

LSST and the Cloud: Astro Collaboration in 2016

Tim Axelrod
LSST Data Management Scientist

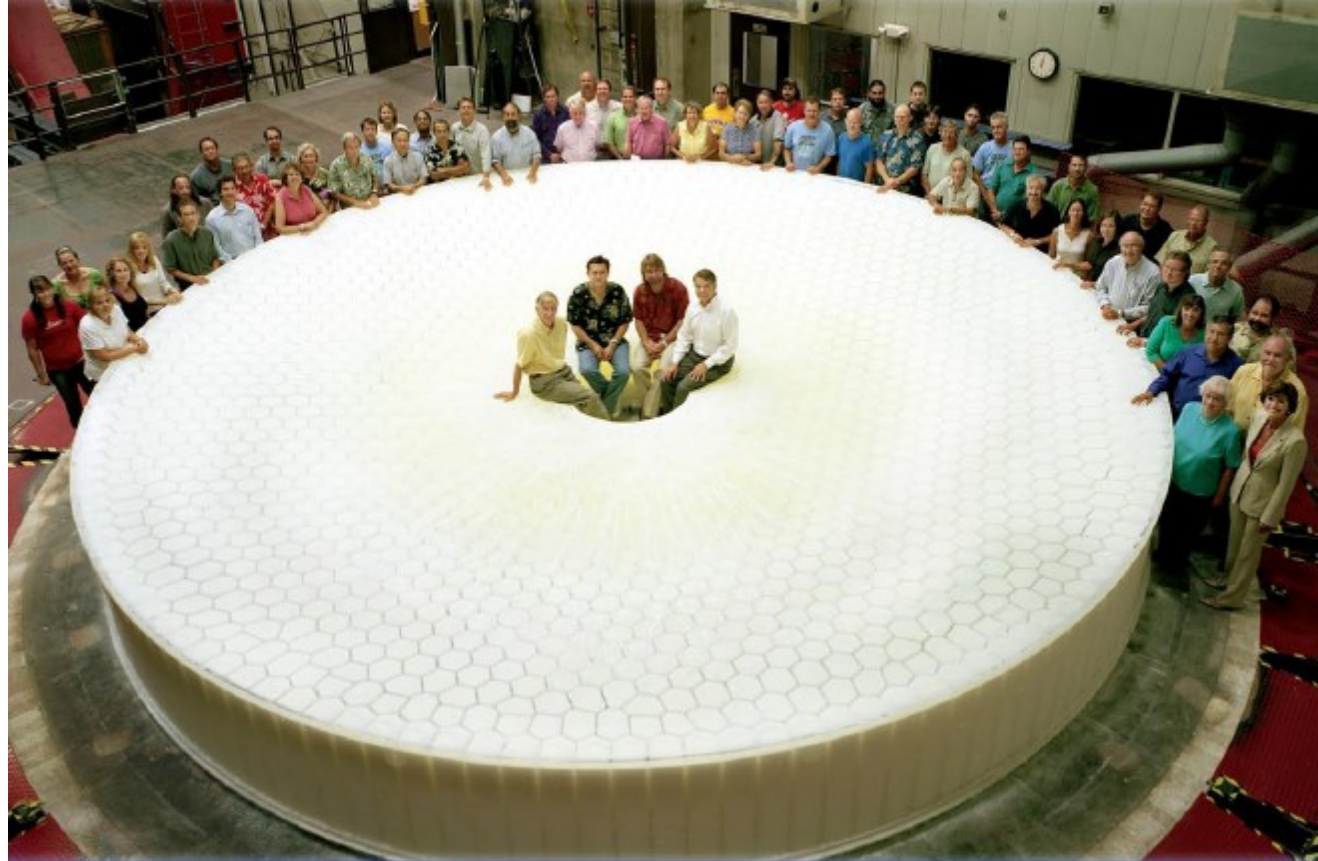
DERCAP
Sydney, Australia, 2009

Overview of Presentation



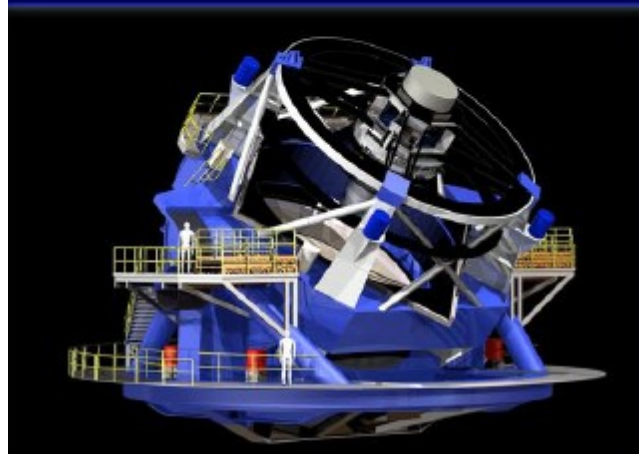
- **LSST - a large-scale Southern hemisphere optical survey**
- **LSST and other surveys**
- **The astronomical landscape in 2016**
 - **Explosion of network bandwidth**
 - **Cloud computing**
 - **Human computing and citizen scientists**
- **Challenges ahead**
 - **The perils of open data**
 - **Staying ahead of the evolving computing landscape**
 - **Nature of astronomical collaborations**

LSST - Large Synoptic Survey Telescope



J. Anthony Tyson, (530) 400-0406, University of California, Davis, tyson@physics.ucdavis.edu

• Željko Ivezić, University of Washington • Michael A. Strauss, Princeton University • Donald W. Sweeney, LSST Corporation • Sidney C. Wolff, LSST Corporation •



Large Synoptic Survey Telescope: Wide+Deep+Fast

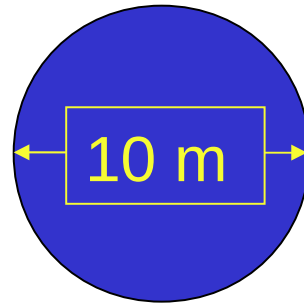


Primary mirror
diameter

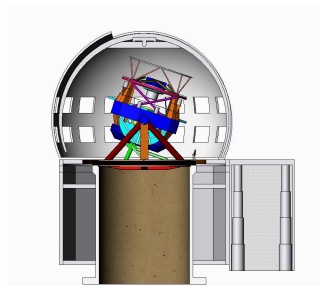
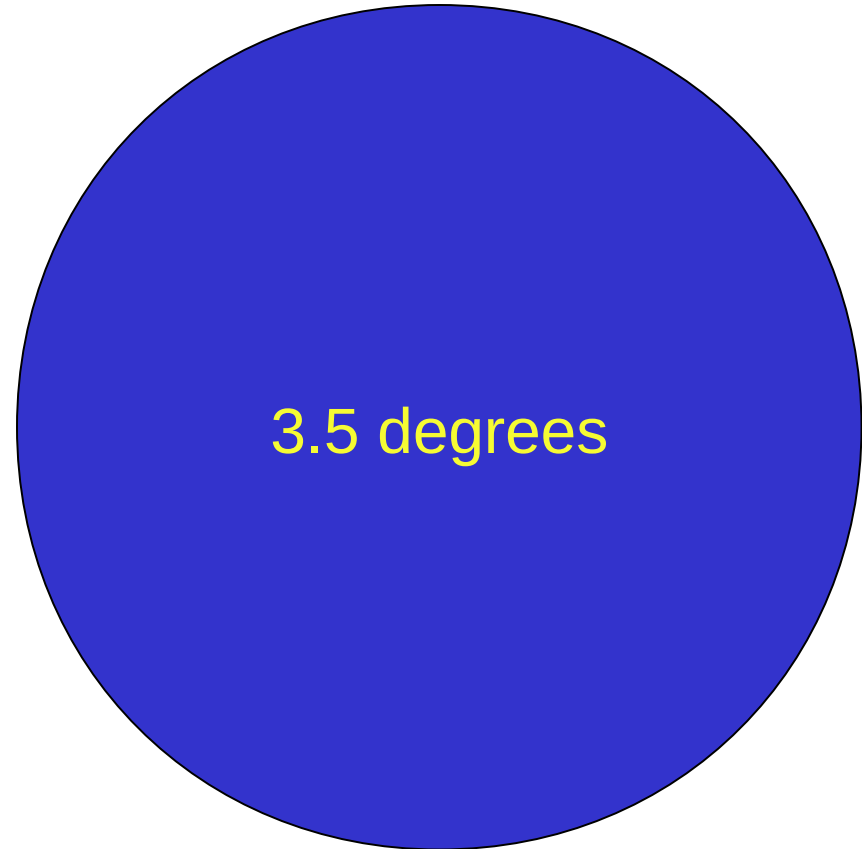
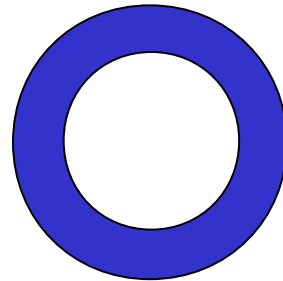
Field of view



Keck
Telescope

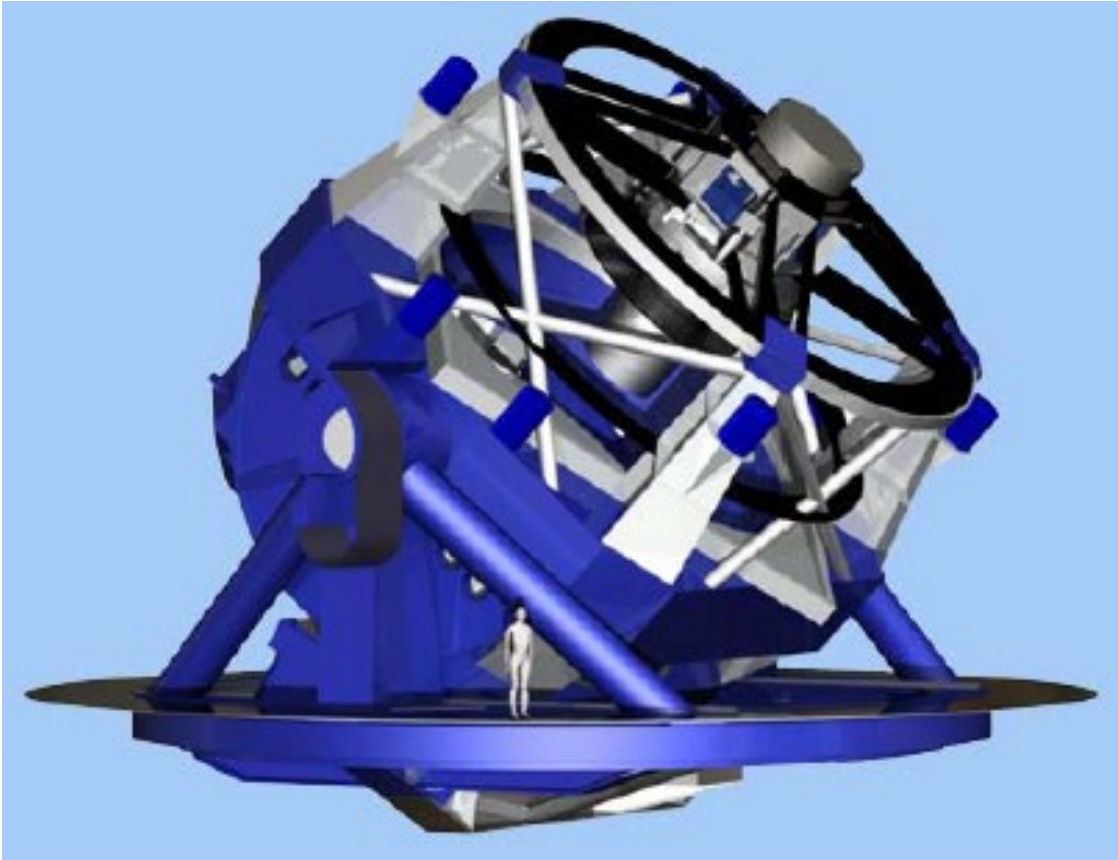


0.2 degrees ●



LSST

LSST - Essential Statistics



- Aperture diameter: 8.4m
- Effective aperture: 6.7m
- FOV: 3.5 deg
- Filters: u, g, r, i, z, y
- 3.2 gigapixels
- 2 sec, 5 electron noise readout
- Observing mode: pairs of 15 sec exposures, separated by 5 sec slew
- Single exposure depth: ~24.5
- Repetitively scan 20000 sq deg
- Site: Cerro Pachon, Chile
- Data flows at 0.5 GB/sec – all night
- 18 TB / night
- First light: ~2016

LSST Institutional Members



Brookhaven National Laboratory (BNL)

California Institute of Technology

Carnegie Mellon University

Chile

Columbia University

Cornell University

Drexel University

Google, Inc.

Harvard-Smithsonian Center for Astrophysics

Institut de Physique Nucléaire et de Physique des Particules (IN2P3)

Johns Hopkins University

Kavli Institute for Particle Astrophysics and Cosmology (KIPAC) - Stanford University

Las Cumbres Observatory Global Telescope Network, Inc.

Lawrence Livermore National Laboratory (LLNL)

Los Alamos National Laboratory (LANL)

National Optical Astronomy Observatory*

Princeton University

Purdue University

Research Corporation for Science Advancement*

Rutgers University

SLAC National Accelerator Laboratory

Space Telescope Science Institute

The Pennsylvania State University

The University of Arizona*

University of California at Davis

University of California at Irvine

University of Illinois at Urbana-Champaign

University of Pennsylvania

University of Pittsburgh

University of Washington*

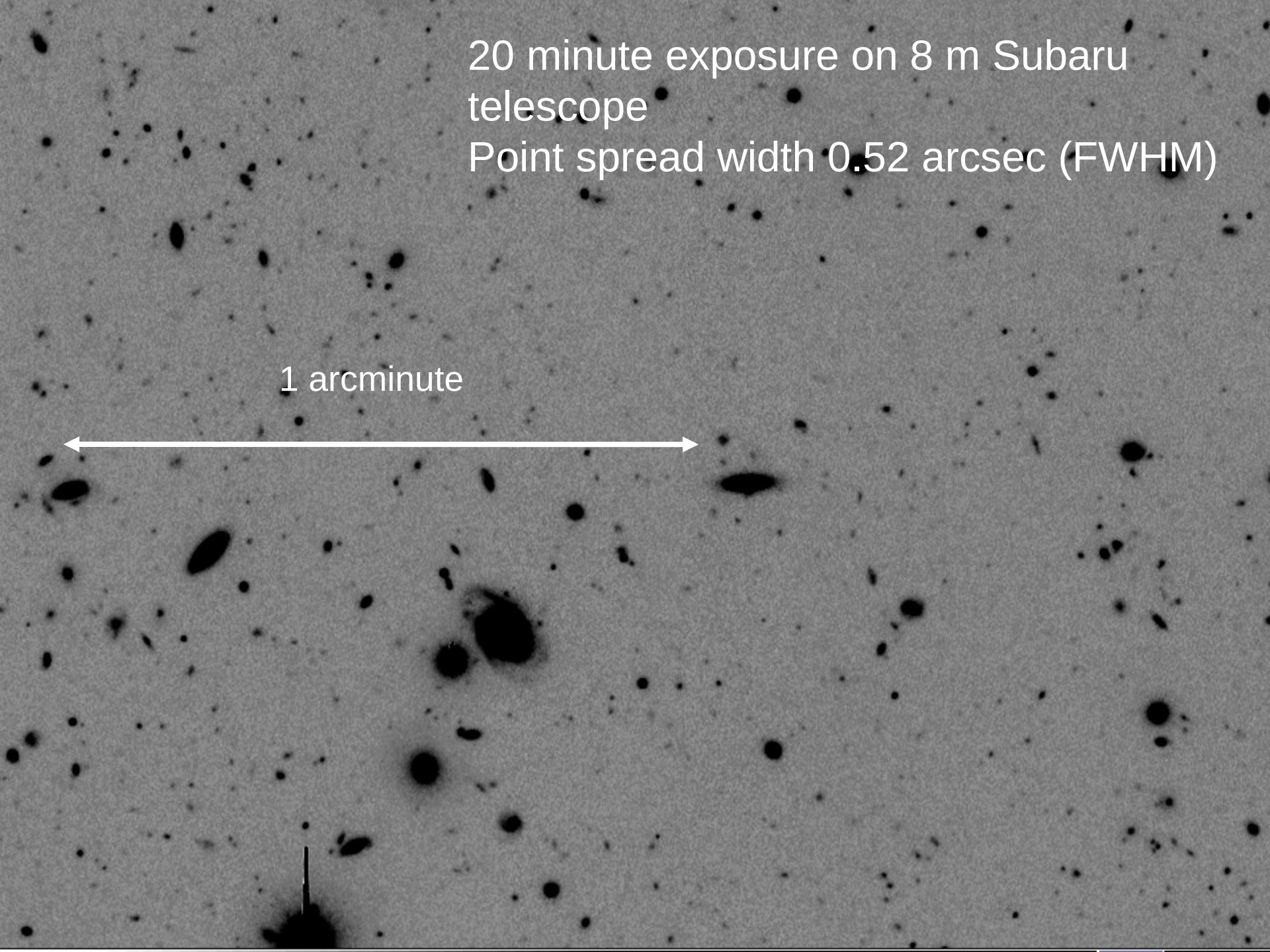
Vanderbilt University

* Founding Member

A Huge Geographical Contrast from High Energy Physics - WHY??

20 minute exposure on 8 m Subaru
telescope
Point spread width 0.52 arcsec (FWHM)

1 arcminute



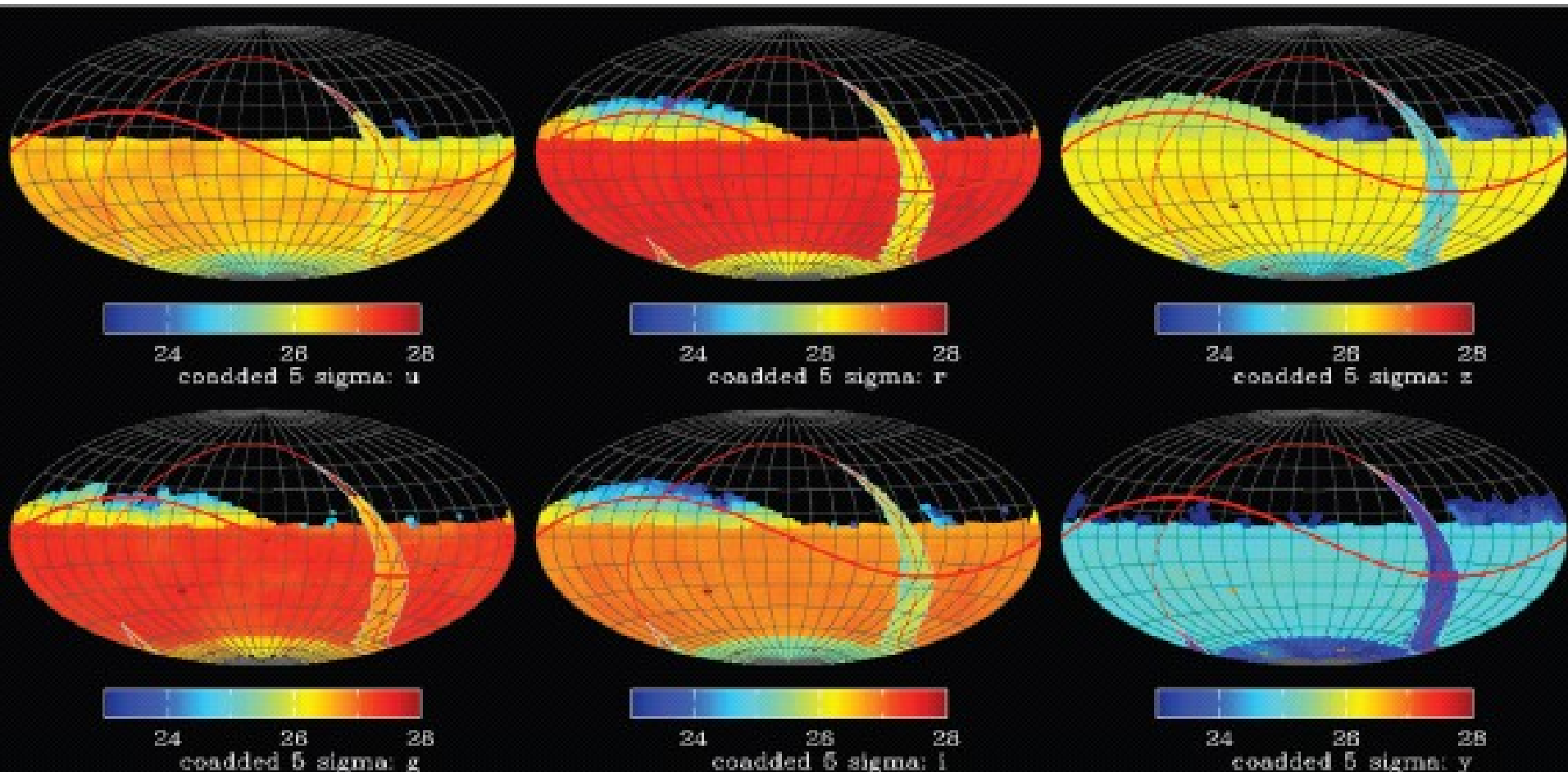
One Survey – Many Science Programs



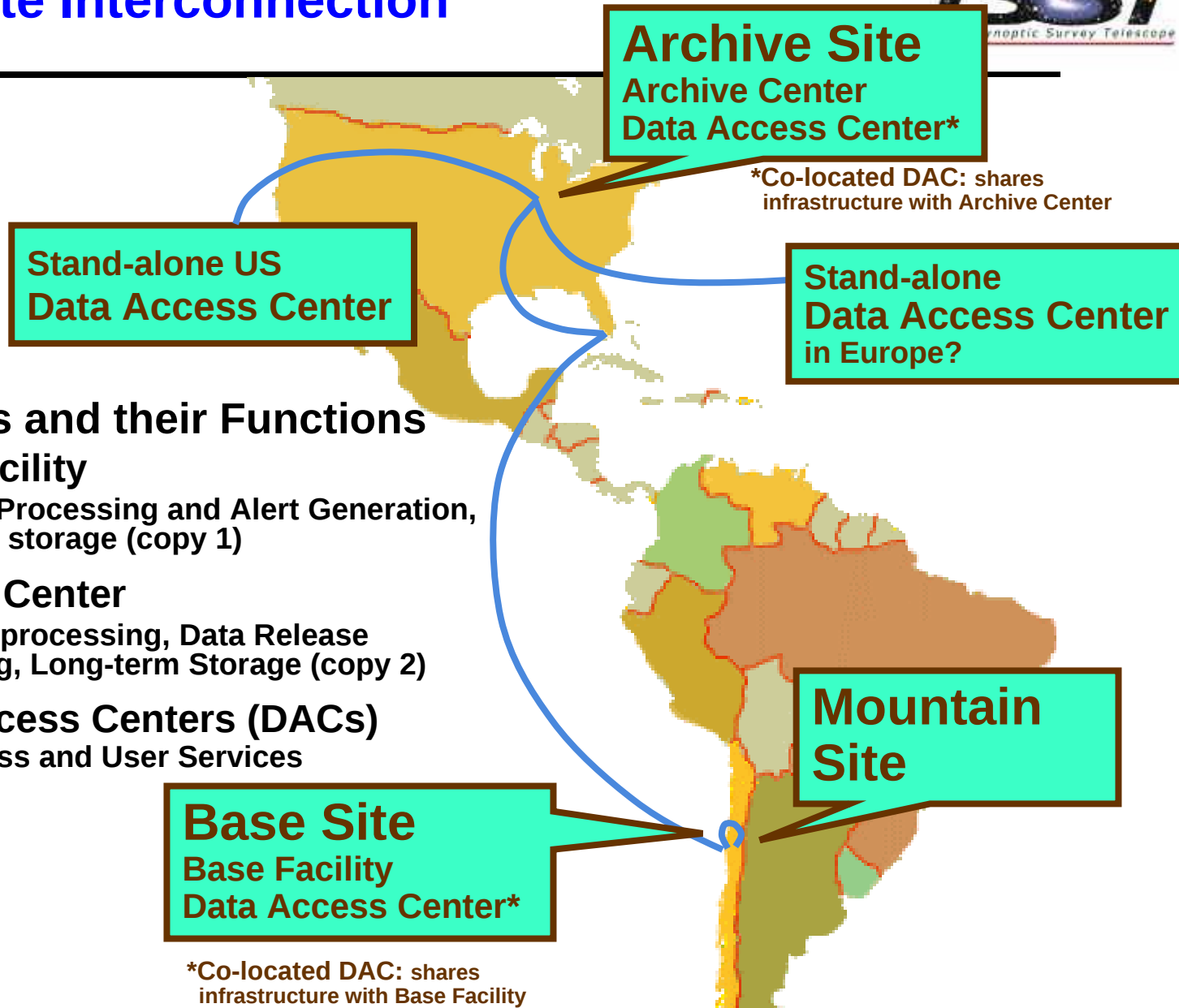
- **The LSST Observatory will produce a data stream which the Data Management System turns into data products.**
 - **0.5 GB/sec all night, every night for 10 years**
 - **104 PB of images at survey end**
 - **2.5 PB science database at survey end**
- **Many science programs are supported by the same data products**
 - **Weak lensing**
 - **Supernovae & transient astrophysics**
 - **Milky Way structure**
 - **Solar System inventory**
 - **Many more in individual science collaborations**

Simulated Results of 10 yr Survey

5.3M Exposures



LSST Site Interconnection



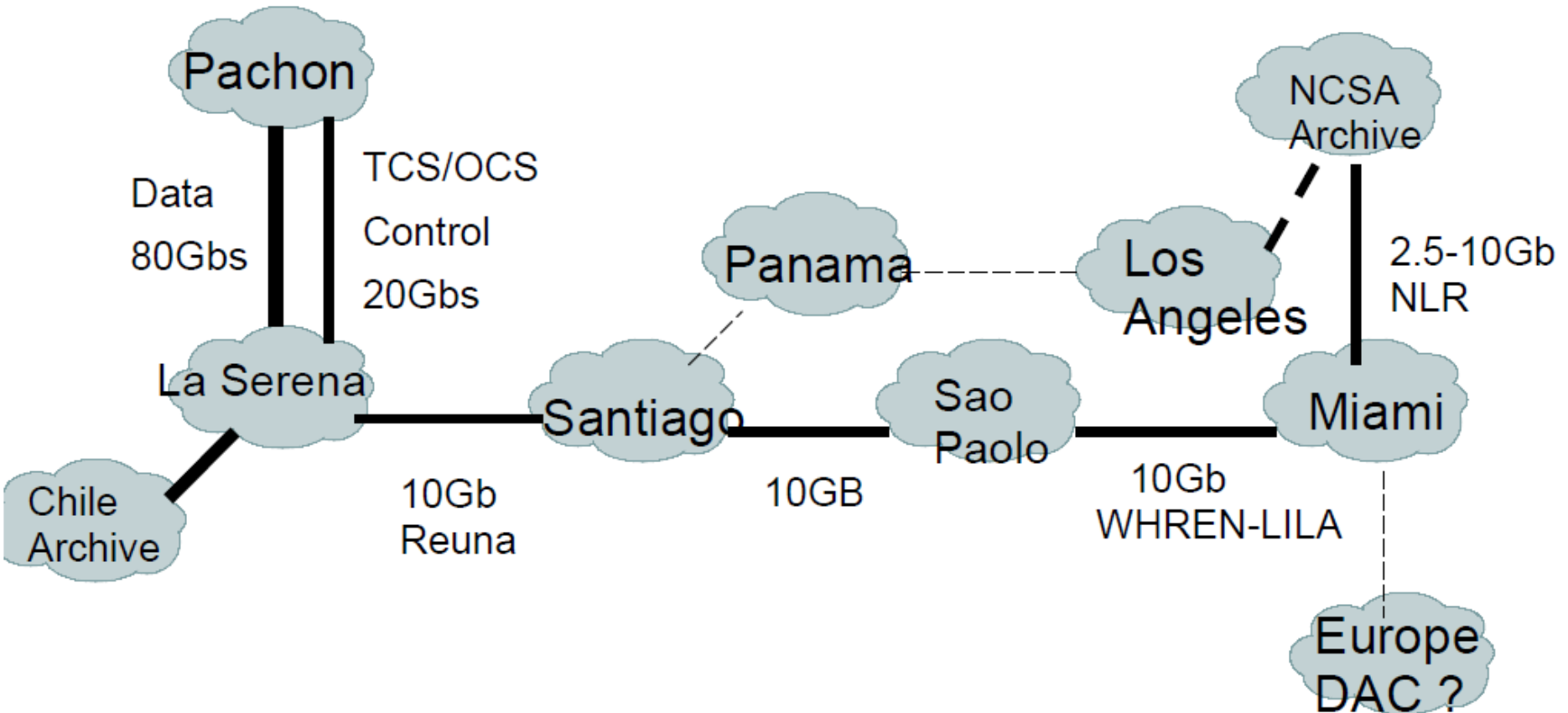
Site Roles and their Functions

- **Base Facility**
Real-time Processing and Alert Generation, Long-term storage (copy 1)
- **Archive Center**
Nightly Reprocessing, Data Release Processing, Long-term Storage (copy 2)
- **Data Access Centers (DACs)**
Data Access and User Services

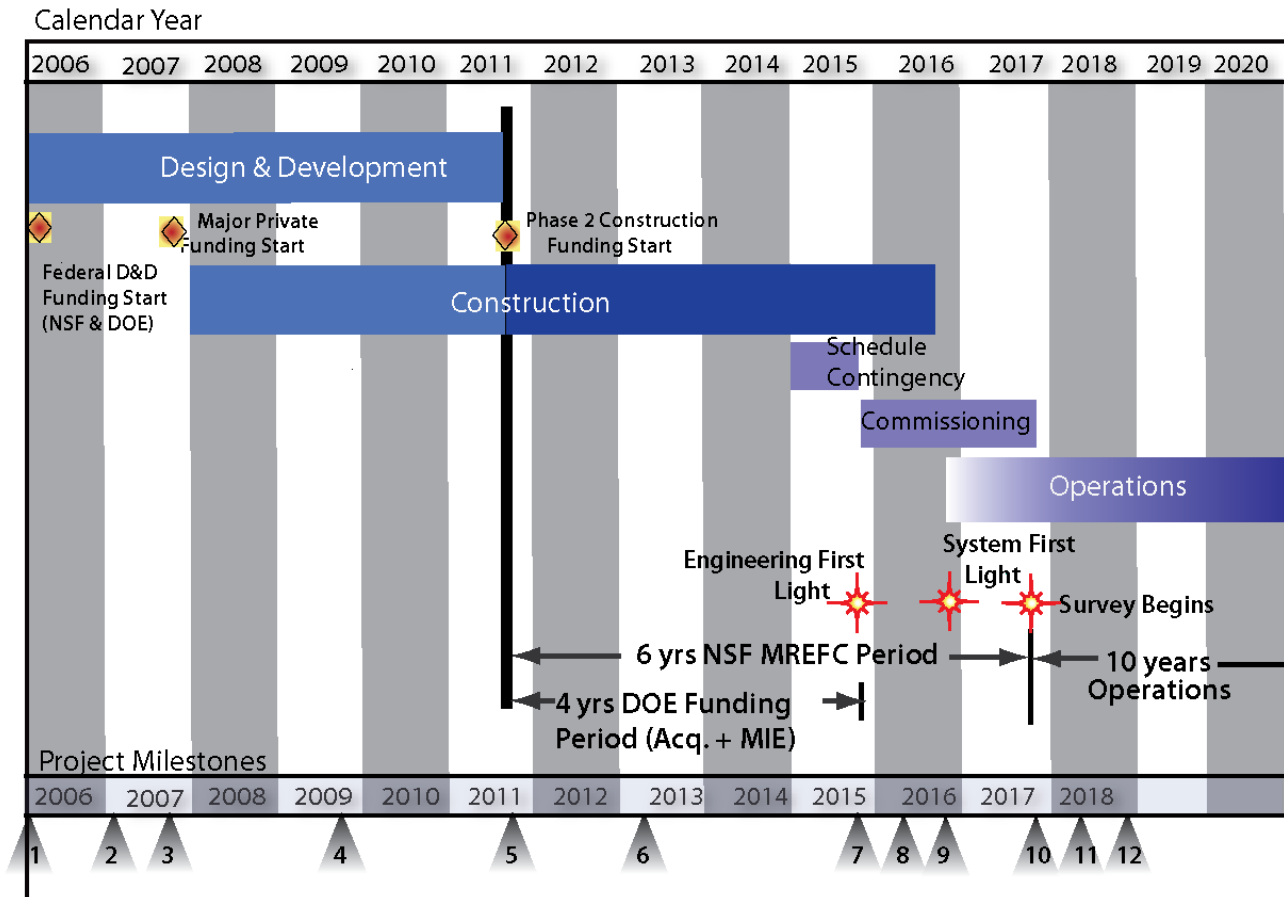
*Co-located DAC: shares infrastructure with Base Facility

October 15-16, 2008 Tucson, AZ

LSST High Speed Networks



Proposed LSST timeline



Major Project Milestones:

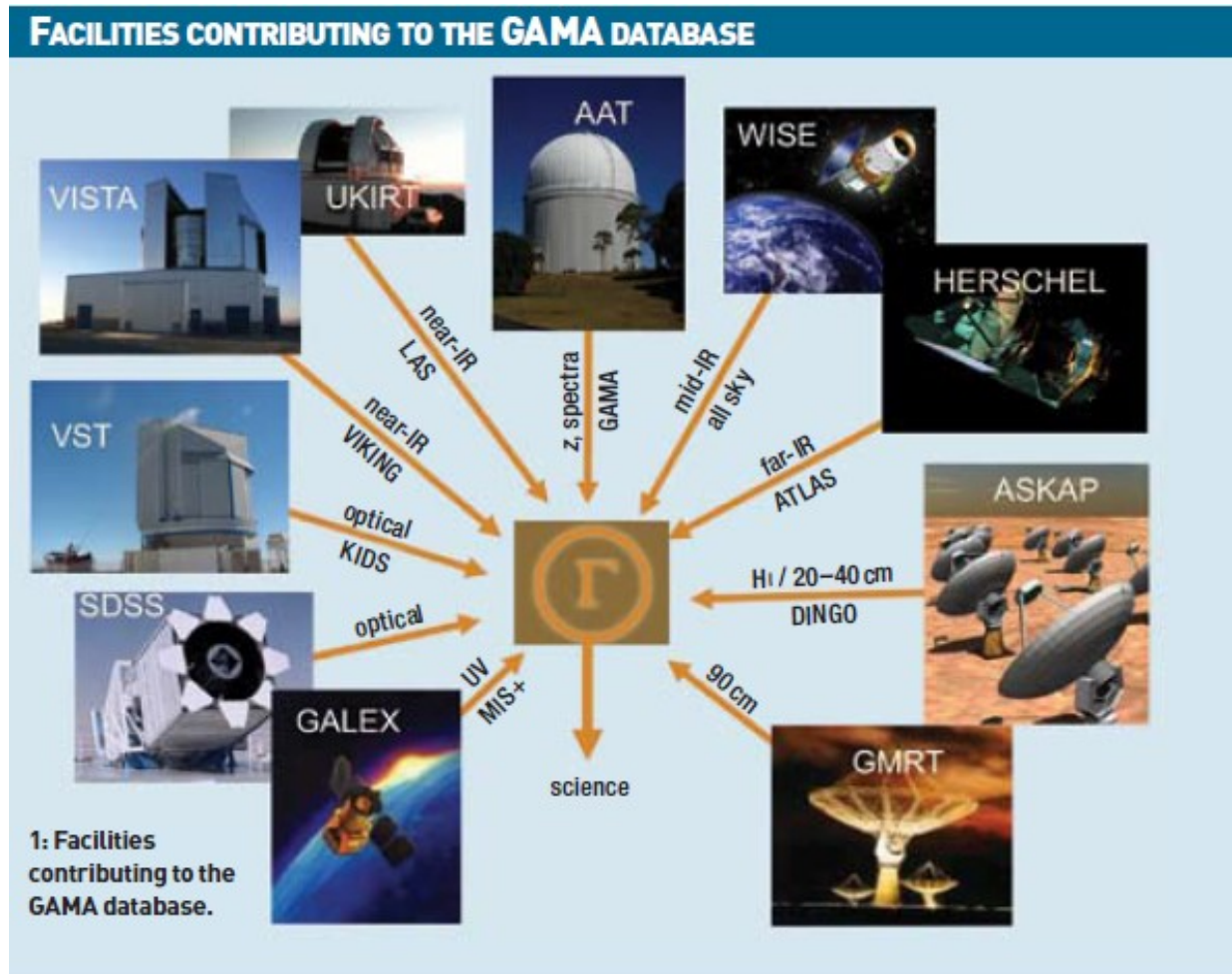
1. Sep 2005 NSF D&D Funding Start
2. Feb 2007 NSF MREFC Proposal Submitted
3. Sep 2007 NSF Conceptual Design Review
4. Oct 2009 NSF Preliminary Design Review
5. Oct 2011 NSF Critical Design Review; Construction Funding Start
DOE Critical Decision 2a Review; DOE Acquisition Funding Start
6. Apr 2013 First Camera Raft Complete
7. Aug 2017 First Engineering Light with Eng Camera
System Integration and Test Begins
8. Mar 2016 Archive Center Complete
9. Sep 2016 System First Light with 3.2 GP Camera
System Science Validation Begins
10. Oct 2017 Full Science Operations Begins
11. Apr 2018 First LSST Data Release
12. Oct 2018 Second LSST Data Release

LSST and Other Surveys - What might we do?



- **Real time spectroscopic followup**
 - Most LSST-detected transients are really faint – 24 or so, need big telescopes for spectroscopy
 - Southern Hemisphere – Chile, S. Africa?
- **Real time transient science combining optical and radio**
 - ASKAP, SKA
 - **The Australian connection!**
- **Pixel-level combination with surveys in other wavelengths**
 - Optimal deblending
 - Detect rare objects, eg lensed supernovae
- **Ad-hoc combination of information with other surveys through the VO**

Projects like GAMA will become more common



GAMA = Galaxy and Mass Assembly – Simon Driver, PI

There is a cloud on the horizon...



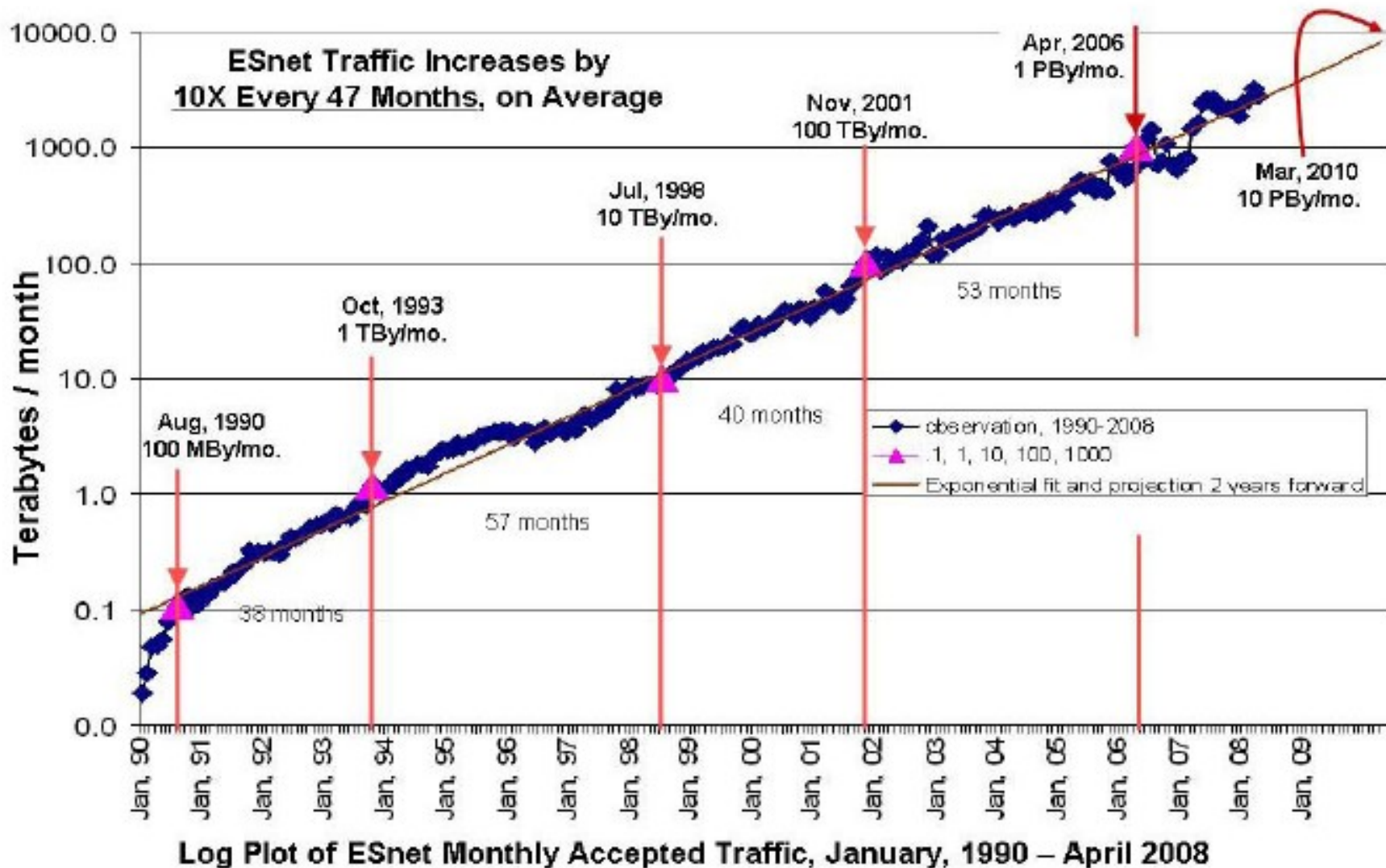
-
- **The cloud on the horizon is the inexorably increasing ratio of data to humans**
 - The science budget is roughly constant
 - The number of funded humans on project budgets is also roughly constant
 - The data flow from experiments is exploding, driven by sensor and computing technology
 - LSST is one example of this, among many
 - **What is the danger?**
 - The entire scientific enterprise is built on the quality of experimental data
 - If it contains unrecognized quality problems, we may fail
 - Experience shows that automated techniques are only partly successful at identifying subtle problems in the data

Three Transformative Trends Will Shape the Landscape in 2016



- **High Speed Networks**
 - The foundation
- **Cloud Computing**
 - Enabled by high speed networks
- **Human Computation**
 - Enabled by both high speed networks and cloud computing
 - Itself a form of cloud computing?
- **We will need to make use of all three to overcome the challenges ahead to data intensive astronomy**

There is a Moore's law for networks too...



CMS Global Data Grid

CMS Experiment



Online System

0.1 - 1.5 GB/s

- 2500 physicists, 40 countries
- 10s of Petabytes/yr by 2008
- 1000 Petabytes in < 10 yrs?

CERN T0

10-40 Gb/s

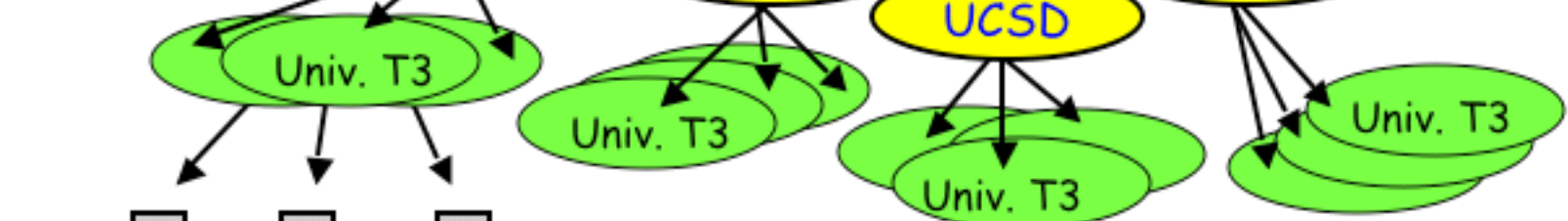
Tier 0



Tier 1



Tier 2



Tier 3



Tier 4

Implications of Network Bandwidth Explosion



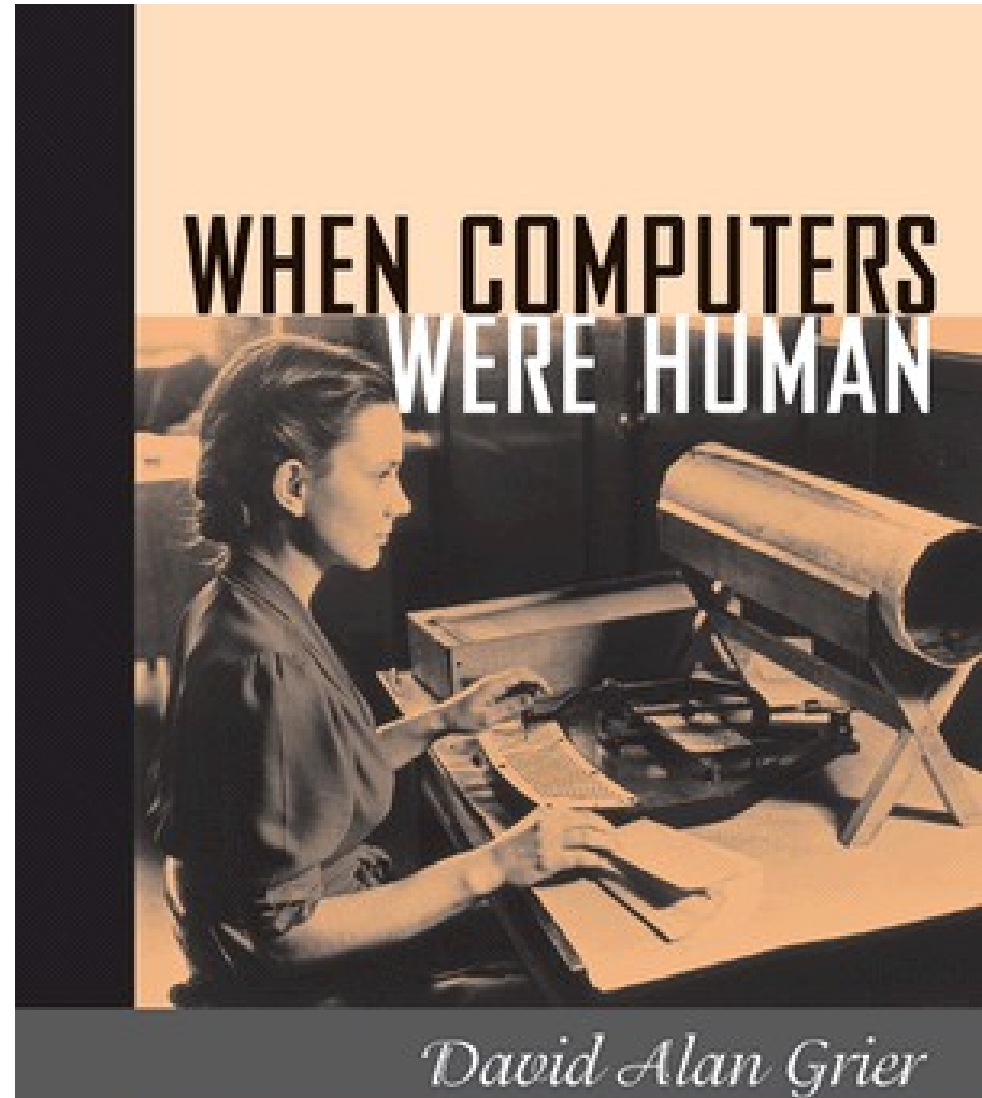
- **Large data gathering experiments at remote locations**
 - Are possible
 - Can stream their data globally in real time
- **Large data repositories can be globally “present” at user sites**
 - The pendulum is swinging back from “must move the computing to the data”
 - Neither the computing or the data is really localized!
- **A concrete example: Reprocess 10% of the LSST survey pixels at a user site in 2016**
 - About 10 PB of raw pixel data
 - For a major user site, we could expect 50 Gb/sec bandwidth
 - Takes 20 days – not negligible, but not silly
 - Processing time probably not network limited

- **Extensively covered by other speakers**
- **Essential characteristics for Astronomy**
 - **Economic model radically different**
 - Computing as commodity
 - Fund from operations rather than construction
 - **Overall capacity driven by demand from enormous user base**
 - In this context, science demands are not so large
 - The capacity will just “be there”
 - **We can hope for a convergence of currently diverse cloud application interfaces**
 - Needed to justify large investment in science application software

Human Computing



- **The first computers were human**
 - Astronomy the earliest application
 - Manhattan Project
- **From Manhattan Project days on, human computers were combined with mechanical/electronic aids**



Changing Notion of Human Computation



- **Original human computers were a substitute for the electronic computers that didn't yet exist**
- **Low computing rate and relatively high error rates forced**
 - **Clever numerical algorithms**
 - **Error detection and correction**
 - **Use of parallelism**
 - Circa 1915 Lewis Richardson fantasized about a network of 64000 human computers in a stadium sized space connected in a spherical topology for modeling weather
- **But none of the essentially human characteristics were used**
 - **Pattern recognition**
 - **Learning**

- **Motivation of Human Computers**
 - Economic
 - Play
 - Participation in a scientific enterprise - “citizen science”
- **Strengths of Human Computers**
 - Still unmatched at recognizing subtle visual patterns
 - They learn!
 - There are potentially really a lot of them
- **Limitations of Human Computers**
 - Biases, unreliability
 - They do get bored...
- **Citizen science is the most natural match to our astronomical needs**

An Example Citizen Science App



GALAXY ZOO 2

Home The Story So Far The Science How To Take Part Classify Galaxies Forum Zoo Media Blog FAQ Contact Us

Register Log In

Aiming for 60 million

The Zoonometer™ has quickly become a popular feature of Galaxy Zoo, so we've decided to make it a permanent one. Exactly how many classifications will be needed to finish the job for Galaxy Zoo 2 will depend on exactly what the results look like when we get into them, but we've decided to think big and aim for 60 million. Get clicking!

NUMBER OF CLASSIFICATIONS

46,453,612

ZOONOMETER

60,000,000
45,000,000
30,000,000
15,000,000
0

8000 clicks per hour, averaged over 8 months

How Might we Integrate Citizen Science Into A Large Survey?



- Finding anomalies and quality problems in the data is our biggest need
- Images
- Patterns in databases
 - Classification is a close second
 - To make this work, our human computers will need some really advanced visualization tools
- Apply high speed computing to allow humans to have configurable “data goggles”
- Must keep it visually interesting, and enjoyable
 - LSST has an active EPO program that has citizen science as a focus
- Working on early prototypes (Lightcurve Zoo)
- Ideas, and especially collaborations, are welcome!

The Perils of Open Data



- **LSST has been committed from the start to completely open data**
- **Existing astronomy projects, like high energy physics projects, are not open**
- **It is difficult to persuade people to pay for what they think they will get for free**
- **It is difficult to persuade people to give up the perceived benefits of keeping their own data proprietary**
- **Given that situation, US funding agencies are reluctant to extend open data beyond the US border**
- **Open data has apparently become a real obstacle to building and operating the LSST**
 - **Is this the reason the geographical map looks so different from High Energy Physics?**

Astronomical Collaborations in 2016



- **Many reasons to collaborate!**
 - **Data usefulness grows through joining with other surveys – will become ever more true**
 - **Many common problems**
 - Software design
 - Data archiving
 - Data curation
 - Etc
- **The computing infrastructure will make the mechanics of collaboration ever easier even as data volumes grow**
- **The challenges are mainly sociological**
 - We need to understand them
 - Our funding agencies need to understand them!