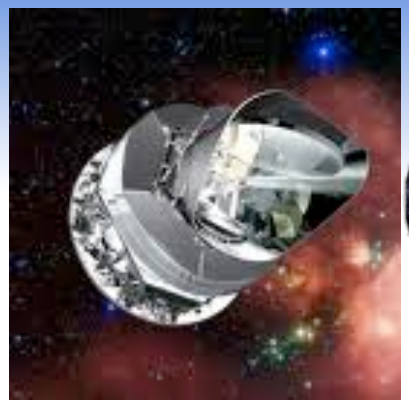
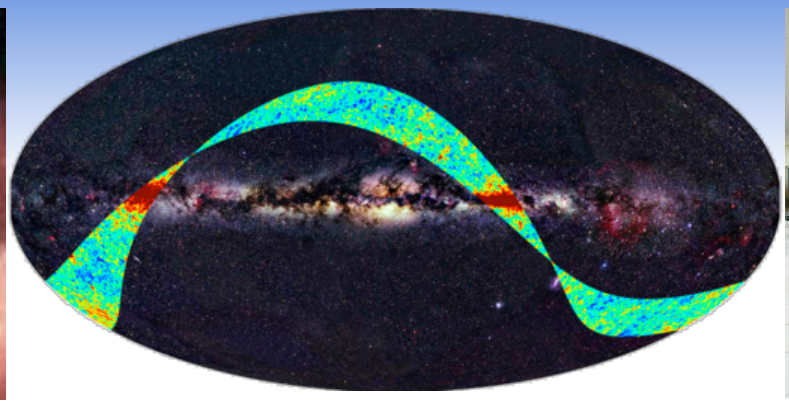




Cosmic Computing

Supporting the Science of the Planck Space Based Telescope

Shane Canon

Lawrence Berkeley National Laboratory

Plenary Address

LISA 2009

November 2009

BERKELEY LAB

NERSC

Office of Science

U.S. DEPARTMENT OF ENERGY

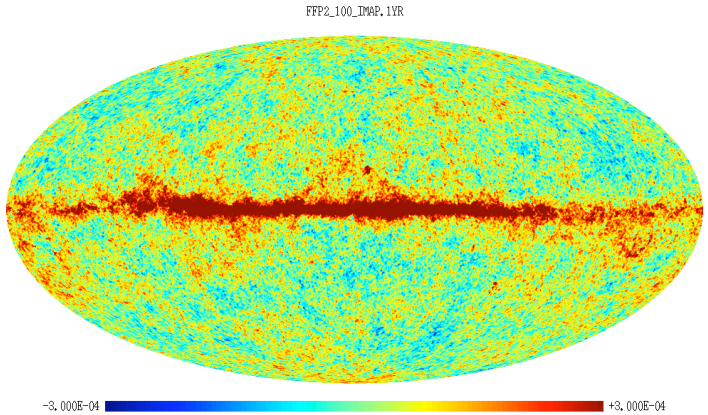
Disclaimer

I am not an astrophysicist, cosmologist, rocket scientist or even a computer scientist. I am not a member of the Planck Collaboration.

So I'm not an expert on much of what I'm about to speak on.

Outline

- The Science
- The Planck Mission
- The Data Pipeline
- NERSC
- Big Data at NERSC
- Big Data Challenges

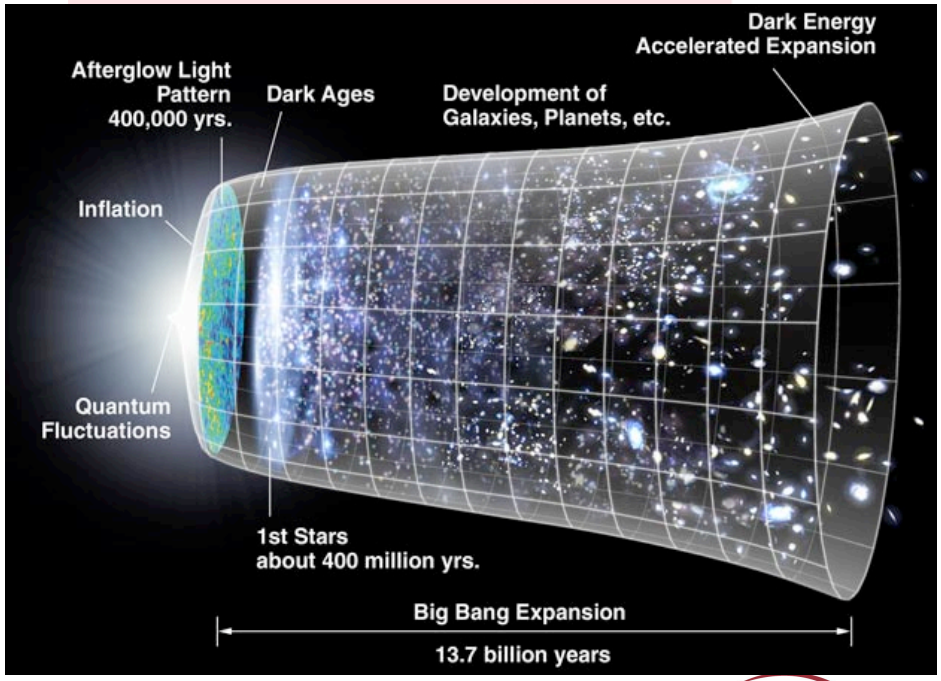


What Is the Cosmic Microwave Background?

About 400,000 years after the Big Bang, the expanding Universe cools through the ionization temperature of hydrogen: $p^+ + e^- \Rightarrow H$.

Without free electrons to scatter off, CMB photons free-stream to us today.

- **COSMIC - filling all of space.**
- **MICROWAVE - redshifted by the expansion of the Universe from 3000K to 3K.**
- **BACKGROUND - primordial photons coming from "behind" all astrophysical sources.**



Credit: NASA

Why Are We Interested In The CMB ?

- It is the earliest possible photon image of the Universe.
- Its existence supports a Big Bang over a Steady State cosmology.
- Tiny fluctuations in the CMB temperature and polarization encode details of
  - cosmology
    - geometry
    - topology
    - composition
    - history
  - ultra-high energy physics
    - fundamental forces
    - quantum field theory beyond the standard model
    - inflation, the dark sector, etc.

Initial Discovery

Accidentally discovered by Penzias and Wilson in 1965 while working on improving transmitters for Bell Labs.

Awarded the Nobel Prize in 1978.

Other Efforts

Ground Based

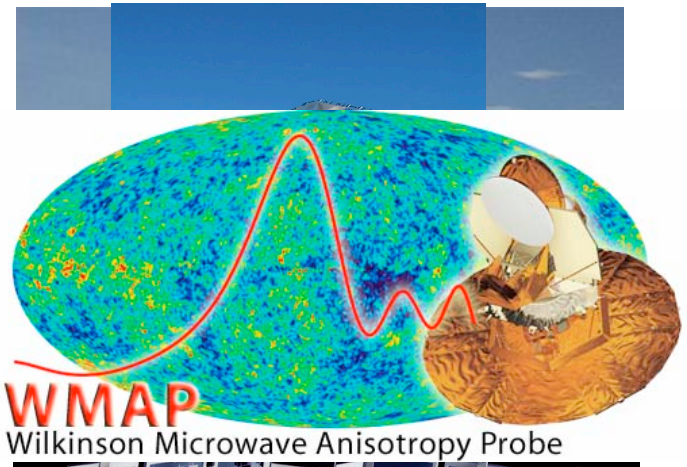
- AMiBA, CBI

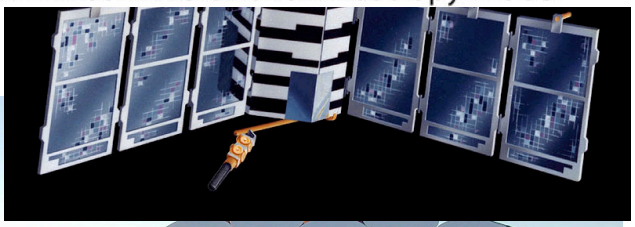
Balloon Based Experiments

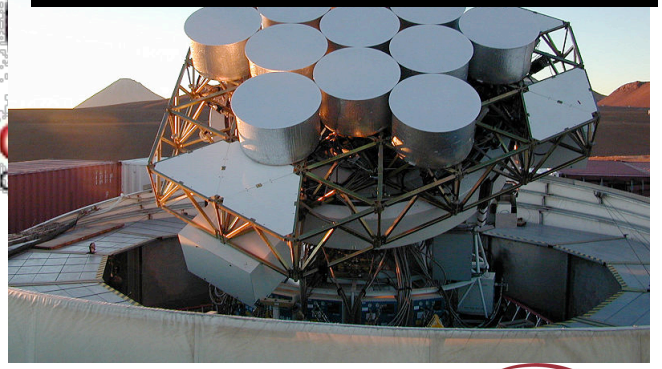
- Boomerang, MAXIMA

Space Based

- COBE, WMAP







Understanding our Universe

Research on the CMB lead to a second Nobel Prize in 2006.

Photo Courtesy of Nobel Foundation 2006

The Royal Swedish Academy of Sciences has decided to award the Nobel Prize in Physics for 2006 jointly to

- **John C. Mather**
  NASA Goddard Space Flight Center, Greenbelt, MD, USA,

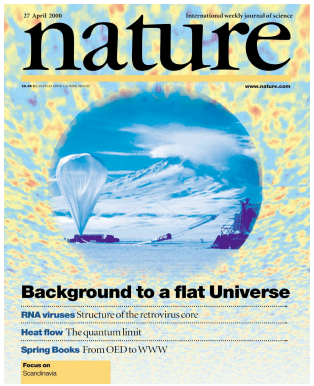
  and
- **George F. Smoot**
  University of California, Berkeley, CA, USA

  "for their discovery of the blackbody form and anisotropy of the cosmic microwave background radiation".

The Concordance Cosmology

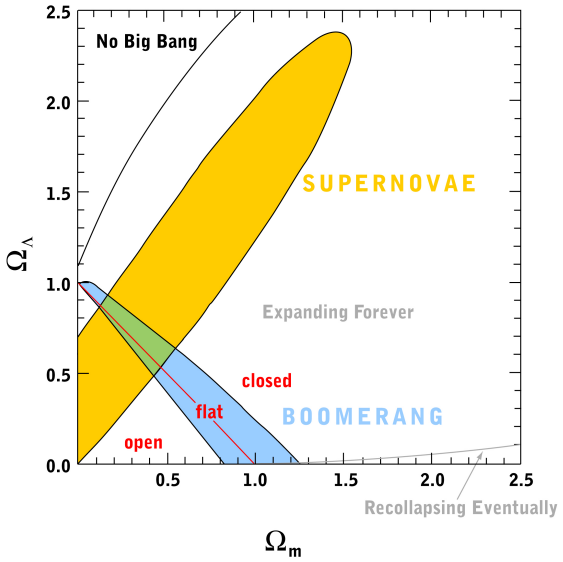
Supernova Cosmology Project (1998):

Cosmic Dynamics (Ω_{Λ} - Ω_m)



BOOMERanG & MAXIMA (2000):

Cosmic Geometry (Ω_{Λ} + Ω_m)



70% Dark Energy + 25% Dark Matter + 5% Baryons

95% Ignorance

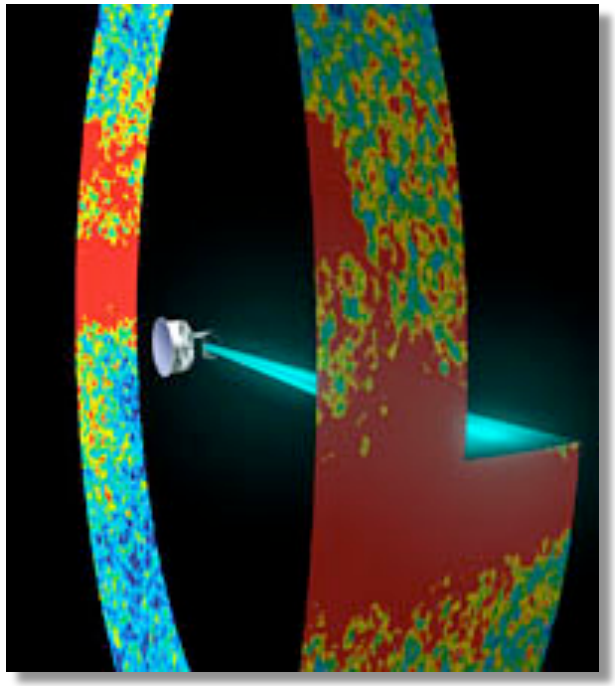
What (and why) is the Dark Universe ?

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Outline

- The Science
- The Planck Mission
- The Data Pipeline
- NERSC
- Big Data at NERSC
- Big Data Challenges



NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Mission Goal

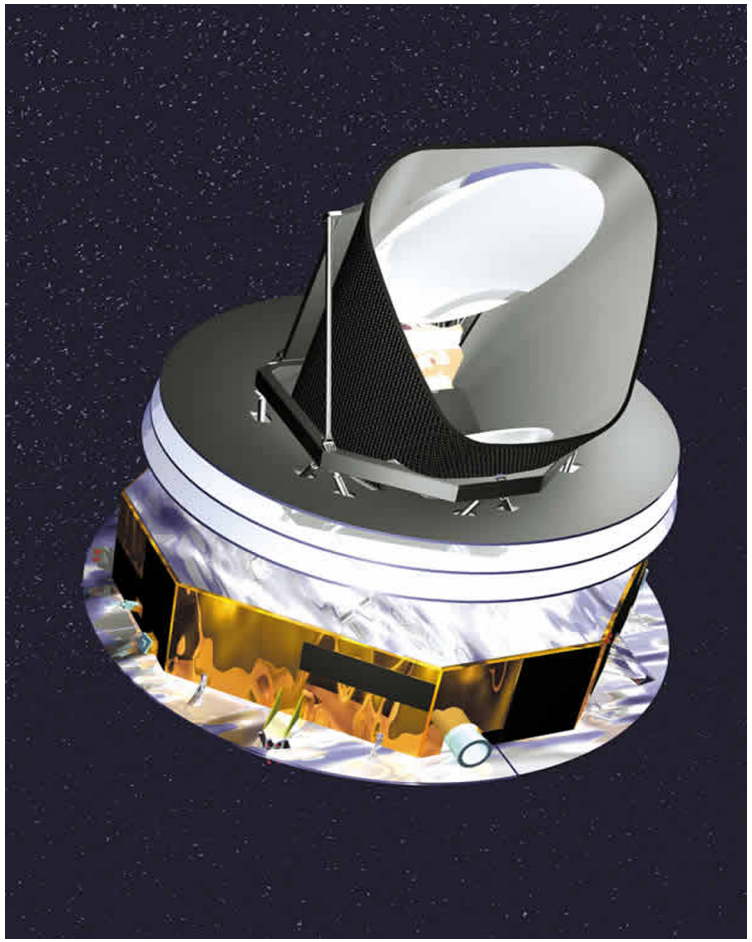
Planck will measure the fluctuations of the CMB with an accuracy set by fundamental astrophysical limits.

Planck will provide the sharpest picture ever of the young Universe — when it was only 380 000 years old — and zeroing-in on theories that describe its birth and evolution.

Planck's Main Objectives

- To determine the large-scale properties of the Universe with high precision including Dark Matter
- To test theories of inflation
- To search for primordial gravitational waves.
- To search for 'defects' in space,
- To study the origin of the structures we see in the Universe today.
- To study our and other galaxies in the microwave.

The Planck Satellite

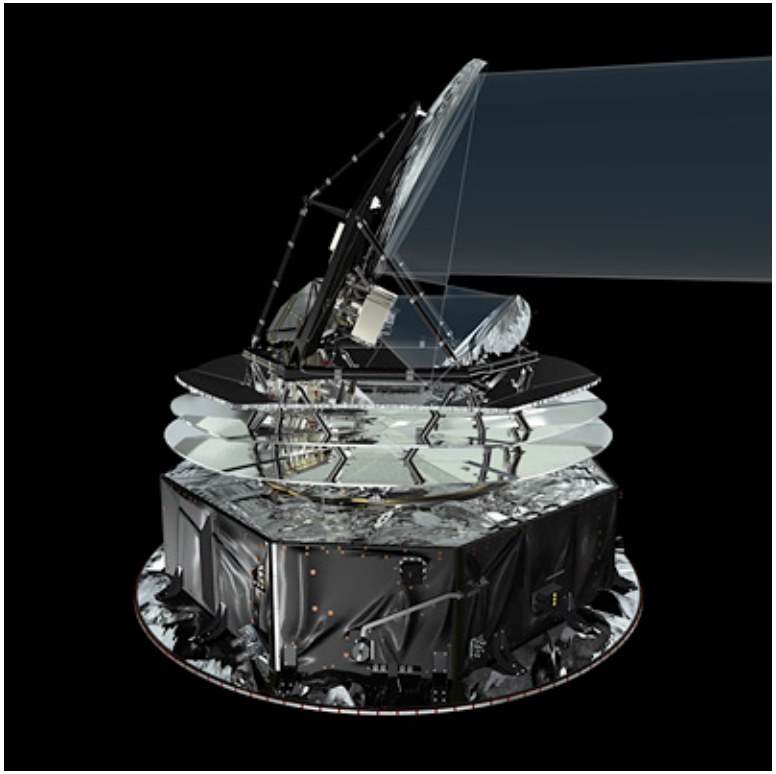


Credit: ESA

- A joint ESA/NASA mission launched in 2009.
- A 2-year+ all-sky survey from L2.
- Survey at 9 microwave frequencies from 30 to 857 GHz.
- The biggest data set to date:
  - $O(10^{12})$ observations
  - $O(10^{8})$ sky pixels
  - $O(10^{4})$ spectral multipoles

The Planck Satellite

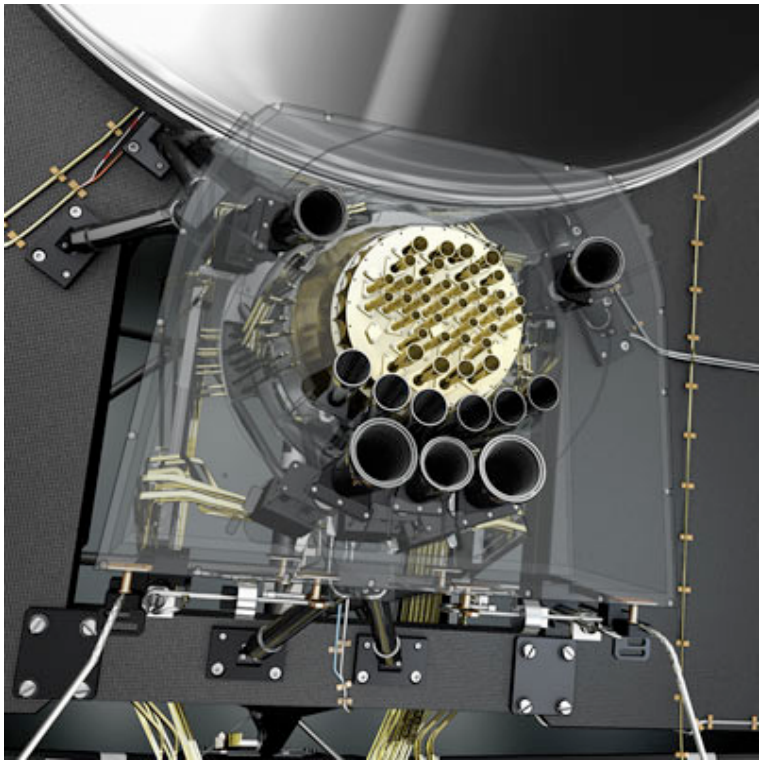
- The Satellite consist of a 1.9mx1.5m telescope which focuses radiation on two arrays of radio detectors (Low and High Frequency Instruments).
- The Satellite also includes shielding and advanced Cryogenic cooling.



Credit: ESA

The Planck Satellite

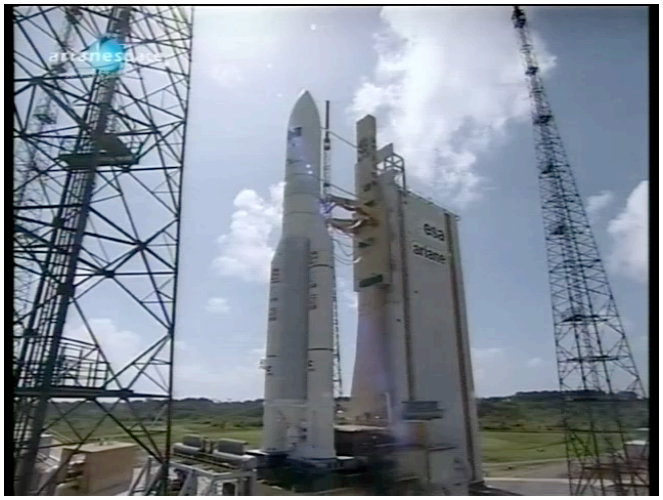
- Low-Frequency
  - 27 to 77 GHz
  - 30 GHz, 44 GHz, and 70 GHz
- High-Frequency
  - 84GHz to 1 THz
  - 100,143,217,353, 545,857 GHz

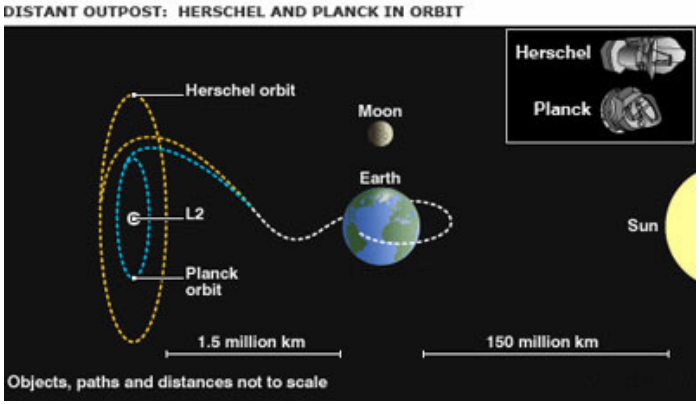


Credit: ESA

Launch

- Launched on May 14, 2009 by the ESA using an Arian 5 rocket.
- In orbit around the second Lagrangian Point (L2) about 1.5 million KM from Earth opposite the Sun.

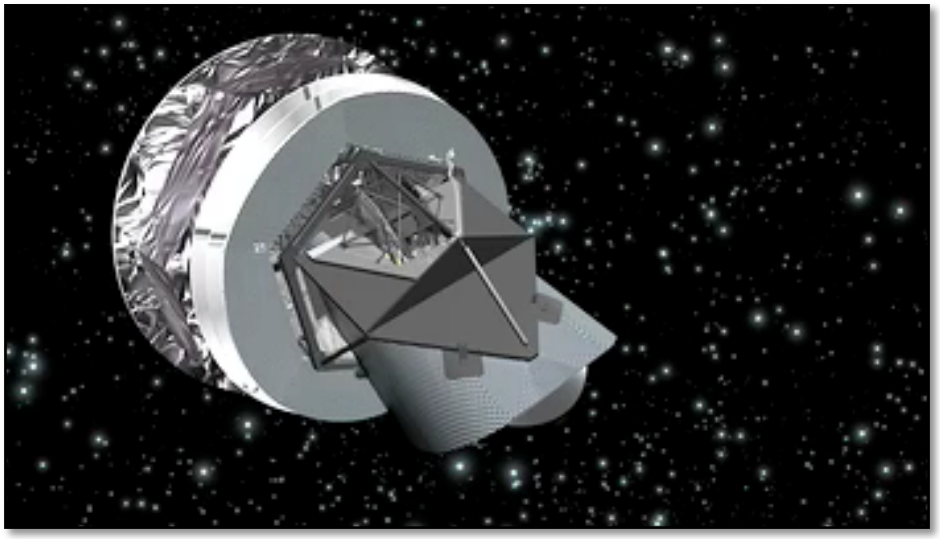




Credit: ESA

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Mapping the Universe

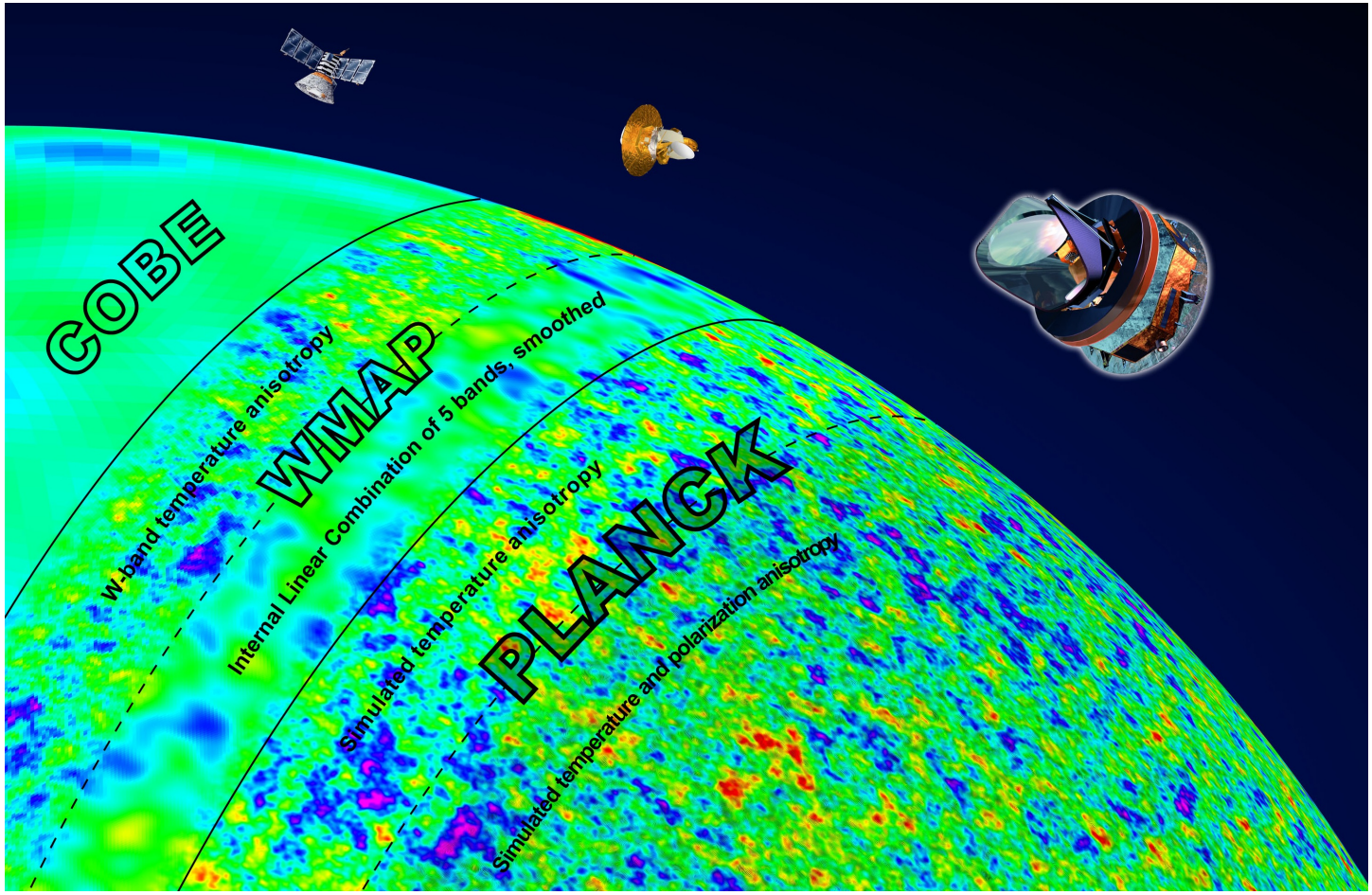


Credit: ESA

Office of Science
U.S. DEPARTMENT OF ENERGY

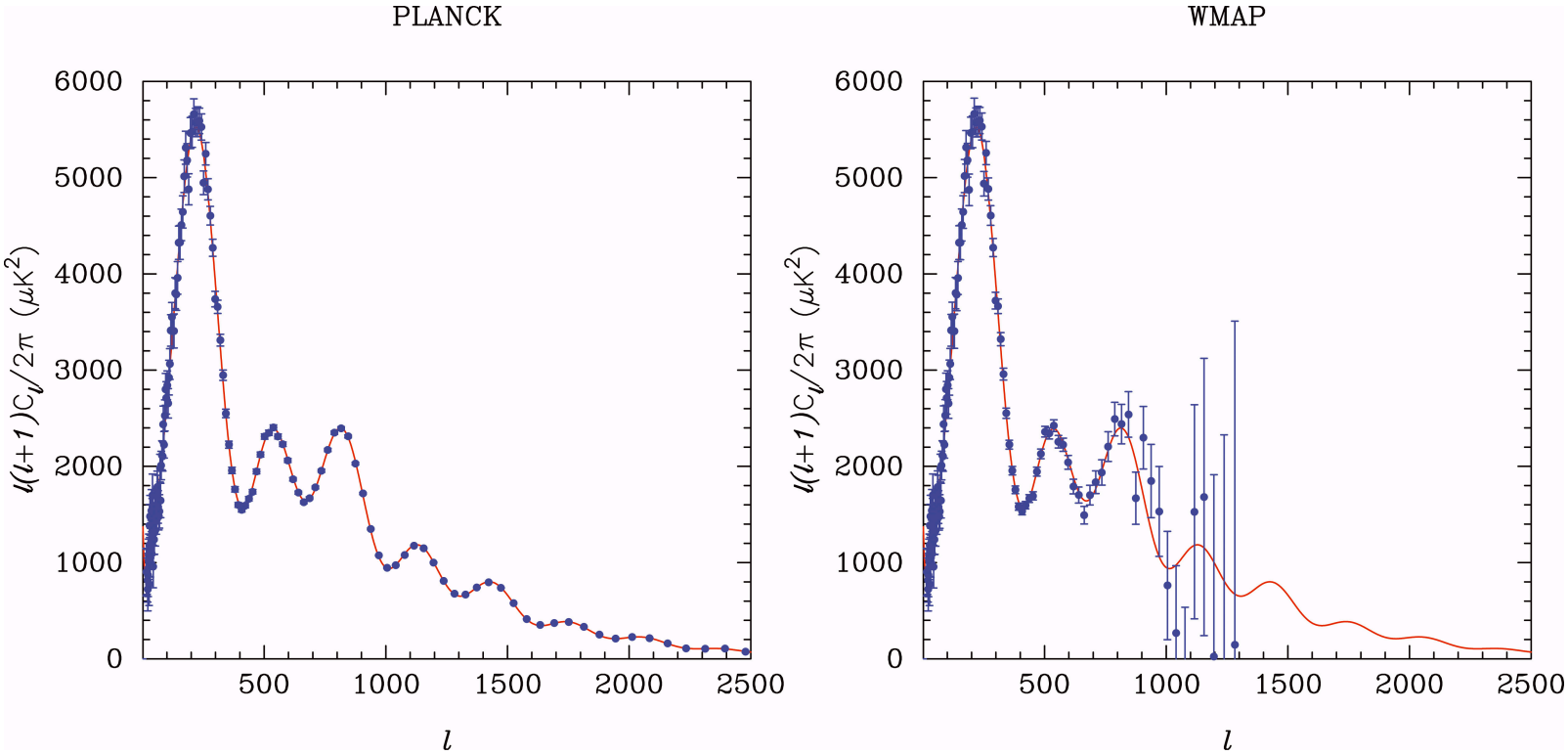
BERKELEY LAB

CMB Satellite Data Evolution



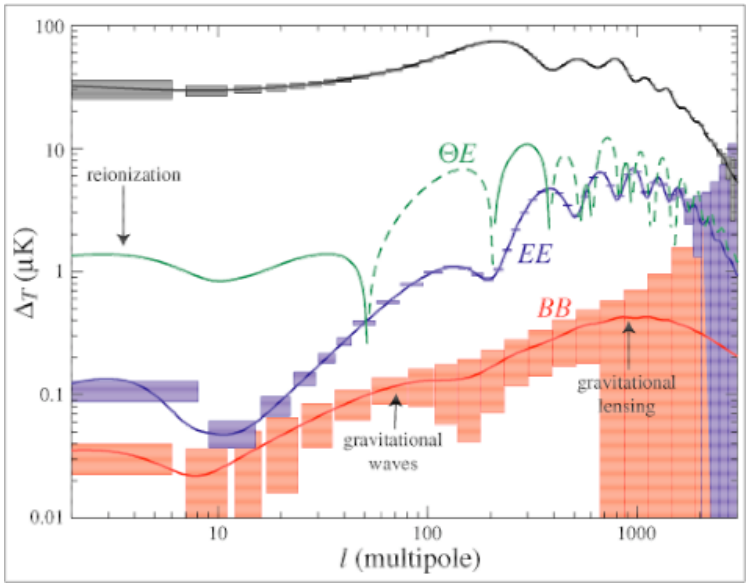
Credit: ESA

Planck versus Previous Instruments



Outline

- The Science
- The Planck Mission
- The Data Pipeline
- NERSC
- Big Data at NERSC
- Big Data Challenges



How the data gets to NERSC

Planck

Ground Station (ESA)

LFI DPC

HFI DPC

Data Processing Centers (Paris and Trieste)

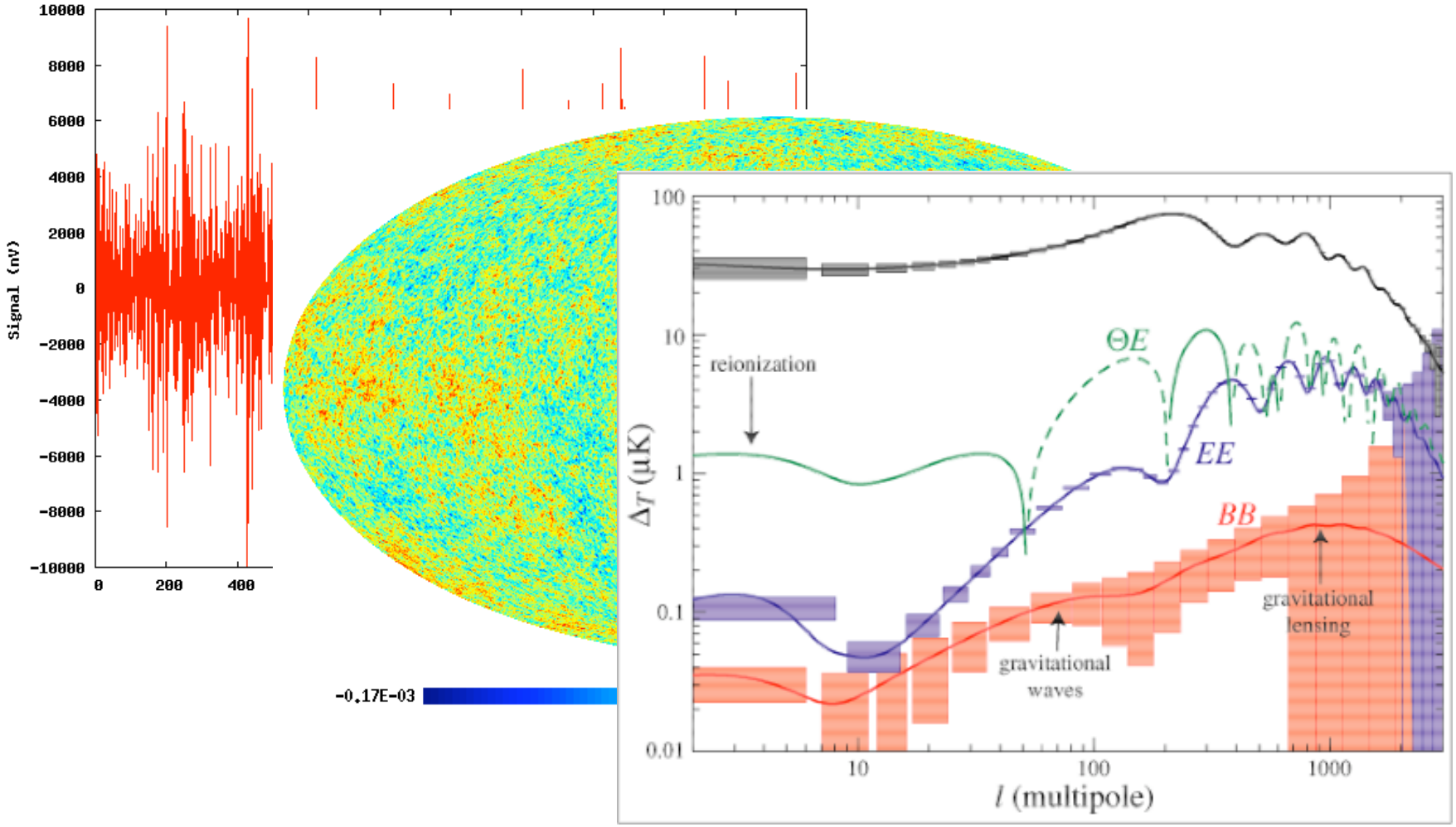
ipac

NASA/JPL

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

What Does The CMB Look Like ?



CMB Data Analysis

- In principle very simple
  - Assume Guassianity and maximize the likelihood
    - of maps given the data and its noise statistics (analytic).
    - of power spectra given the maps and their noise statistics (iterative).
- In practice very complex
  - Foregrounds, asymmetric beams, non-Gaussian noise, etc.
  - Algorithmic scaling with data volume.
    - Correlated data precludes divide-and-conquer.
    - Data simulation scales as *at least* $O(N_t)$, and usually significantly more.
    - Maximum likelihood map-making scales as $O(N_i\, N_t \log N_t)$.
    - Maximum likelihood power spectrum estimation scales as $O(N_i\, N_l\, N_p^3)$.
    - Monte Carlo power spectrum estimation scales as O(simulation) + O(map-making) per realization

The CMB Data Challenge

Extracting fainter signals (polarization mode, angular resolution) from the data requires:
- larger data volumes to provide higher signal-to-noise.
- more complex analyses to remove fainter systematic effects.

Experiment	Date	Time Samples	Sky Pixels	Gflop/Map
COBE	1989	10^9	10^3	1*
BOOMERanG	2000	10^9	10^5	10^3
WMAP	2001	10^{10}	10^6	10^4
Planck	2009	10^{11}	10^7	10^5
PolarBear	2012	10^{12}	10^6	10^6
QUIET-II	2015	10^{13}	10^6	10^7
CMBpol	2020+	10^{14}	10^8	10^8

- 1000x data increase over next 15 years
  - need to continue to scale on the bleeding edge through the next 10 M-foldings !

CMB Data Analysis Evolution

Data volume & computational capability dictate our analysis approach.

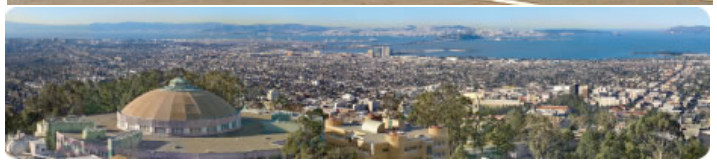
Date	Data	System	Map	Power Spectrum
2000	B98	Cray T3E x 700	Explicit Maximum Likelihood (Matrix Invert - N_p^3)	Explicit Maximum Likelihood (Matrix Cholesky/Tri-solve - N_p^3)
2002	B2K2	IBM SP3 x 3,000	Explicit Maximum Likelihood (Matrix Invert - N_p^3)	Explicit Maximum Likelihood (Matrix Invert/Multiply - N_p^3)
2003-7	Planck subsets	IBM SP3 x 6,000	PCG Maximum Likelihood (FFT - $N_t \log N_t$)	Monte Carlo (Sim + Map - many N_t)
2007+	Planck full	Cray XT4 x 40,000	PCG Maximum Likelihood (FFT - $N_t \log N_t$)	Monte Carlo (SimMap - many N_t)

Planck Map Making

- MADmap code
  - PCG solver for maximum likelihood map given noise statistics
- 2005: First map-making of one year of data from all detectors at one frequency
  - 75 billion observations mapped to 150 million pixels
  - First science code to use all 6,000 CPUs of Seaborg
- 2007: First map-making of data from all detectors at all frequencies (FFP)
  - 750 billion observations mapped to 150 million pixels
  - Using 16,000 cores of Franklin
  - IO & its scaling become very significant issues
    - Write-dominated simulations & read-dominated analyses
- 2008: First on-the-fly simulation capability
  - Single frequency 1 year sim/map is now a 512-way 30 minute debug job !
- 2009: First on-the-fly Monte Carlo sim/map
  - 100x FFP

Outline

- The Science
- The Planck Mission
- The Data Pipeline
- NERSC
- Big Data at NERSC
- Big Data Challenges



NERSC Mission

- NERSC is the Flagship computing Center for the Department of Energy's Office of Science.
- NERSC is National facility operated by Lawrence Berkeley National Lab
- Focused on delivering production High-Performance Computing to non-classified research.

Office of Science
U.S. Department of Energy

Office of Science
U.S. DEPARTMENT OF ENERGY

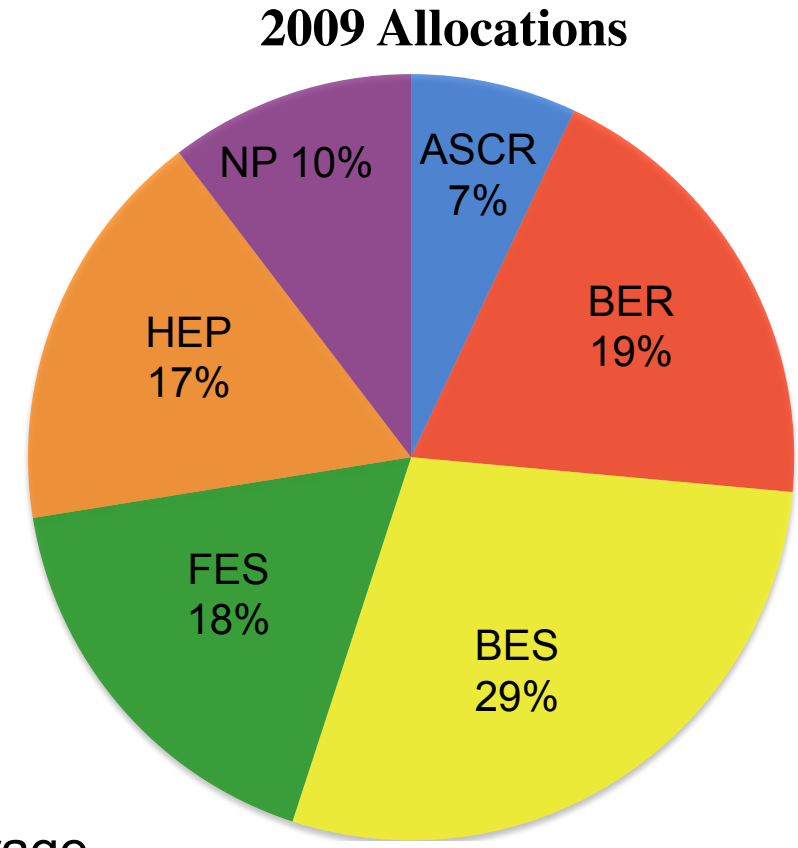
BERKELEY LAB

NERSC History

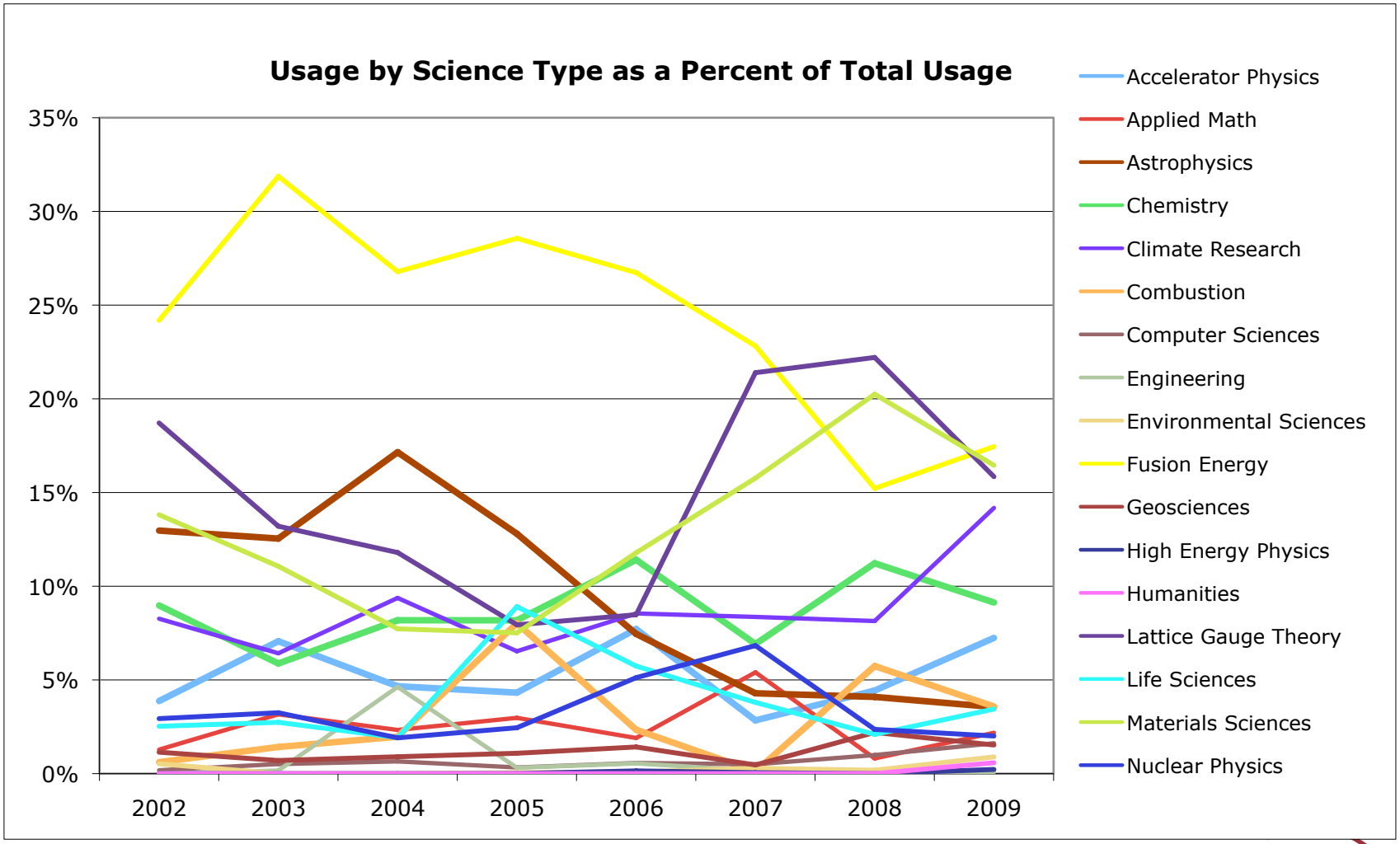
- Began in 1974 at Lawrence Livermore to support fusion science
- Moved to Berkeley Lab in 1996
- Today NERSC is the Department of Energy's Office of Science Flagship computing center.

NERSC is the Production Facility for DOE SC

- NERSC serves a large population
  - Approximately 3000 users,
  - 400 projects, 500 code instances
- Focus on "unique" resources
  - –High end computing systems
  - –High end storage systems
    - Large shared file system
    - Tape archive
  - –Interface to high speed networking
    - ESNet soon to be 100 Gb/s
- Allocate time / storage
  - –Current processor hours and tape storage



Constantly Changing Mix



ASCR's Computing Facilities

NERSC

LBNL

- Hundreds of projects
- 2010 allocations:
  - 70-80% SC offices control; ERCAP process
  - 10-20% ASCR (new ALCC program)
  - 10% NERSC reserve
- Science covers all of DOE/SC science

Leadership Facilities

ORNL and ANL

- Tens of projects
- 2010 allocations:
  - 70-80% ANL/ORNL managed; INCITE process
  - 10-20% ACSR (new ALCC program)
  - 10% LCF reserve
- Science areas limited to those at largest scale; not limited to DOE/SC

NERSC Systems

Large-Scale Computing System
Franklin (NERSC-5): Cray XT4
- 9,740 nodes; 38,288 Opteron cores,
- 8 GB of memory per node
- 26 Tflop/s sustained SSP (355 Tflops/s peak)

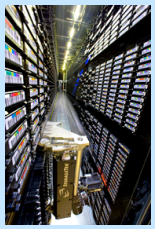
NERSC-6 (XT5) planned for 2010 production
- 3-4x NERSC-5 in application performance

Clusters

- Bassi IBM Power5 (888 cores)
- Jacquard LNXI Opteron (712 cores)
- New Nehalem / IB Cluster
- PDSF (HEP/NP)
  - Linux cluster (~1K cores)

NERSC Global Filesystem (NGF)
- 400 TB; 5.5 GB/s
- GPFS based

HPSS Archival Storage
- 60 PB capacity
- 10 Sun robots
- 130 TB disk cache

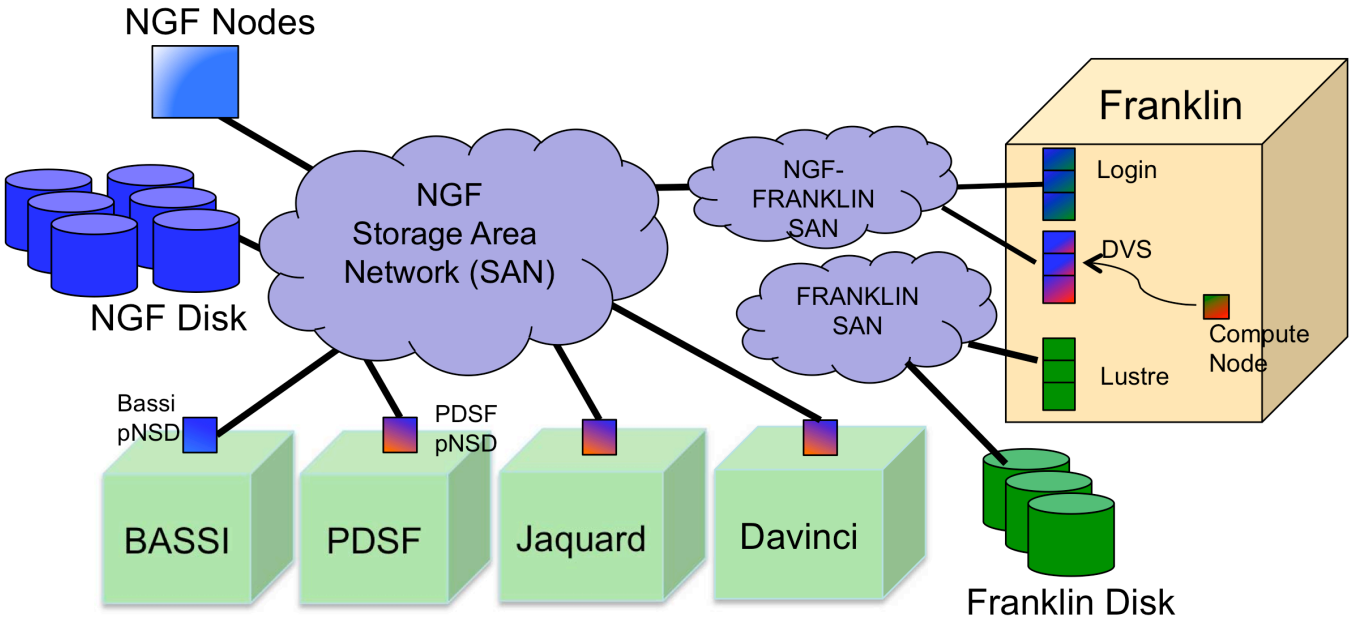


Analytics / Visualization
- Davinci (SGI Altix)

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

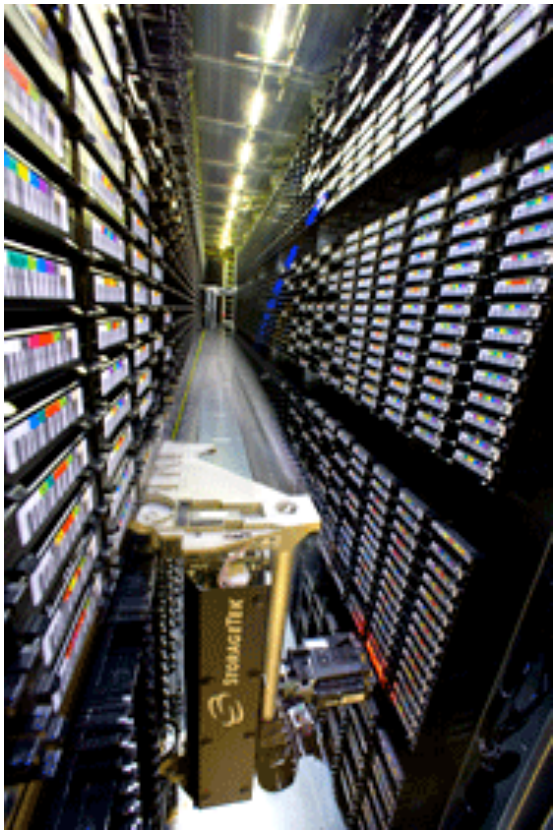
NERSC Global File System (NGF)



- A facility-wide, high performance, parallel file system
  - Uses IBM's GPFS technology for scalable high performance
  - Makes users more productive

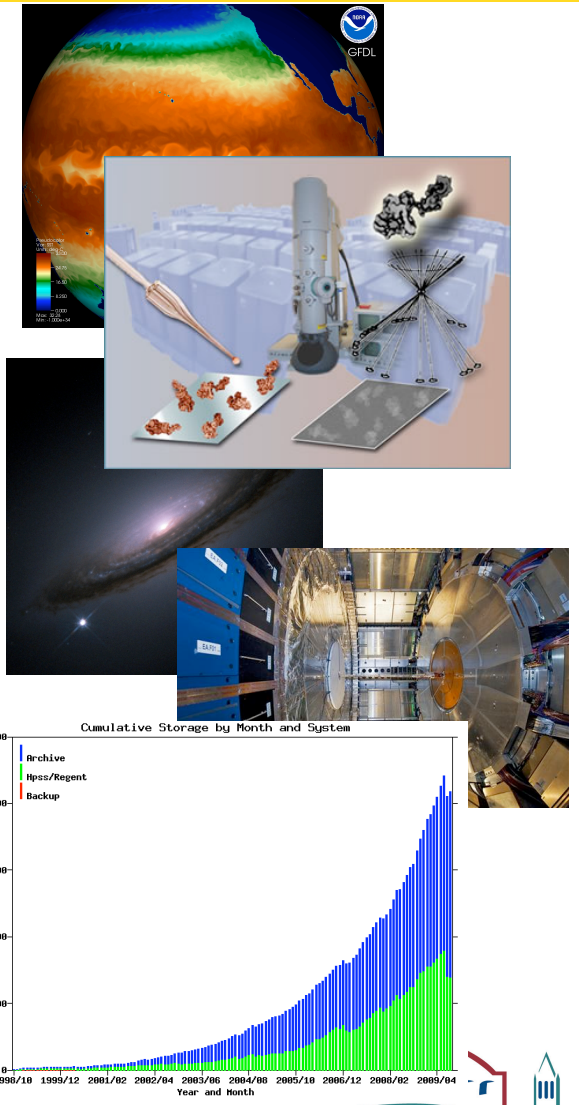
Outline

- The Science
- The Planck Mission
- The Data Pipeline
- NERSC
- Big Data at NERSC
- Big Data Challenges



Data Driven Science

- Ability to generate data is challenging our ability to store, analyze, & archive it.
  - Some observational devices grow in capability with Moore's Law.
  - Data sets are growing exponentially.
- Petabyte (PB) data sets soon will be common:
  - *Climate:* next IPCC estimates 10s of PBs
  - *Genome:* JGI alone will have .5 PB this year and double each year
  - *Particle physics*: LHC projects 16 PB / yr
  - *Astrophysics*: LSST, others, estimate 5 PB / yr
- Redefine the way science is done?
  - One group generates data, different group analyzes
- Turning point: in 2003 NERSC changed from being a data source to a data sink



Selected NERSC Data Intensive Projects

Project	Category	Compute Hours	Storage RUs
Supernovae Factory	Astrophysics	14k	1.8M
Palomar Transient Factory	Astrophysics	20k	600k
CMB: PLANCK +	Astrophysics	680k	500k
STAR	Nuclear Physics	-	8M
KamLAND	Nuclear Physics	-	4M
ALICE	Nuclear Physics	10k	2.2M
PCMDI	Climate	20k	2M
CCSM	Climate	12M	2M
20th Century ReAnalysis	Climate	8M	4M
John Bell	Chem/Comb/Math	5.5M	7.5M
Lattice QCD	High Energy Phys	1.4M	2M
JGI	Biological Science	10k	2M

Cosmic Microwave Background

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

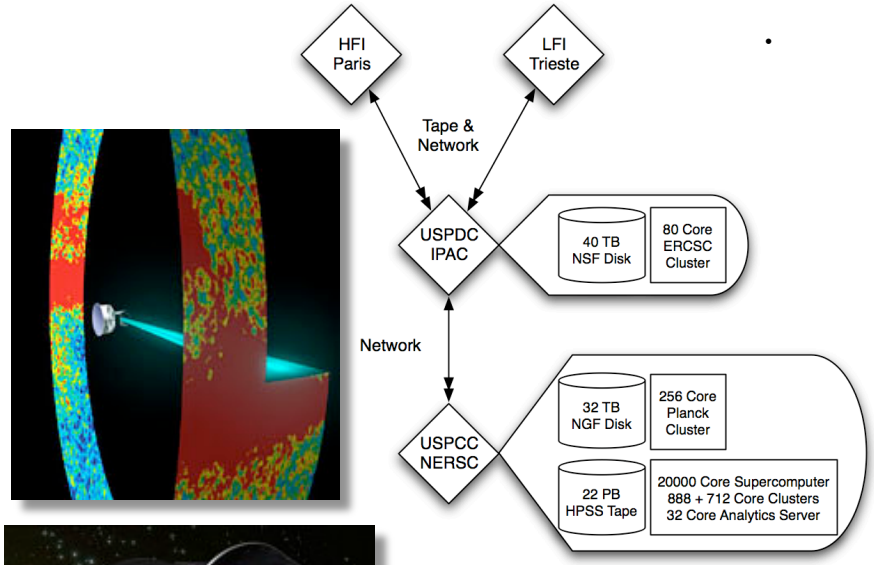
***Objective*: Analyze data from the Planck satellite -- definitive Cosmic Microwave Background (CMB) data set.**

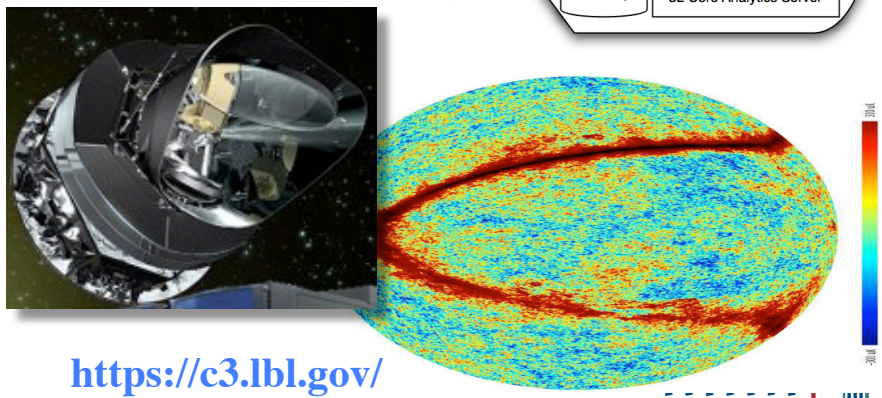
***Implications*: CMB: image of the universe at 400k years, relic radiation from Big Bang**

***Accomplishments:* NERSC provides the components of the data pipeline for noise reduction, map-making, power spectrum analysis, and parameter estimation**

- **2006 Nobel Prize in Physics**
- **32 TB final data set size, ~400 users**
- **data sets analyzed as a whole because complex data correlations; no "divide and conquer"**
- **Launched May09, first "light" Sept09**
- **Also ~10k-core XT4 MonteCarlo calibration runs, produce ~10X data**
- **Anticipate Moore's law growth in data set size for *15 years***

PI: J. Borrill (LBNL)





https://c3.lbl.gov/

BERKELEY LAB

KamLAND

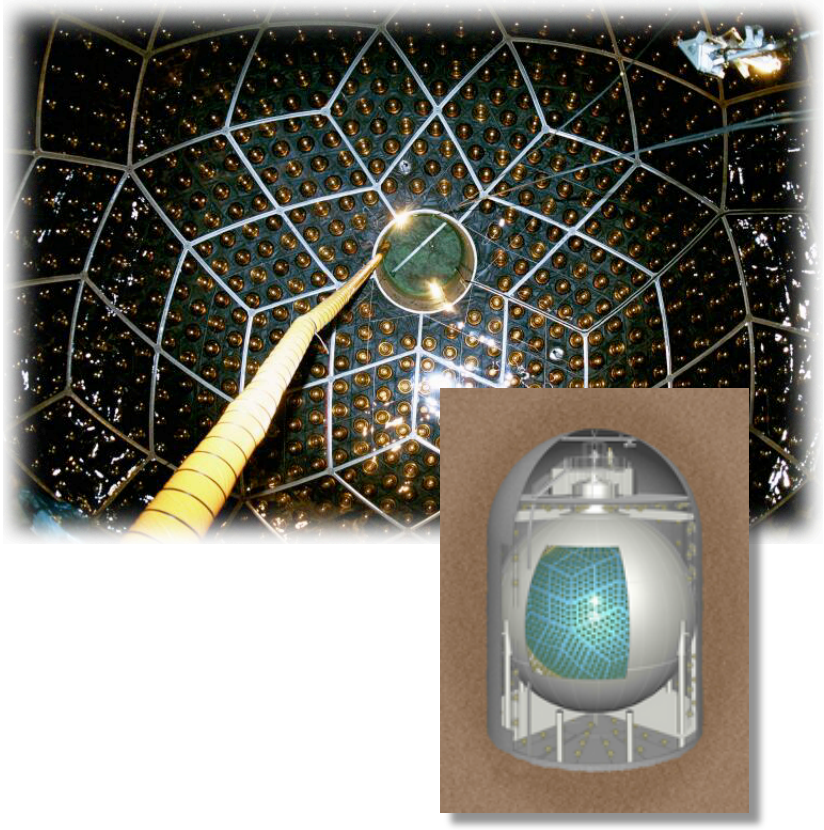
***Objective*: Archive, analyze all stages of the US data from Kamioka Liquid Scintillator Anti-Neutrino Detector**

***Implications*: Substantially increased our scientific knowledge of neutrinos**

***Accomplishments:* Many significant physics milestones – neutrino oscillation, precise value for the neutrino oscillation parameter, etc.**

- **NERSC resources instrumental in reactor neutrino analysis and the preparations for the solar phase;**
- **Currently recording data at trigger rate of 100Hz, data rate of 200GB/day, 365 days/yr**
- **0.6 PB of data stored from 6 years; plan to read large fraction of this in 2010**

PI: S. Freedman (UCB)



http://kamland.lbl.gov/

ALICE

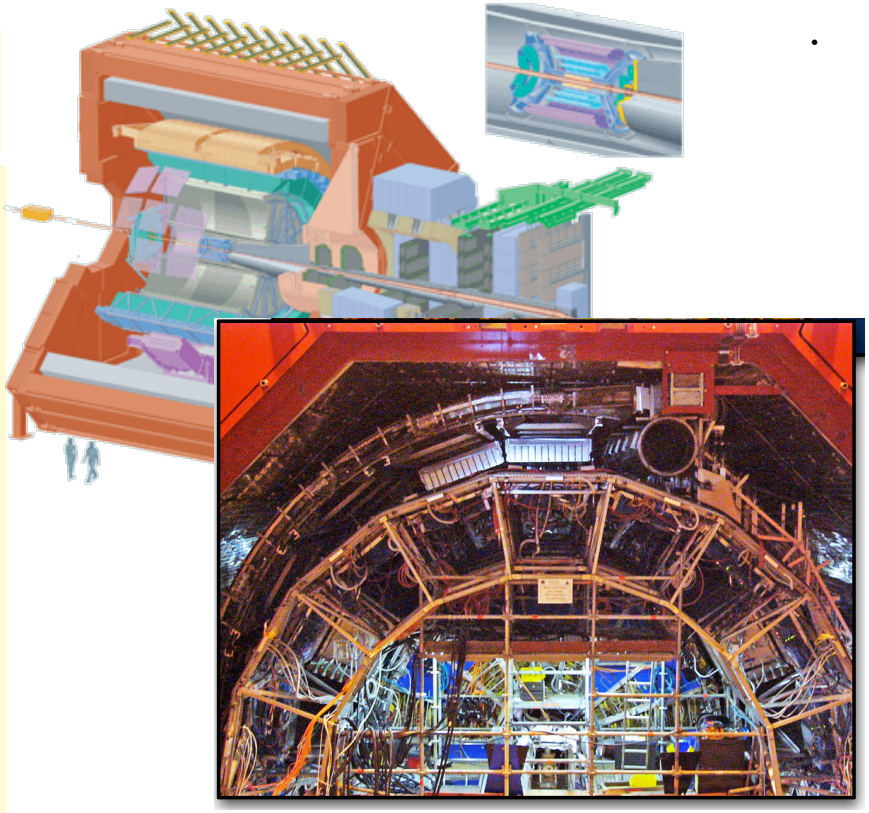
***Objective*: Data analysis and simulations for the ALICE heavy-ion detector experiment at the LHC.**

***Implications*: Understanding of dense QCD matter.**

***Notes:* Uses (primarily) NERSC's PDSF cluster + LLNL + Grid resources;**

- **Expect ~600TB of data distributed over 1GB files, ~25% of USA obligation in 2010.**
- **Challenge of providing direct-charged resources for experimentation that might be delayed.**
- **Simulation resources to reconstruct and analyze detector events prior to the experiment.**
- **Longer term: Estimate 3.8 PB of disk space and 5.31 PB of HPSS in 2013, accessible by international community.**

PI: P. Jacobs (LBNL)



http://aliceinfo.cern.ch/Collaboration/

20th Century Climate Reanalysis

NERSC NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

***Objective*: Use an Ensemble Kalman filter to reconstruct global weather conditions in six-hour intervals from 1871 to the present.**

***Implications*: Validate tools for future projections by successfully recreating – and explaining – climate anomalies of the past.**

***Accomplishments:* First complete database of 3-D global weather maps for the 19th to 21st centuries.**

- **Provide missing information about the conditions in which extreme climate events occurred.**
- **Reproduced 1922 Knickerbocker storm, comprehensive description of 1918 El Niño**
- **Data can be used to validate climate and weather models**

PI: G. Compo (U. Colorado)

Bull. Am. Meteorological Soc. (2009)

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Cloud-Resolving Climate Model

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

***Objective*: Climate models that fully resolve key convective processes in clouds; ultimate goal is 1-km resolution.**

***Implications*: Major transformation in climate/ weather prediction, likely to be standard soon, just barely feasible now.**

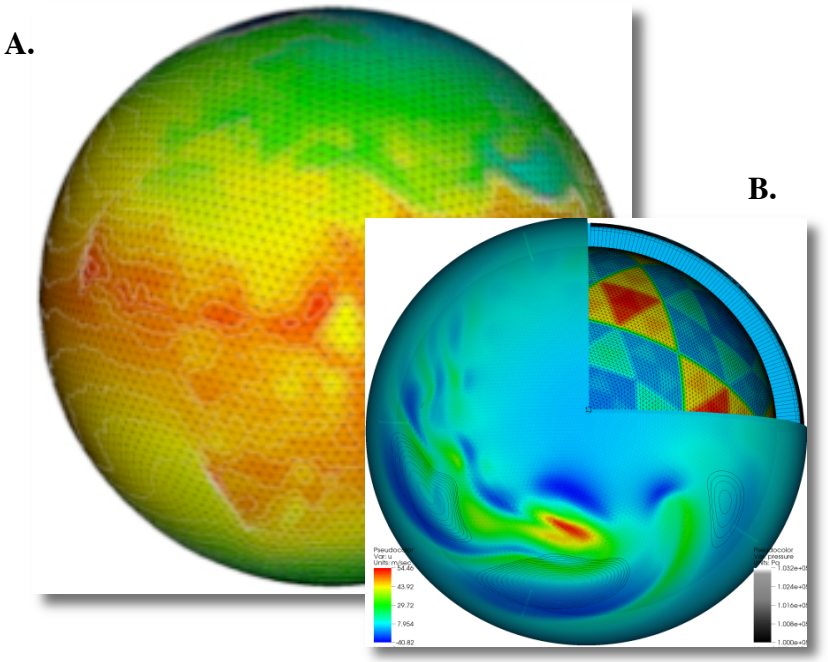
***Accomplishments:* Developed a coupled atmosphere-ocean-land model based on geodesic grids.**

- **Multigrid solver scales perfectly on 20k cores of Franklin using grid with 167M elements.**
- **Invited lecture at SC09.**

NERSC:

- **3-km 24-hr run, 30k cores = 10TB output**
- **NERSC/LBNL played key role in developing critical I/O code & Viz infrastructure to enable analysis of ensemble runs and icosohedral grid.**

PI: D. Randall, Colo. St



A.Surface temperature showing geodesic grid. B.Composite plot showing several variables: wind velocity (surface pseudocolor plot), pressure (b/w contour lines), and a cut-away view of the geodesic grid.

Joint Genome Institute

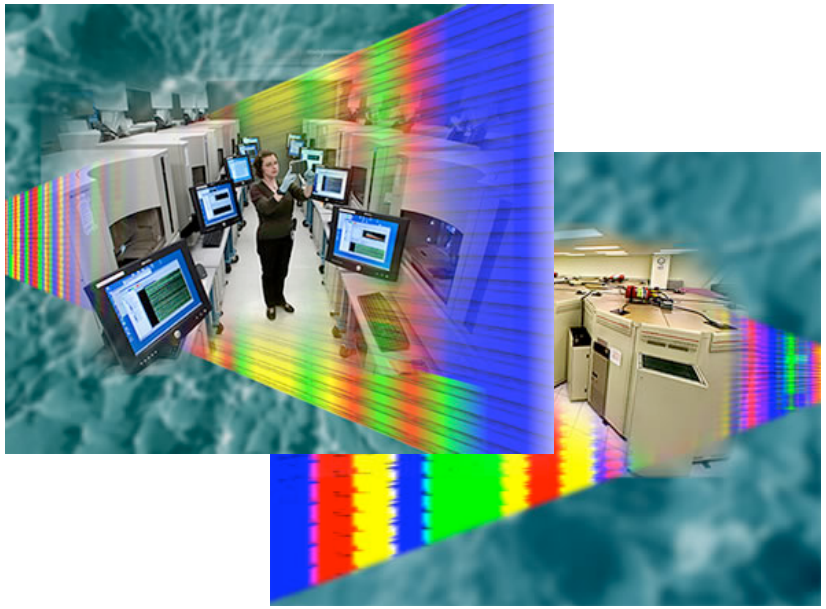
***Objective*: Archive all production and R&D data from three sequencing platforms at JGI**

***Implications*: One of the world's largest public DNA sequencing facilities.**

***Accomplishments:* NERSC, JGI staff collaborated to set up nightly back-up pipeline using ESnet's new Bay Area MAN.**

- **Archiving sequencing data at NERSC allowed JGI to scale up infrastructure with minimal additional DOE investment.**
- **Data import expected to grow nearly exponentially in 2010; impossible to maintain data onsite at the JGI HQ.**
- **NERSC/DOE JGI collaboration to develop improved techniques for data access, handling.**
- **Note: additional Microbial Genome project**

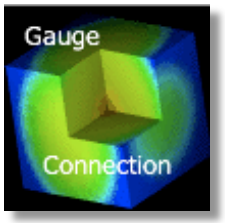
PI: E. Rubin (LBNL)

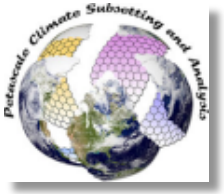


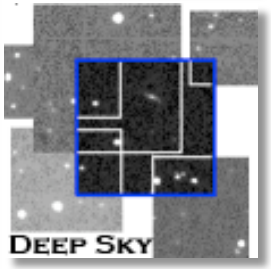
JGI is producing sequence data at increasing rate: 2 million files per month of trace data (25 to 100 KB each) plus 100 assembled projects per month (50 MB to 250 MB); total about 2 TB per month on average.

Science Gateways

- Create scientific communities around data sets
  - NERSC HPSS, NGF accessible by broad community for exploration, scientific discovery, and validation of results
  - Increase value of existing data
- S*cience gateway: custom* hardware, software to provide remotely data/computing services
  - Deep Sky – "Google-Maps" for astronomical image data
    - Discovered 36 supernovae in 6 nights during the PTF Survey
    - 15 collaborators worldwide worked for 24 hours non-stop
  - GCRM – Interactive subselection of climate data (pilot)
  - Gauge Connection – Access QCD Lattice data sets
  - Planck Portal – Access to Planck Data
- New models of computational access
  - Projects with mission-critical time constraints require guaranteed turn-around time.
  - Reservations for anticipated needs: Computational Beamlines
  - Friendly interfaces for applications and workflows







Deep Sky Science Gateway

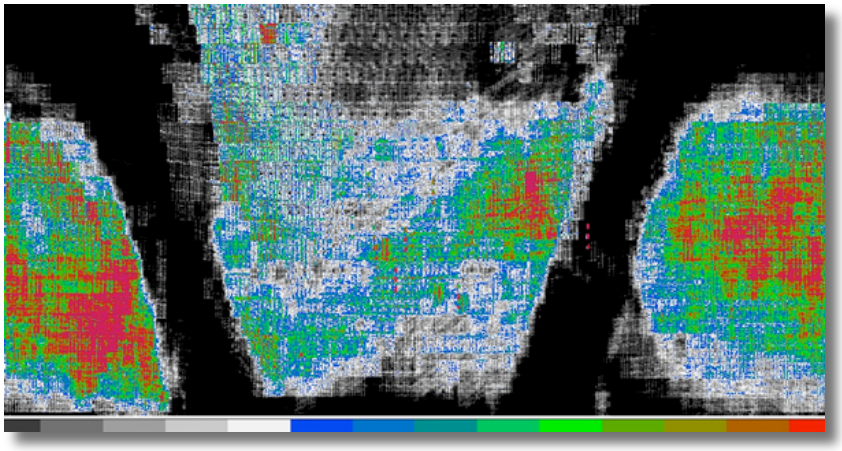
***Objective*: Pilot project to create a richer set of compute- and data-resource interfaces for next-generation astrophysics image data, making it easier for scientists to use NERSC and creating world-wide collaborative opportunities.**

***Implications*: Efficient, streamlined access to massive amounts of data – some archival, some new -- for broad user communities.**

***Accomplishments:* Open-source Postgres DBMS customized to create Deep Sky DB and interface: www.deepskyproject.org**

- **90TB of 6-MB images stored in HPSS / NGF (biggest NGF project now)**
  - **-- images + calibr. data, ref. images, more**
  - **-- special storage pool focused on capacity not bandwidth**
- **Like "Google Earth" for astronomers?**

PI: C. Aragon (NERSC)



Map of the sky as viewed from Palomar Observatory; color shows the number of times an area was observed

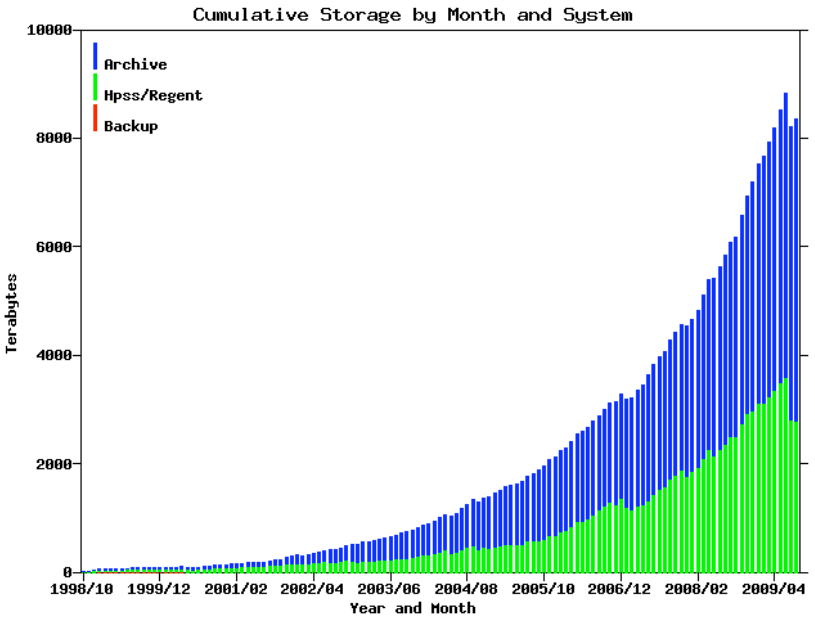
Testbeds, etc.

- GPU/Accelerator Testbed
  - Large-memory (.5 TB RAM) with Nvidia Tesla (1 TF) GPU accelerators
  - Experiment with GPU accelerated sequence matching and OpenCL/CUDA programming model
  - Gain experience with administration of this kind of platform
- Cloud Computing Testbed (NERSC/ANL: Magellan)
  - Distributed, multi-institution dynamically expandable computing resource
  - Experiment with cost effectiveness of cloud computing paradigm, including Amazon EC2 evaluation
- Solid State/FLASH Accelerated I/O
  - Later
- FPGA Accelerator Testbed (LBL Computing Research Division)
  - Convey HC1 FPGA accelerator with 80GB/s vector memory subsystem: can be programmed with “custom personalities” for bioinformatics applications

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Outline

- The Science
- The Planck Mission
- The Data Pipeline
- NERSC
- Big Data at NERSC
- Big Data Challenges



Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Keys to Success for Working with Large Data

- Common community data models
- Parallelism at all levels (file systems, clients, application)
- Design for failures and data integrity
- Large I/O operations are critical for good bandwidth. Avoid I/O if possible (B/S ~ 1000x FLOPS in cost)
- Include data provenance and metadata (Data from observation and experiment gain value over time - compared with simulation)

Enabling Technologies

- Scalable archive systems – HPSS has evolved over a decade and scales to multi PB archives
- Parallel File Systems (GPFS, Lustre, Panasas, etc) – POSIX compliant, scalable file systems
- Data Management Middleware – HDF5, NetCDF, MPI-IO can help organize data and assist in scaling
- Data Analysis and Visualization tools – VisIT, root, VTK

MapReduce and Hadoop

- MapReduce/Hadoop
  - Initially championed by Google
  - Hadoop is an open source implementation
  - Designed to scale
  - Fault Tolerance built into the application framework
  - Data redundancy and check-summing
  - Move the work to the data (coupling between the file system and task manager)
  - Well suited to unstructured data

Enterprise Storage Trends

According to COMPUTERWORLD

- Virtualization
- **10G versus Fibre Channel**
- Disk versus Tape
- Deduplication
- **Flash-based Storage**

And

- Magnetic Disk capacity continue to grow

Magnetic Storage Trends

- Heated Assisted Magnetic Recording (HAMR) and Bit-patterned media could help push magnetic media to 50 Tb/ in^2.
- **Bad News:** Capacity scales with aerial density, BW scales with $\sqrt{}$.
- Also, SAS replacing FC for Enterprise market

Converged Fabrics

- Converged Enhanced Ethernet enables SAN and IP over a single fabric
  - Adds flow control and congestion management to Ethernet
  - Makes Ethernet a “loss-less” fabric
- InfiniBand also supports these capabilities
  - InfiniBand SRP for storage
  - Native RDMA support
- Even if a system doesn’t use a converged fabric this trends impacts the evolution of SANS and Networks

Flash Technology Trends

- Solid state storage predicted to match disk storage in $/GB by 2014 timeframe (but possibly much later)
- However, impact could be felt sooner. (Will R&D investment levels in magnetic media continue if SSDs take over consumer market?)
- $/IOPS is competitive, especially for ~1 TB- solutions
- Moore's Law Challenge with shrinking…
- Phase Change Memory or SONOS could replace NAND (faster and more reliable)

*Chart courtesy of Micron, Pictures courtesy of Intel, Samsung, FusionIO website

Gaps and Challenges for Large Data

- Capacity continues to outpace bandwidth for Magnetic storage
  - Flash Storage may help
  - Better Middleware can help to efficiently utilize the bandwidth
- Need an ecosystem of next generation utilities
  - Still relying on serial based utilities (first step is to multithread for multi-core)
  - Example: copying, tar and compressing data
- Data Integrity
  - Bit Error rates haven't improved for HDD (1 in 10^{14} or 10^{15} bits)
  - Data integrity checks on read are becoming a must and will only increase in importance
  - Exa-scale will require end-to-end checks
  - See CERN study
- Life-cycle management
  - From curation to preservation
  - Data can long outlive the experiment that generated it.

Summary

- Planck has now begun its mission to provide the next generation of data on the CMB. These results will further improve our understanding of the Universe.
- Experiments like Planck rely on processing and managing large data sets.
- Big Data is enabling the next wave of Science in a variety of fields (Astrophysics, Particle Physics, Climate, Bioinformatics)
- Big Data computing is still in its infancy. There are many challenges on the horizon (from life-cycle management to building an ecosystem of tools).
- The other challenges are social in building communities around data with common data models and standards.

Acknowledgments

Special thanks to Julian Borrill
Member of the Planck Collaboration, PI for Planck at NERSC, NERSC Power User
Thanks to the Planck Collaboration for images and animation
Other images courtesy of Wikimedia Commons

Contact Info:

Shane Canon

Canon at nersc dot gov

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC

Berkeley Lab