Health Sciences Driving UCSD Research Cyberinfrastructure

Invited Talk

UCSD Health Sciences Faculty Council

UC San Diego

April 3, 2012

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

Follow me at http://lsmarr.calit2.net



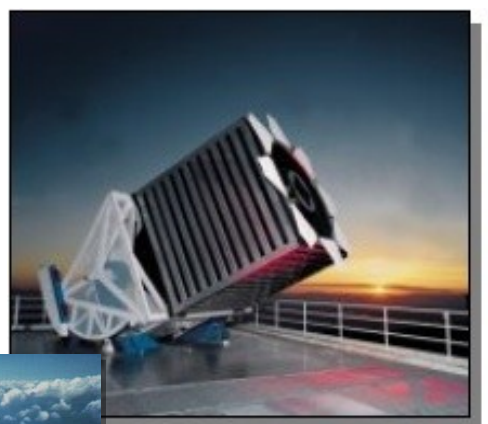
UCSD Researcher Research Cyberinfrastructure Needs

- **UCSD Researchers Surveyed in 2008 to Determine Their Unmet CI Needs**
- **Answer: DATA – Help!**
  - ***Data Infrastructure*** **(Storage, Transmission, Curation)**
  - ***Data Expertise*** **(Management, Analysis, Visualization, Curation)**

Diverse Sources of Data



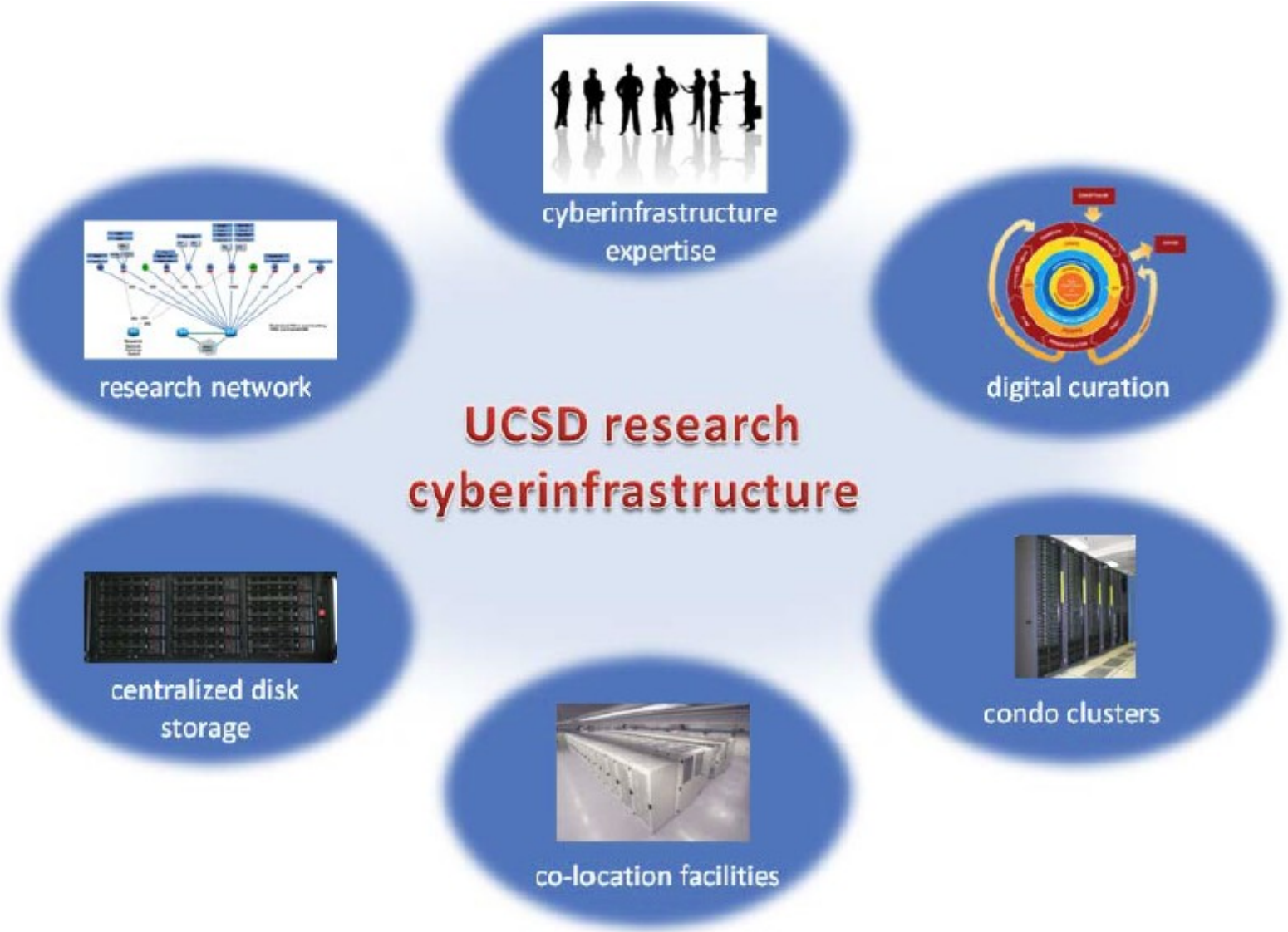






Source: Mike Norman, SDSC

"Blueprint for a Digital University"



Report 2009

http://rci.ucsd.edu

SDSC

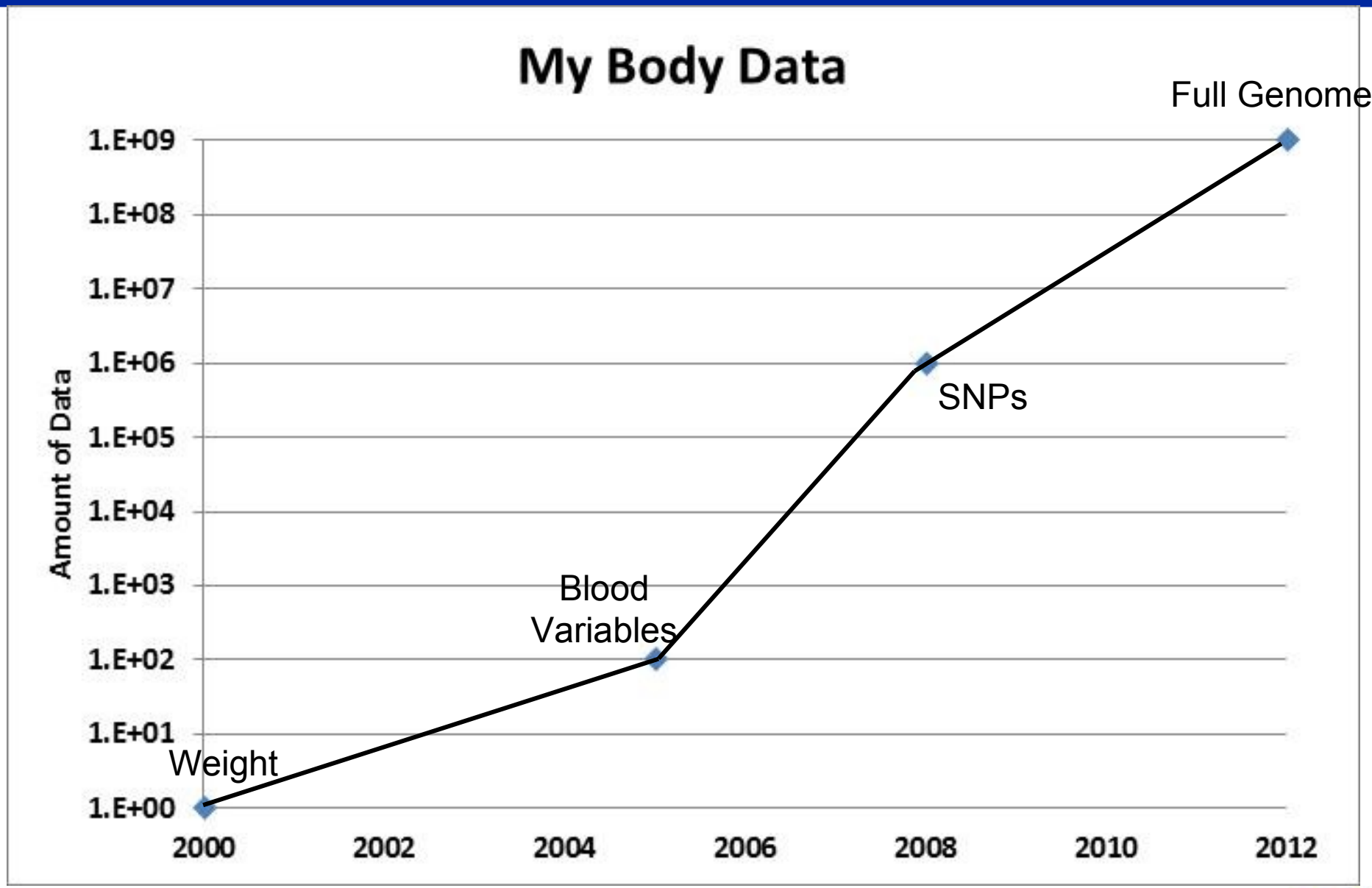
it²

UCSD RCI
Provider Organizations

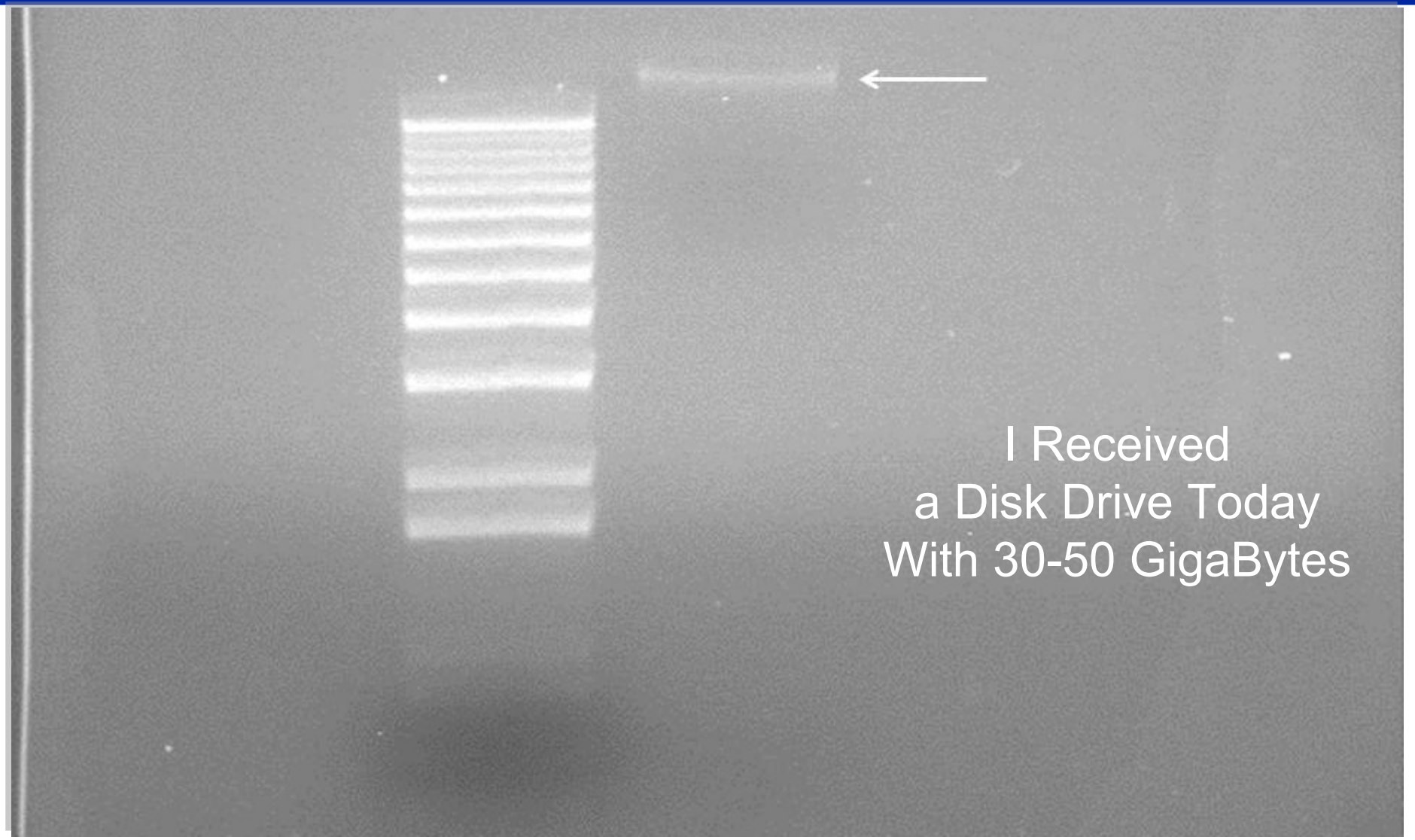
RCI element	SDSC	UCSD Libraries	ACT	Calit2
Co-Location	Lead			
Storage	Lead	Partner		Partner
Curation	Partner	Lead		
Computing	Lead			
Networking	Partner		Lead	Partner

Source: Mike Norman, SDSC

From One to a Billion Data Points Defining Me: The Exponential Rise in Body Data in Just One Decade



First Stage of Metagenomic Sequencing of My Gut Microbiome at J. Craig Venter Institute



Gel Image of Extract from Smarr Sample-Next is Library Construction
Manny Torralba, Project Lead - Human Genomic Medicine
J Craig Venter Institute
January 25, 2012

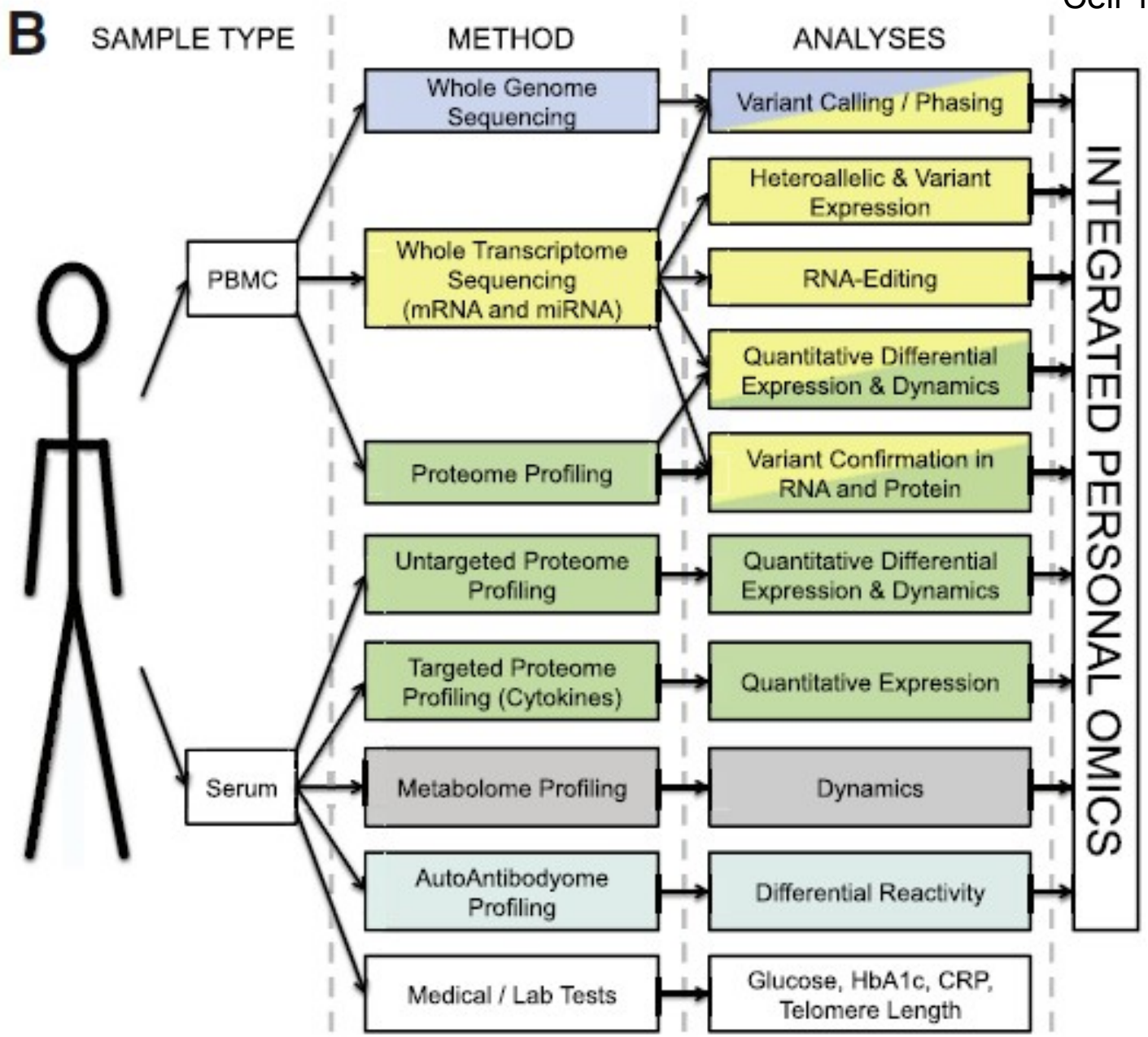
The Coming Digital Transformation of Health



www.technologyreview.com/biomedicine/39636

Integrative Personal Omics Profiling Reveals Details of Clinical Onset of Viruses and Diabetes

Cell 148, 1293–1307, March 16, 2012



- **Michael Snyder, Chair of Genomics Stanford Univ.**
- **Genome 140x Coverage**
- **Blood Tests 20 Times in 14 Months**
  - **tracked nearly 20,000 distinct transcripts coding for 12,000 genes**
  - **measured the relative levels of more than 6,000 proteins and 1,000 metabolites in Snyder's blood**

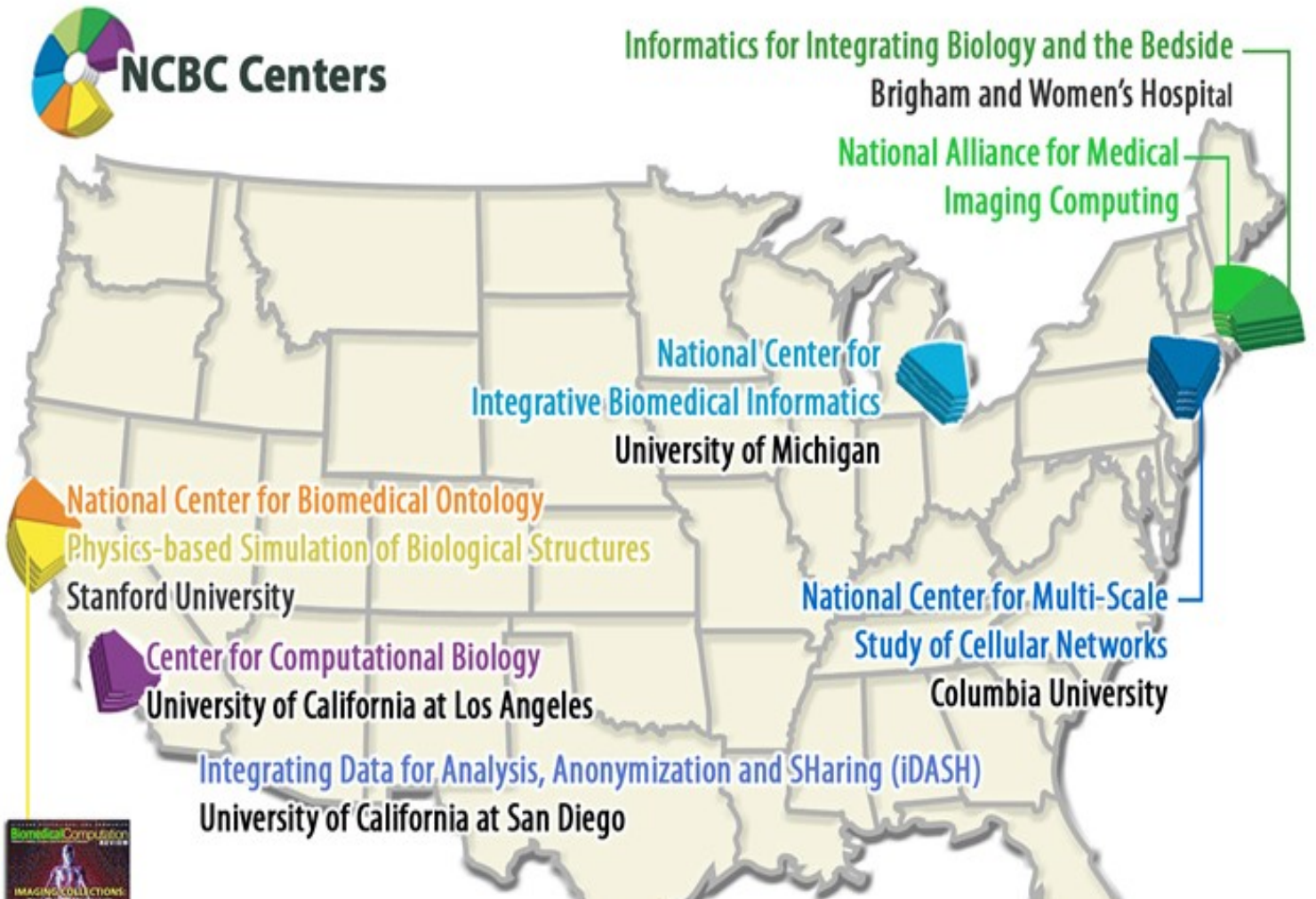
NIH Roadmap

Source: Lucila Ohno-Machado, UCSD SOM

National Centers for Biomedical Computing

Home | NCBC Summary | Calendar | All Hands Meetings | Biological Projects | Biositemaps Projects | Working Group Archive

Search for NCBC resources in the new Resource Discovery System (RDS)





Outcome of NIH Botstein-Smarr Report (1999)

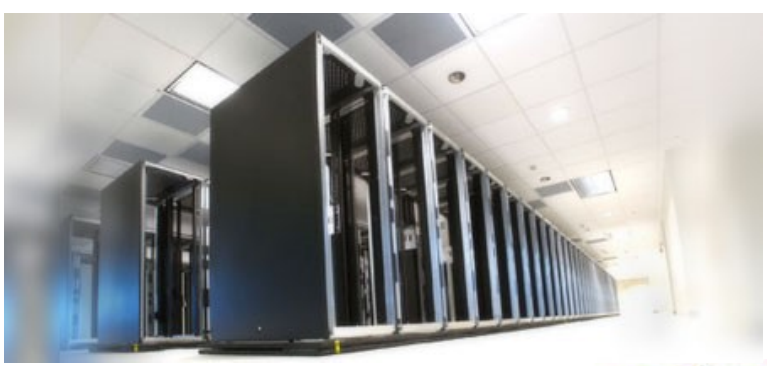
http://acd.od.nih.gov/agendas/060399_Biomed_Computing_WG_RPT.htm

integrating Data for Analysis, Anonymization, and SHaring (iDASH)

- **Data Exported for Computation Elsewhere**
  - Users download data from iDASH
- **Computation Comes to the Data**
  - Users access data in iDASH
  - Users upload algorithms into iDASH
- **iDASH Exportable Cyberinfrastructure**
  - Users download infrastructure



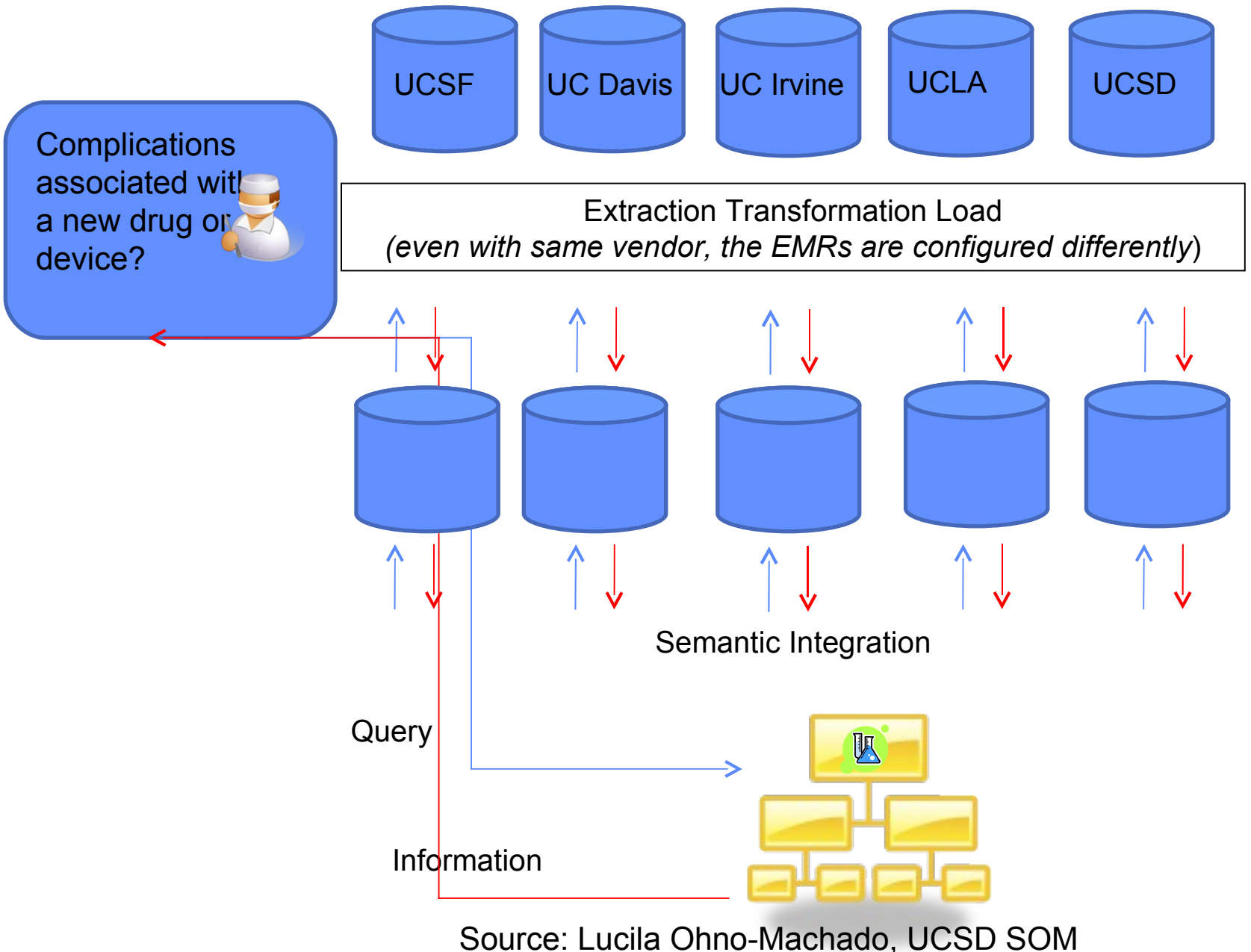
Private Cloud at SD Supercomputer Center
Medical Center Data Hosting
HIPAA certified facility



Source: Lucila Ohno-Machado, UCSD SOM

funded by NIH U54HL108460

Data + Ontologies + Tools





Source: Lucila Ohno-Machado, UCSD SOM

Personalized Care and Population Health

- **Genomics**
  - SNP-based therapy (cancer)
- **'Phenomics'**
  - Electronic Health Records
  - Personal monitoring
    - Blood pressure, glucose
  - Behavior
    - Adherence to medication, exercise
- **Public Health and Environment**
  - Air quality, food
  - Surveillance



Source: DOE

Source: Lucila Ohno-Machado, UCSD SOM

NCMIR's Integrated Infrastructure of Shared Resources

NCMIR

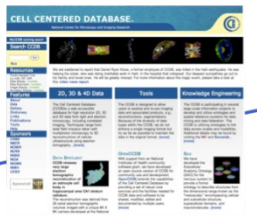
Telemicroscopy Resources:
Osaka Univ., Oxford Univ.,
The Jackson Labs

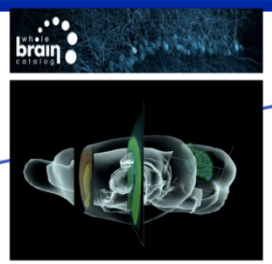






Neuroinformatics Resources





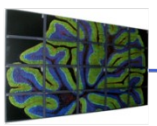


Shared Infrastructure

Development Systems

Storage Resources

OptIPortal Tiled Display







California Institute of Telecommunications and Information Technology

UCSD

NBCR

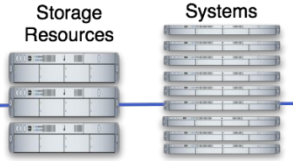
SDSC

CRBS

Storage Resources

Production Systems

Triton Petascale Resource





San Diego Supercomputer Center

Scientific Instruments

Ultra-wide Field Intra-vital LM Instruments



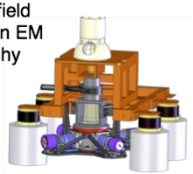
Super-resolution PALM



Serial Blockface Scanning EM



Ultra-wide field Serial Section EM Tomography



Multi-color EM, EFTEM



Wide-field STEM/TEM

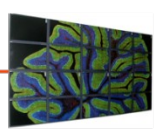


End User Workstations

Graphics Workstations



OptIPortal Tiled Display



Local SOM Infrastructure

NCMIR Data Storage Servers



Critical Servers (Mail, Web, DB)



Large Memory Servers





NBCR Compute Clusters



GPU Compute Clusters

SDSC

it²

Source: Steve Peltier, NCMIR

Ideker Lab Workflow

Leichtag/Sequencer



Storage



Skaggs/Users

















Calit2/Storage

SDSC/Triton

SDSC

Source: Chris Misleh, Calit2/SOM

it²

Next Generation Genome Sequencers Produce Large Data Sets

Data Volume	Total	Final	Comment
HiSeq 2000 200G run			
Image Data	32 TB	0	
Intensity Data	2 TB	0	Optionally transferred
Base Call / Quality Score Data	250 GB	250 GB	1 byte/base (raw) assuming qseq generation offline
Alignment Output	6 TB (3 TB)	1.2 TB	Remove intermediate files
GA_{IIx} 50G run			
Image Data	6.9	0	Optionally transferred
Intensity Data	0.93	0.93	
Base Call / Quality Score Data	0.17	0.17	
Alignment Output	1.2TB	1.2 TB	

Source: Chris Misleh, SOM

Moving to Shared Enterprise Data Storage & Analysis Resources: SDSC Triton Resource & Calit2 GreenLight

http://tritonresource.sdsc.edu

Source: Philip Papadopoulos, SDSC, UCSD

SDSC
Large Memory
Nodes
- 256/512 GB/sys
- 8TB Total
- 128 GB/sec
- ~ 9 TF



x28



x256

SDSC Shared
Resource
Cluster
- 24 GB/Node
- 6TB Total
- 256 GB/sec
- ~ 20 TF

ROCKS



SDSC Data Oasis
Large Scale Storage
- 2 PB
- 50 GB/sec
- 3000 – 6000 disks
- Phase 0: 1/3 PB, 8GB/s



UCSD Research Labs





N x 10Gb/s



Calit2 GreenLight

Campus
Research
Network

SDSC

it²

SOM Use of SDSC Triton Resource

- **10 SOM PIs Received Substantial Allocations**
  - **100K CPU-hours or more**
- **8 SOM PIs / Labs Currently Using Triton with Time Purchased from Grant Funds**
- **30+ Active Trial Accounts**
- **Supporting ~6 Next Generation Sequencing Projects with PIs from SOM, SIO, and 2 Outside Research Institutes (TSRI, LIAI)**

SDSC

it²

Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis

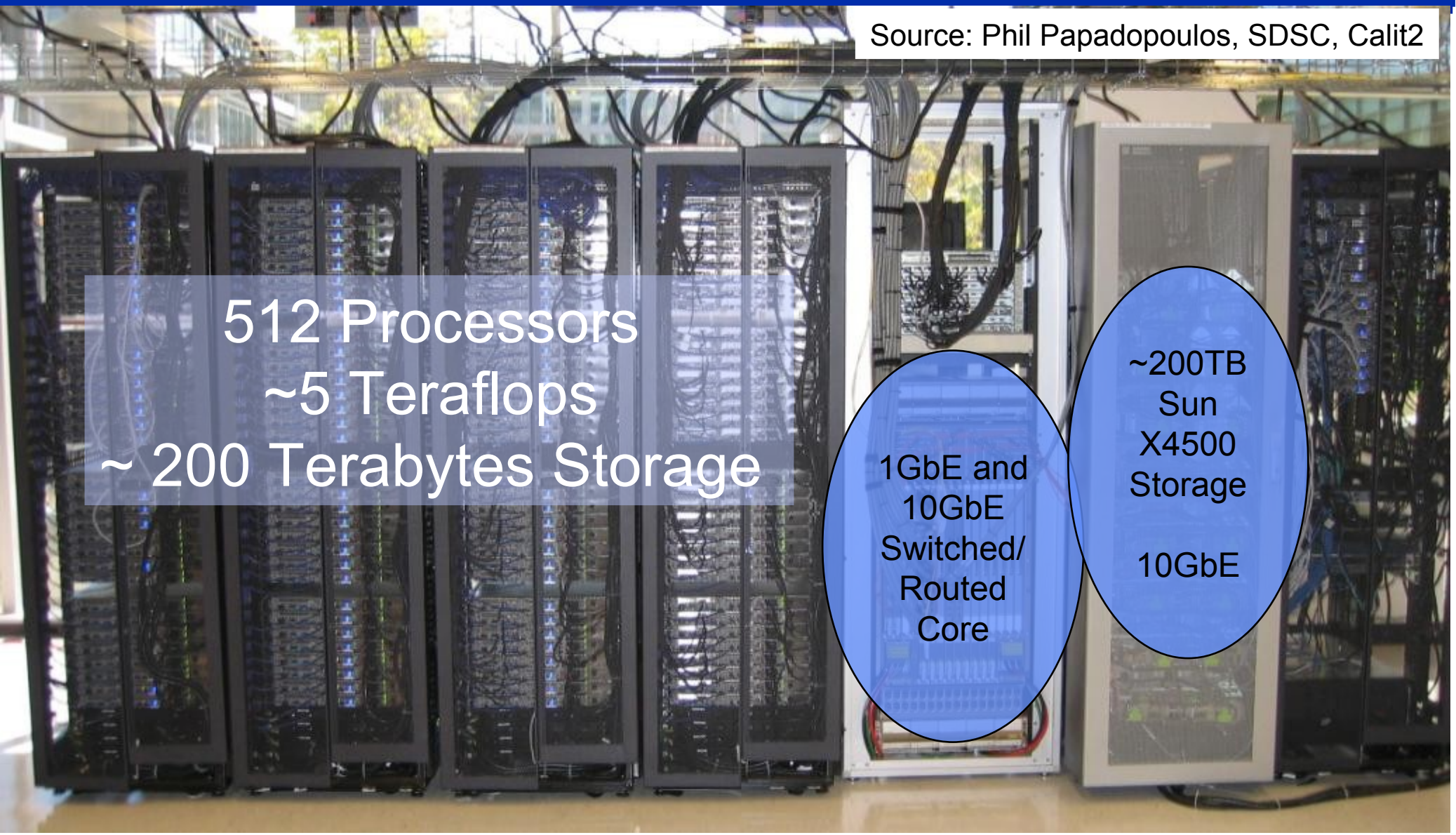


SDSC

Gordon and Betty MOORE FOUNDATION

it²

Calit2 Microbial Metagenomics Cluster- Next Generation Optically Linked Science Data Server



4000 Users
From 90 Countries

Gordon and Betty MOORE FOUNDATION

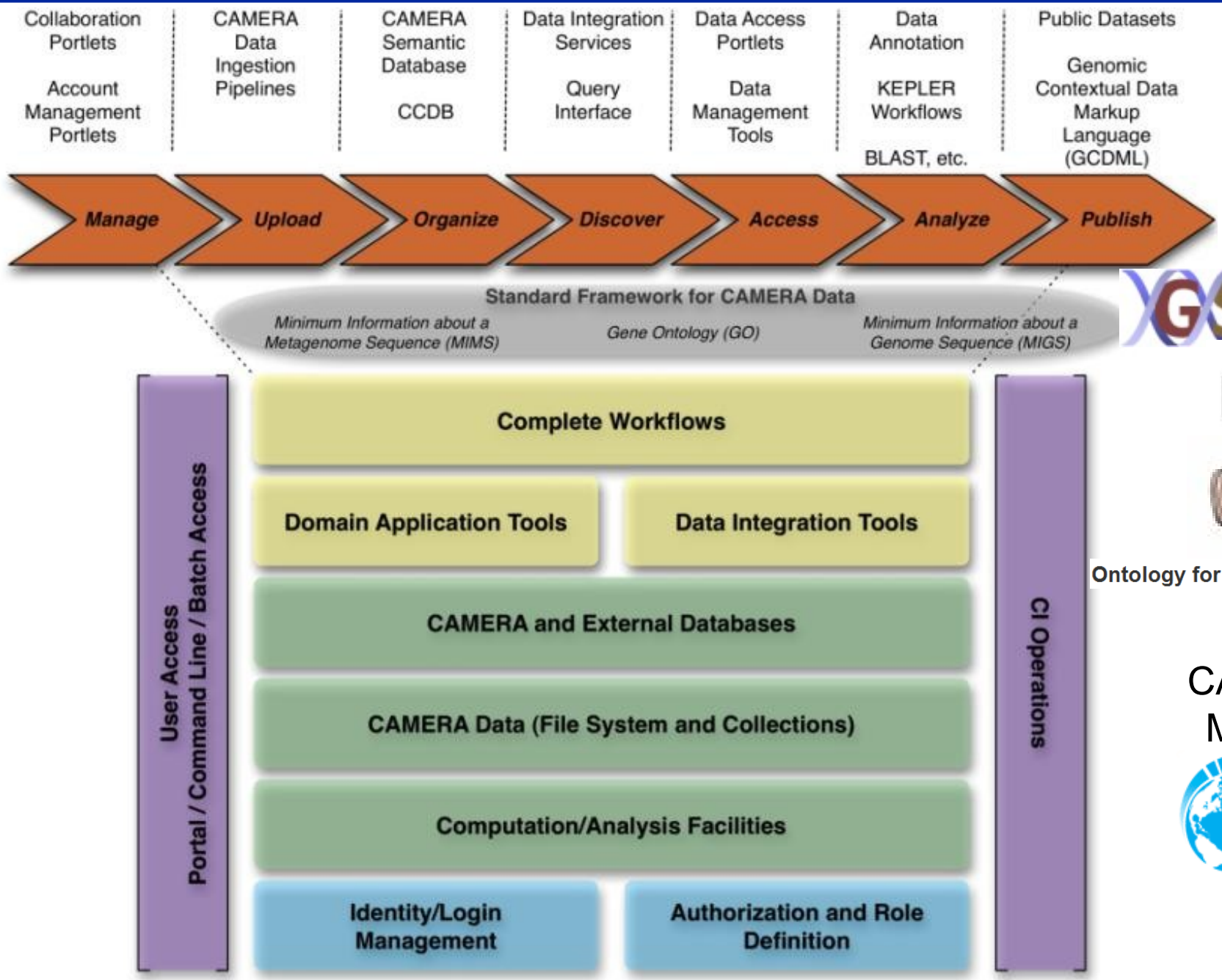
ROCKS

SDSC

J. Craig Venter INSTITUTE

it²

Creating CAMERA 2.0 -
Advanced Cyberinfrastructure Service Oriented Architecture



GSC Genomic Standards Consortium

Envo

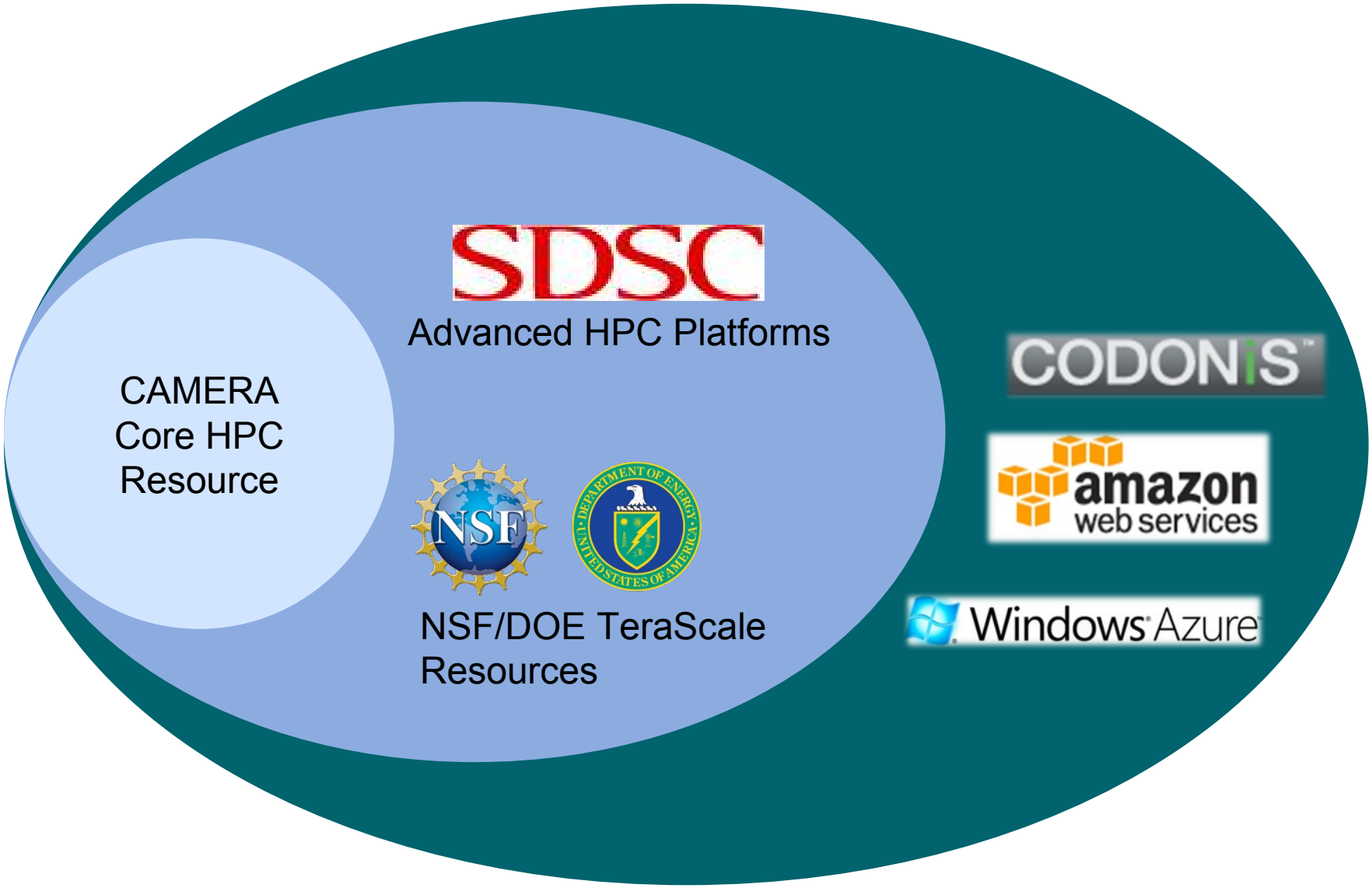
OBI
Ontology for Biomedical Investigations

Source:
CAMERA CTO
Mark Ellisman

camera
Marine Microbial Ecology

it2

Access to Computing Resources Tailored by User's Requirements and Resources



Source: Jeff Grethe, CAMERA

NSF Funds a Data-Intensive Track 2 Supercomputer: SDSC's Gordon-Coming Summer 2011

- **Data-Intensive Supercomputer Based on SSD Flash Memory and Virtual Shared Memory SW**
  - *Emphasizes MEM and IOPS over FLOPS*
  - *Supernode has Virtual Shared Memory:*
    - *2 TB RAM Aggregate*
    - *8 TB SSD Aggregate*
    - *Total Machine = 32 Supernodes*
    - *4 PB Disk Parallel File System >100 GB/s I/O*



- **System Designed to Accelerate Access to Massive Data Bases being Generated in Many Fields of Science, Engineering, Medicine, and Social Science**

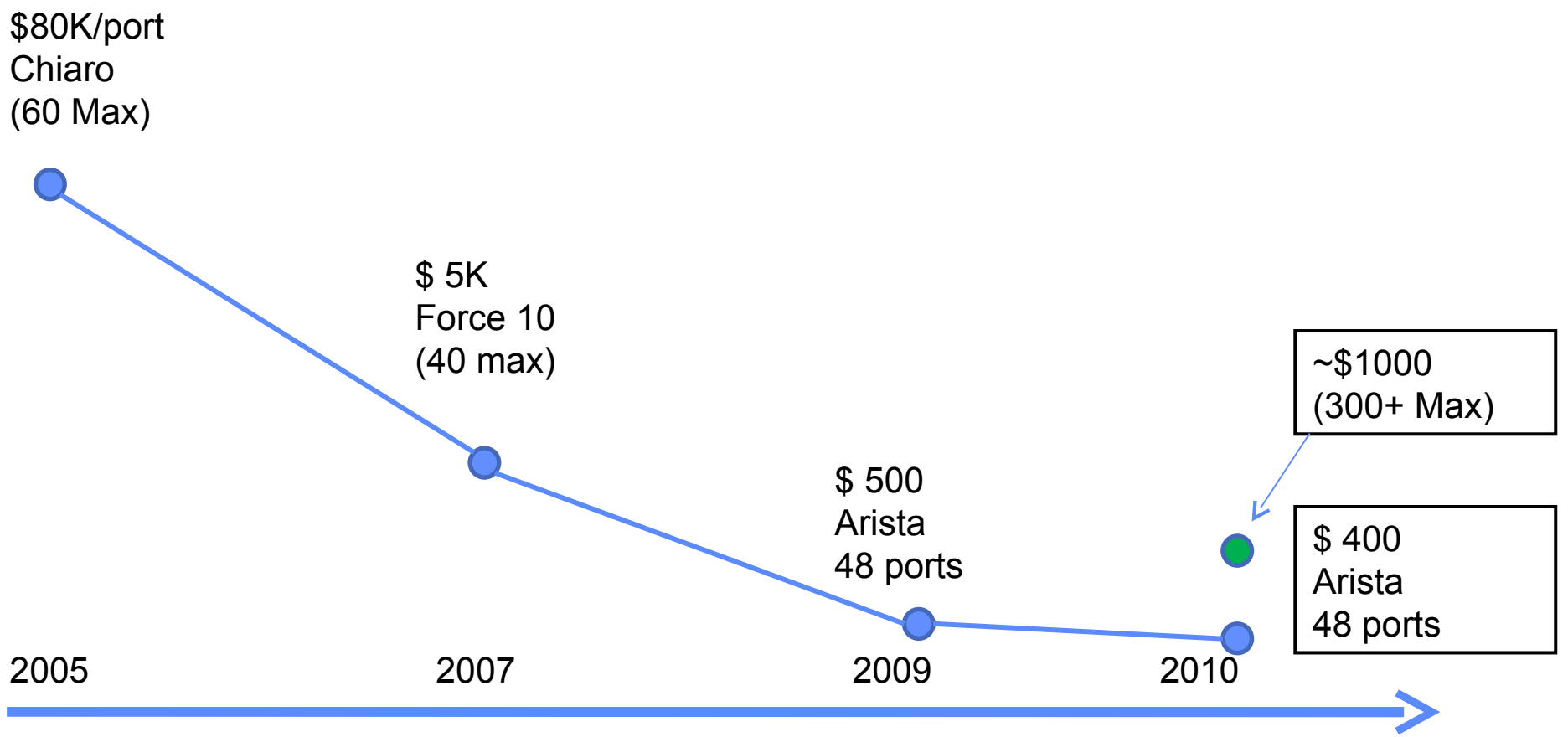
SDSC

Source: Mike Norman, Allan Snavely SDSC

it²

Rapid Evolution of 10GbE Port Prices Makes Campus-Scale 10Gbps CI Affordable

- Port Pricing is Falling
- Density is Rising – Dramatically
- Cost of 10GbE Approaching Cluster HPC Interconnects

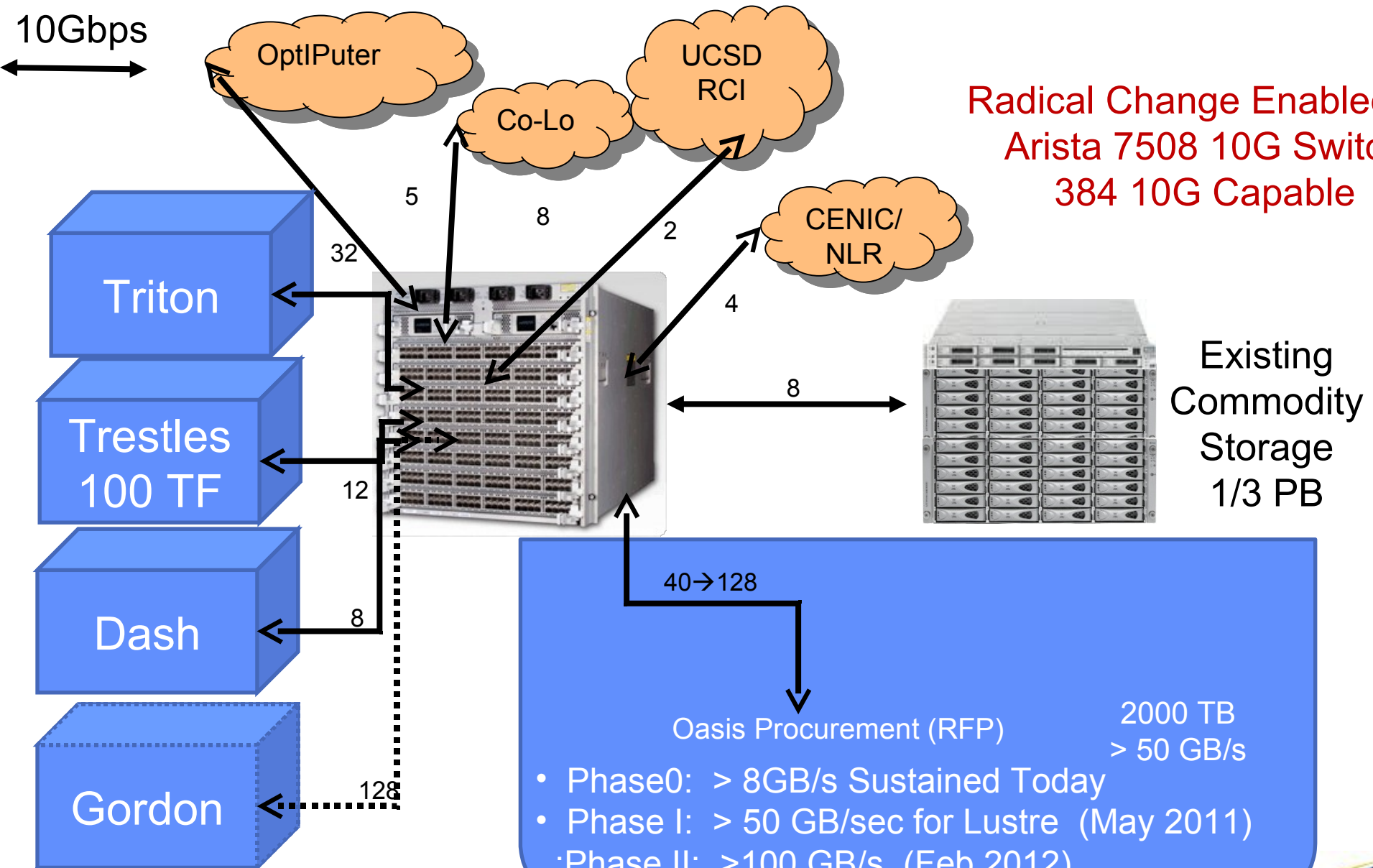


SDSC

Source: Philip Papadopoulos, SDSC/Calit2

it²

10G Switched Data Analysis Resource: SDSC's Data Oasis – Scaled Performance



Radical Change Enabled by Arista 7508 10G Switch 384 10G Capable

Existing Commodity Storage 1/3 PB

SDSC

Source: Philip Papadopoulos, SDSC/Calit2

it²

2012 RCI Initiatives

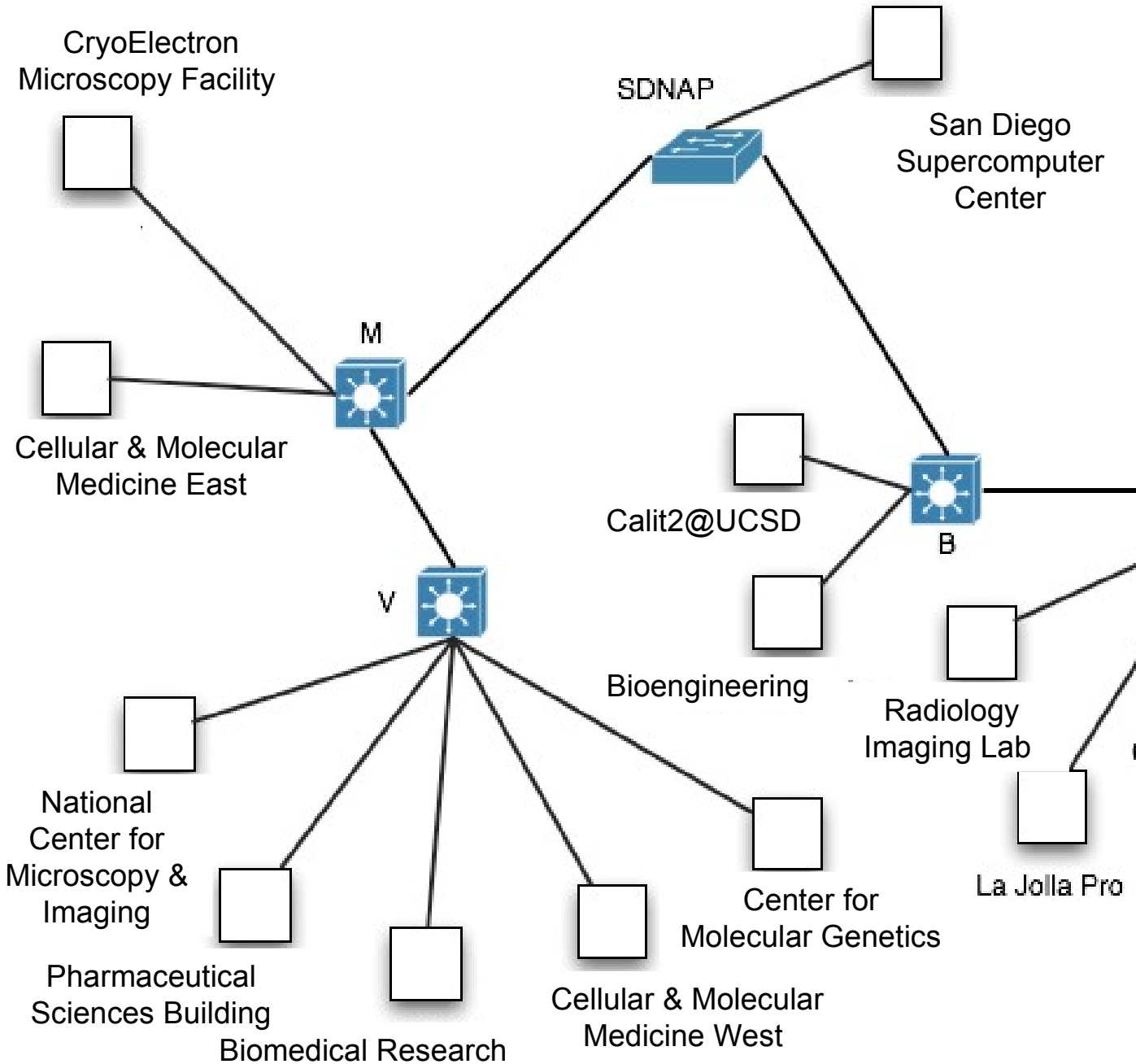
- **RCI is Preparing an Attractive Storage Offering for All UCSD Researchers to Encourage Adoption**
  - **"Wide and Deep"**
  - **On-Ramp to Digital Curation Efforts**
- **SOM Possesses Many of the Most Data-Intensive Instruments on Campus (NGS, MassSpec, MRI)**
  - **Effort to Connect Them to RCI Resources This Year**
- **SDSC Working with DBMI to Define a HIPPA-compliant Cloud Computing Resource that Would Leverage or Extend RCI Resources**
- **RCI Implementation Team Needs your Input and Collaboration (email Richard Moore @ SDSC)**

SDSC

Source: Mike Norman, SDSC

it²

Potential UCSD Optical Networked Biomedical Researchers and Instruments



- **Connects at 10 Gbps :**
  - **Microarrays**
  - **Genome Sequencers**
  - **Mass Spectrometry**
  - **Light and Electron Microscopes**
  - **Whole Body Imagers**
  - **Computing**
  - **Storage**

Developing Detailed Plan