOPY and IMAGING RESEARCH at San Diego, an NIH supported resource center

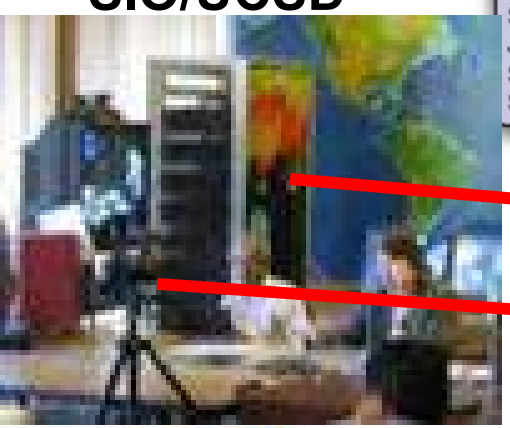
it2

Eliminating Distance to Unify Remote Laboratories

www.calit2.net/articles/article.php?id=660



SIO/UCSD



25 Miles

Venter Institute

OptIPuter Visualized Data

HDTV Over Lambda

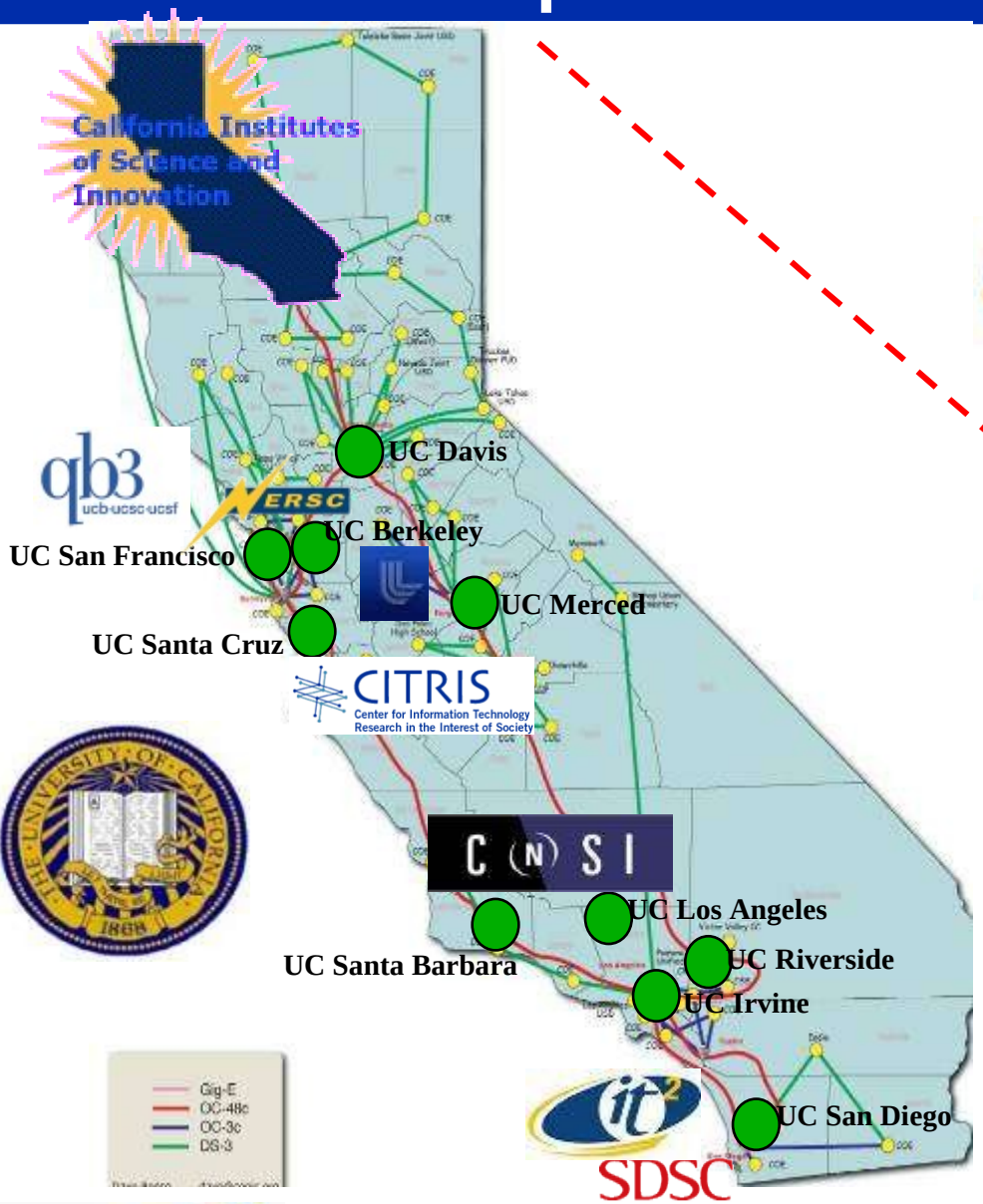


NASA Goddard

OptIPuter

it2

Calit2/SDSC Proposal to Create a UC Cyberinfrastructure of "On-Ramps" to National LambdaRail Resources



OptIPuter + CalREN-XD + TeraGrid = "OptiGrid"

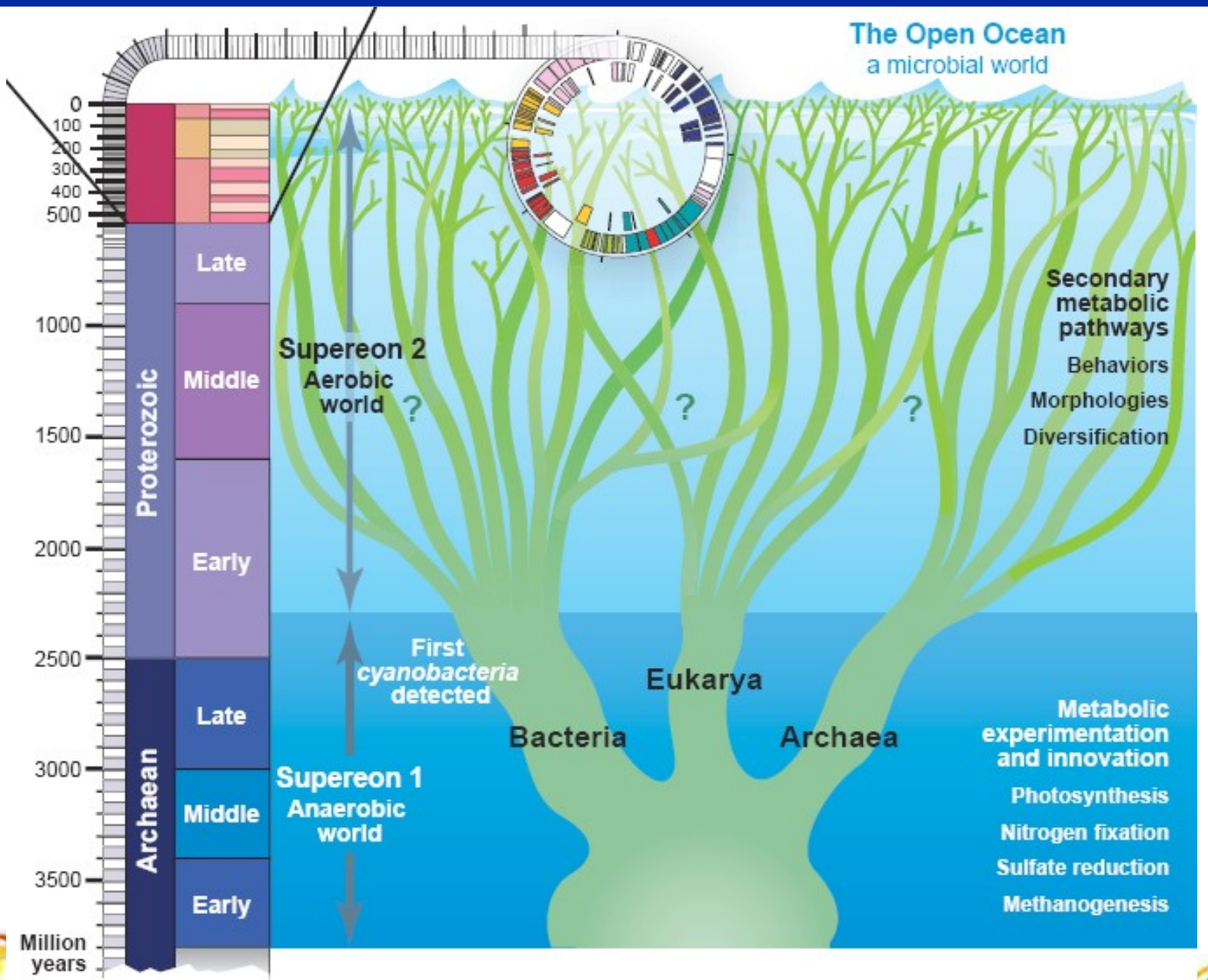


Creating a Critical Mass of End Users on a Secure LambdaGrid

OptIPuter

Source: Fran Berman, SDSC , Larry Smarr, Calit2

Looking Back Nearly 4 Billion Years In the Evolution of Microbe Genomics



Science Falkowski and Vargas 304 (5667): 58