

# **Big Bang, Big Data, Big Iron: High Performance Computing for Cosmic Microwave Background Data Analysis**

Julian Borrill

Computational Cosmology Center, Berkeley Lab  
Space Sciences Laboratory, UC Berkeley

*Cosmologists are often in error,  
but never in doubt.*

- Lev Landau

# **A Brief History Of Cosmology**

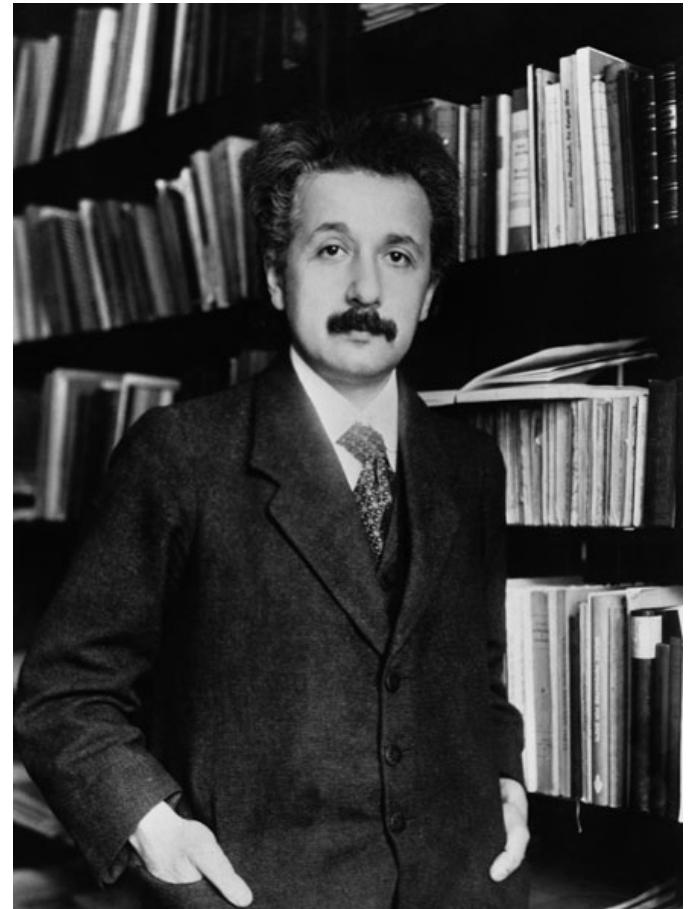
# 1916 – General Relativity

- General Relativity
  - Space tells matter how to move
  - Matter tells space how to bend

$$G_{\mu\nu} = 8 \pi G T_{\mu\nu}$$

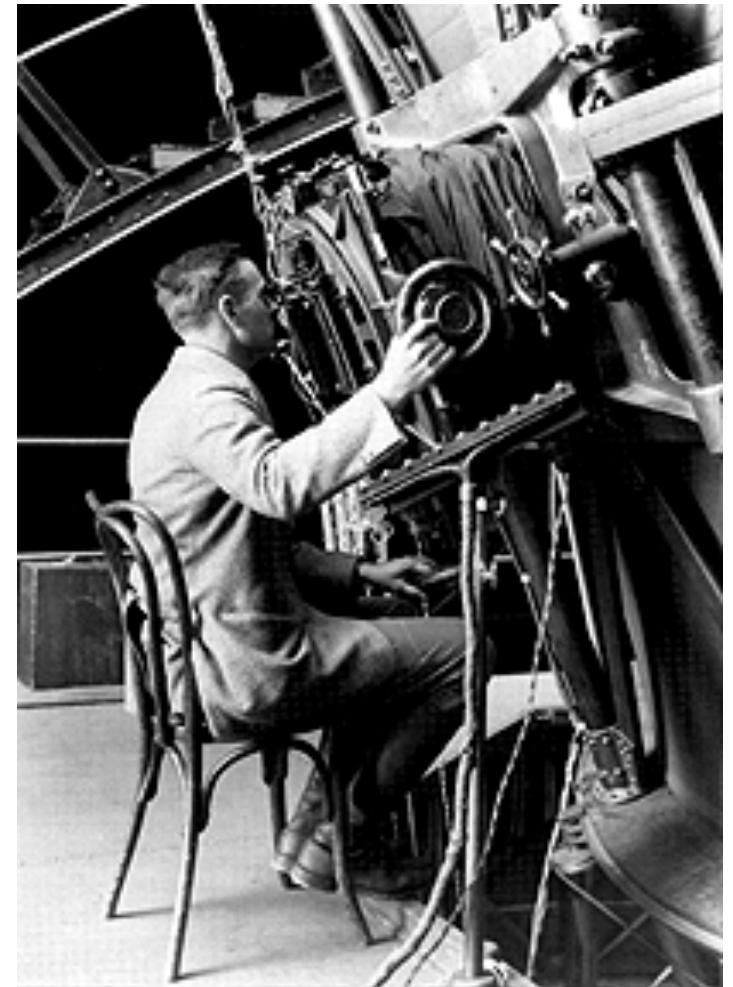
Space      Matter

- But this implies that the Universe is dynamic and everyone *knows* it's static ...
- ... so Einstein adds a Cosmological Constant (even though the result is unstable equilibrium)



# 1929 – Expanding Universe

- Using the Mount Wilson 100-inch telescope Hubble measures nearby galaxies'
  - velocity (via their redshift)
  - distance (via their Cepheid variables)and finds velocity proportional to distance.
- Space is expanding!
- The Universe is dynamic after all.
- Einstein calls the Cosmological Constant "my biggest blunder".



# 1930-60s – Steady State vs Big Bang

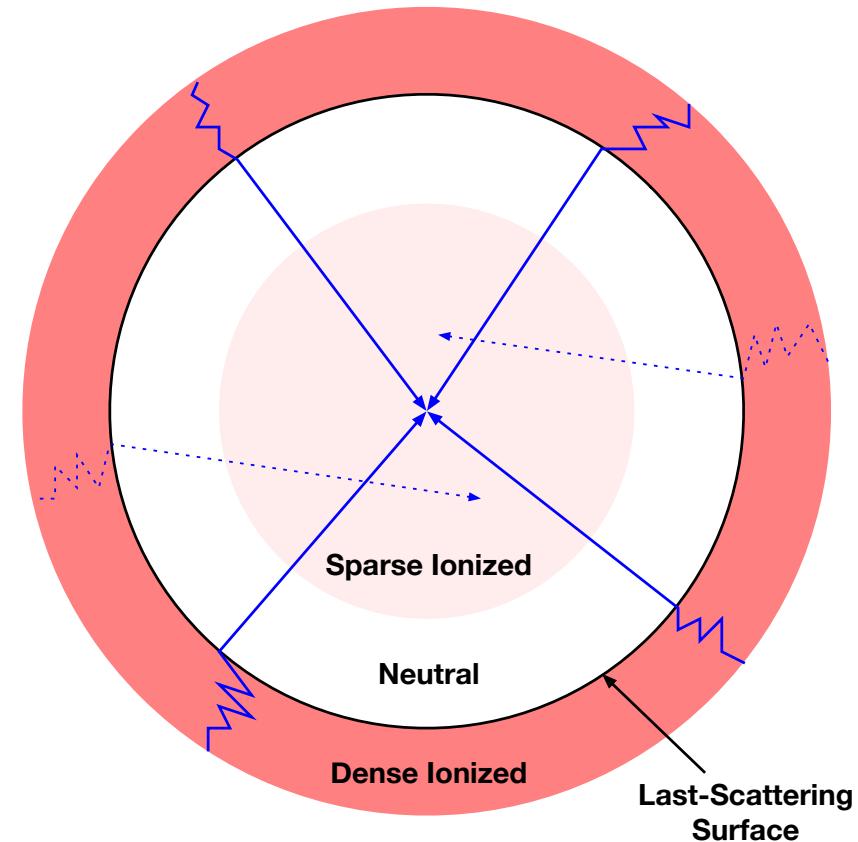
- What does an expanding Universe tells us about its origin and fate?
  - Steady State Theory:
    - new matter is generated to fill the space created by the expansion, and the Universe as a whole is unchanged and eternal (past & future).
  - Big Bang Theory:
    - the Universe (matter and energy; space and time) is created in a single explosive event, resulting in an expanding and hence cooling & rarifying Universe.

# Big Bang Cosmology

- In a Big Bang cosmology, the early Universe is hot, dense, ionized and expanding; primordial photons are constantly scattering off the free electrons.
- As the Universe expands it cools, eventually dropping below the ionization temperature of hydrogen ( $\sim 3000\text{K}$ ).
- Protons and free electrons form hydrogen atoms and the Universe becomes neutral. From this time on the primordial photons simply free-stream throughout the Universe.
- The ongoing expansion of the Universe redshifts these photons, progressively reducing their temperature.
- In 1948 Alpher & Herman predicted that this remnant radiation would be a 5K black body today.

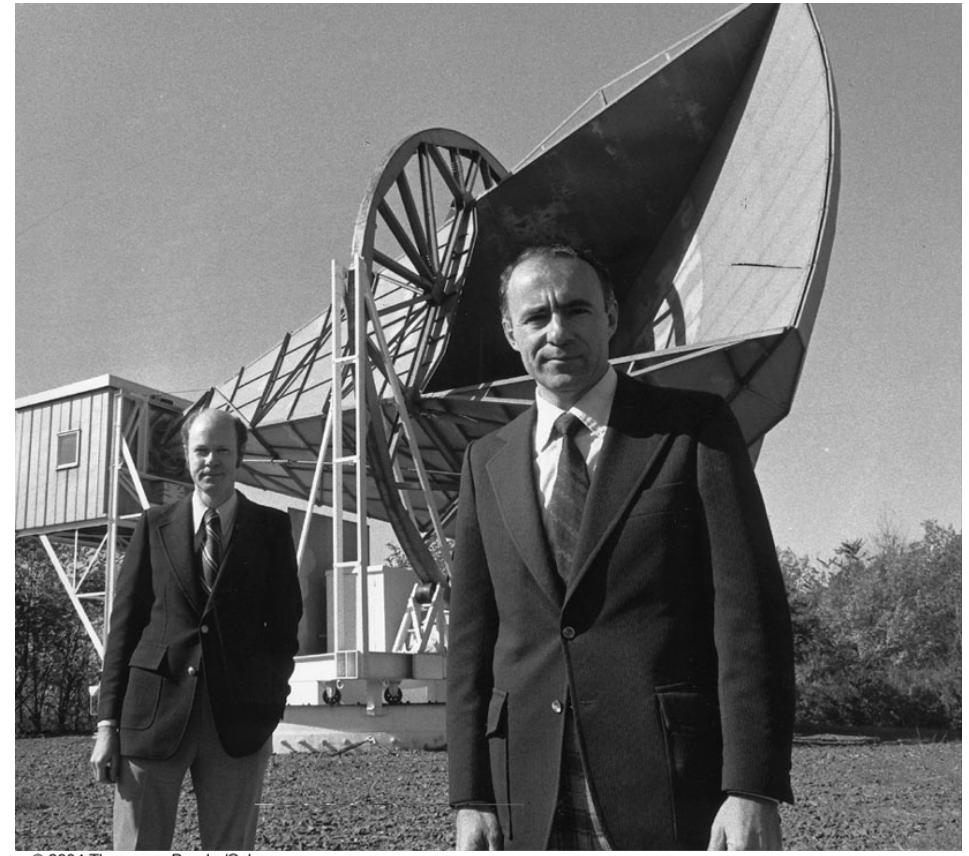
# 1948 – Cosmic Microwave Background

- In a Big Bang Universe the hot, expanding Universe eventually cools through the ionization temperature of hydrogen:  $p^+ + e^- \Rightarrow H$ .
- Without free electrons to scatter off, the photons free-stream to us.
- Alpher, Herman & Gamow predict a residual photon field at 5 – 50K
- 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.



# 1964 – First CMB Detection

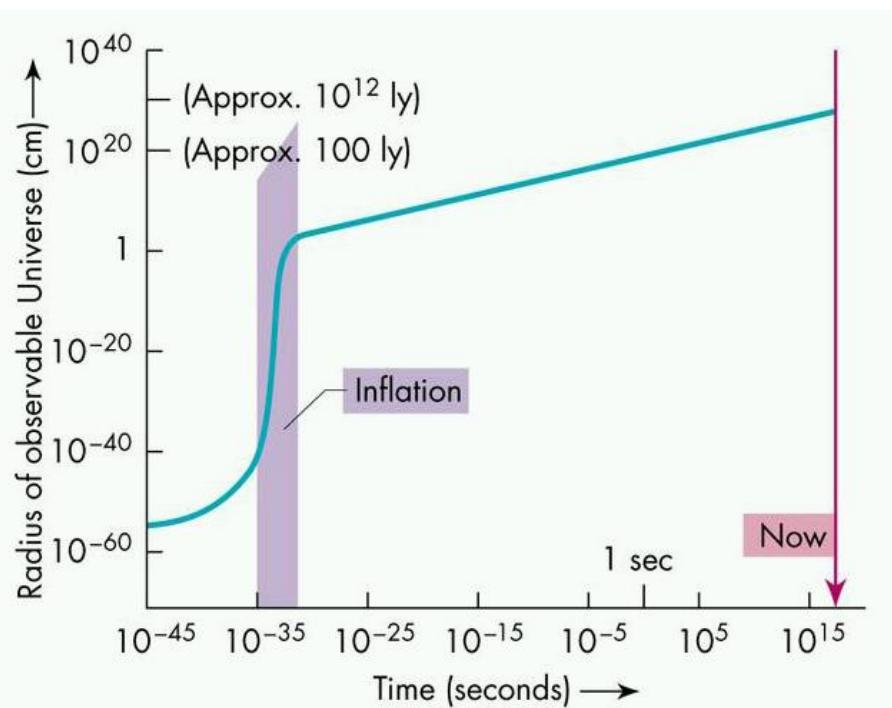
- Penzias & Wilson find a puzzling signal that is constant in time and direction.
- They determine it isn't a systematic – not terrestrial, instrumental, or due to a “white dielectric substance”.
- Dicke, Peebles, Roll & Wilkinson explain to them that they're seeing the Big Bang.
- Their accidental measurement kills the Steady State theory and wins them the 1978 Nobel Prize in physics.



© 2004 Thomson - Brooks/Cole

# 1980 – Inflation

- Increasingly detailed measurements of the CMB temperature show it to be uniform to better than 1 part in 100,000.
- At the time of last-scattering any points more than  $1^\circ$  apart on the sky today are out of causal contact, so how could they have exactly the same temperature? This is the horizon problem.
- Guth proposes a very early epoch of exponential expansion driven by the energy of the vacuum.
- This also solves the flatness & monopole problems.



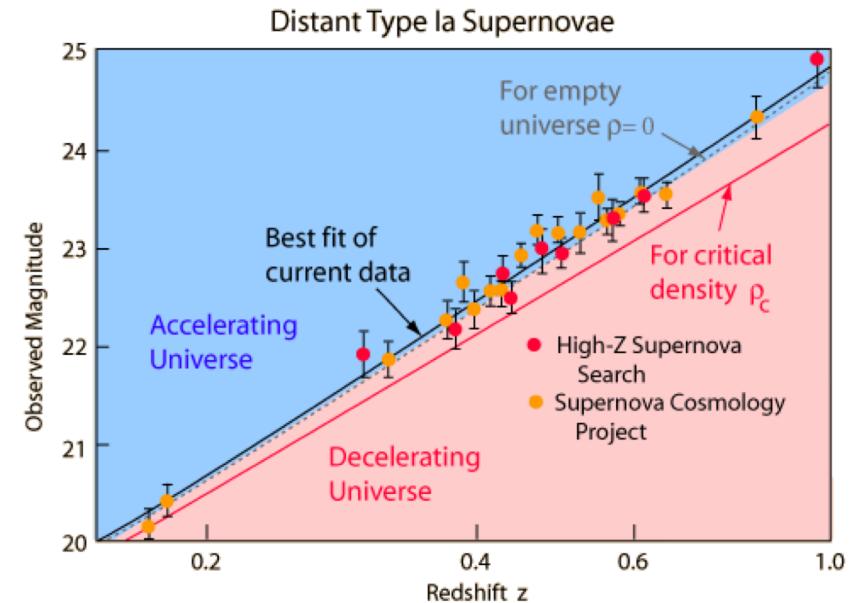
# 1992 – CMB Fluctuations

- For structure to exist in the Universe today there must have been seed density perturbations in the early Universe.
- Despite its apparent uniformity, the CMB must therefore carry the imprint of these fluctuations.
- After 20 years of searching, fluctuations in the CMB temperature were finally detected by the COBE satellite mission.
- COBE also confirmed that the CMB had a perfect black body spectrum, as a residue of the Big Bang would.
- Mather & Smoot share the 2006 Nobel Prize in physics.



# 1998 – The Accelerating Universe

- Both the dynamics and the geometry of the Universe were thought to depend solely on its overall density:
  - Critical ( $\Omega=1$ ): expansion rate asymptotes to zero, flat Universe.
  - Subcritical ( $\Omega<1$ ): eternal expansion, open Universe.
  - Supercritical ( $\Omega>1$ ): expansion to contraction, closed Universe.
- Measurements of supernovae surprisingly showed the Universe is accelerating!
- Acceleration (maybe) driven by a Cosmological Constant!
- Perlmutter/Riess & Schmidt share 2011 Nobel Prize in physics.



# 2000 – The Concordance Cosmology

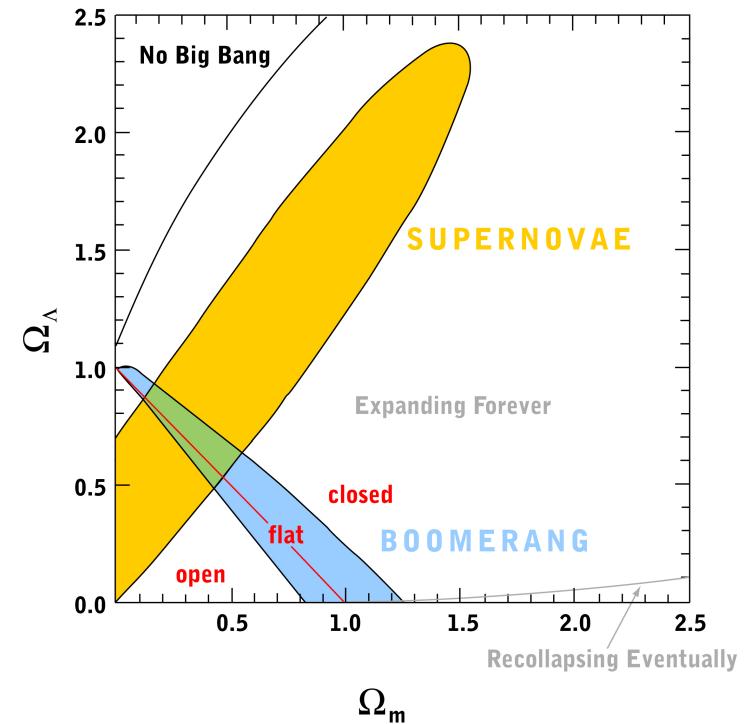
- The BOOMERanG & MAXIMA balloon experiments measure small-scale CMB fluctuations, demonstrating that the Universe is flat.
- CMB fluctuations encode cosmic geometry:  $(\Omega_\Lambda + \Omega_m)$
- Type 1a supernovae encode cosmic dynamics:  $(\Omega_\Lambda - \Omega_m)$
- Their combination breaks the degeneracy in each.

The Concordance Cosmology:

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

=> 95% ignorance!

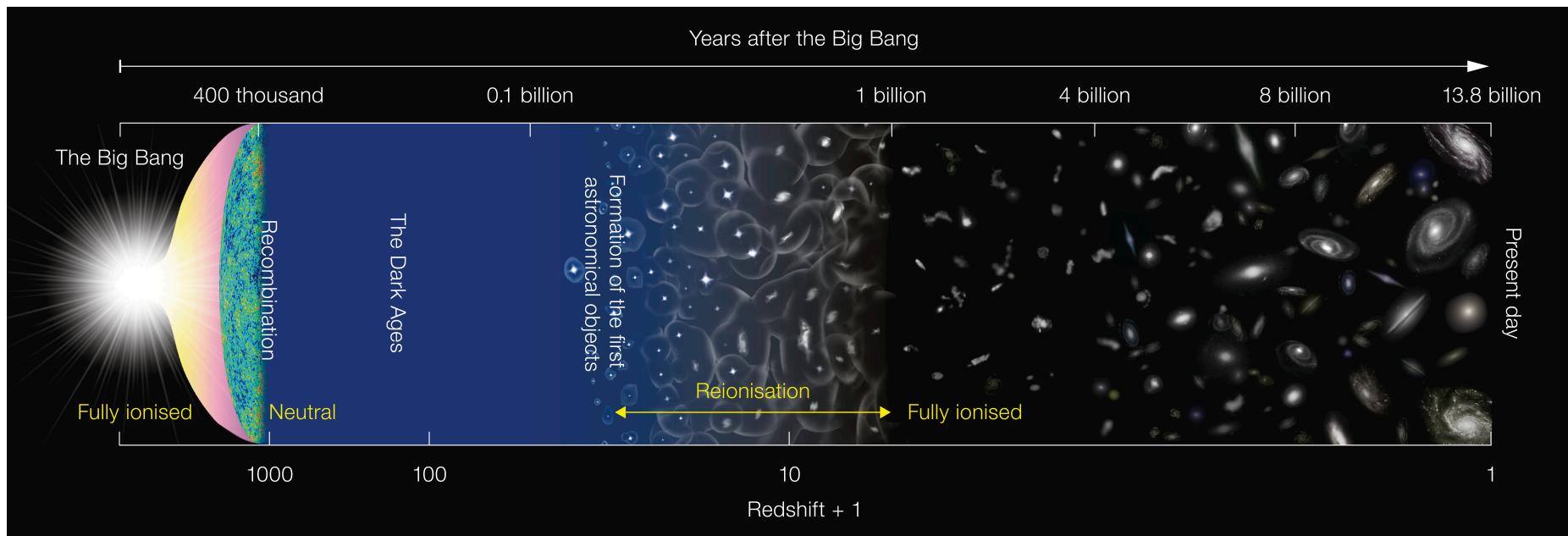
- What and why is the Dark Universe?



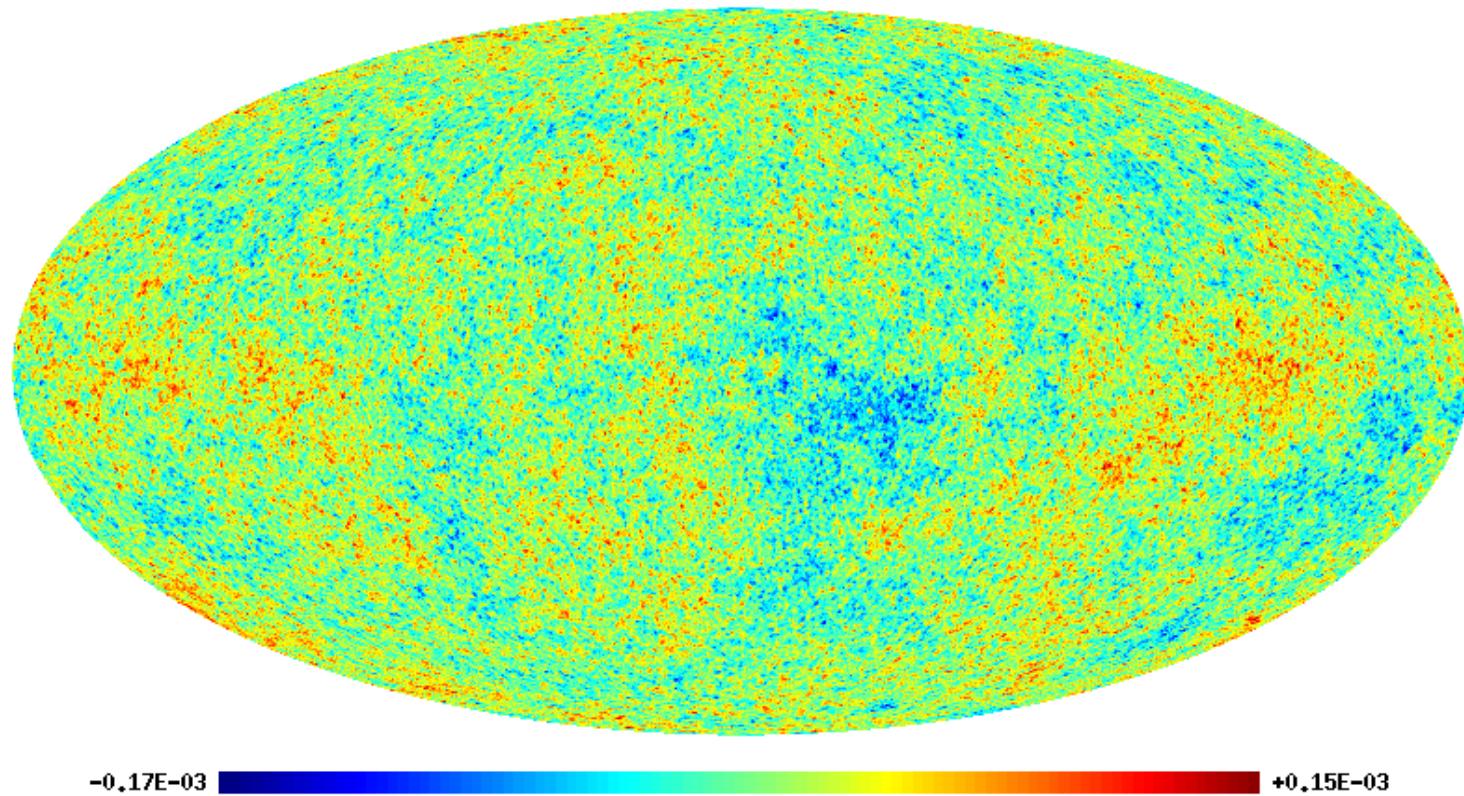
# **The Cosmic Microwave Background**

# CMB Science

- Primordial anisotropies: fundamental physics and cosmology using the early Universe as the ultimate particle accelerator.
- Secondary anisotropies: formation and evolution of structure in the later Universe from gravitational and electromagnetic distortions.
- mm-wave survey: observations of “contaminating” microwave emission from the galaxy and solar system.

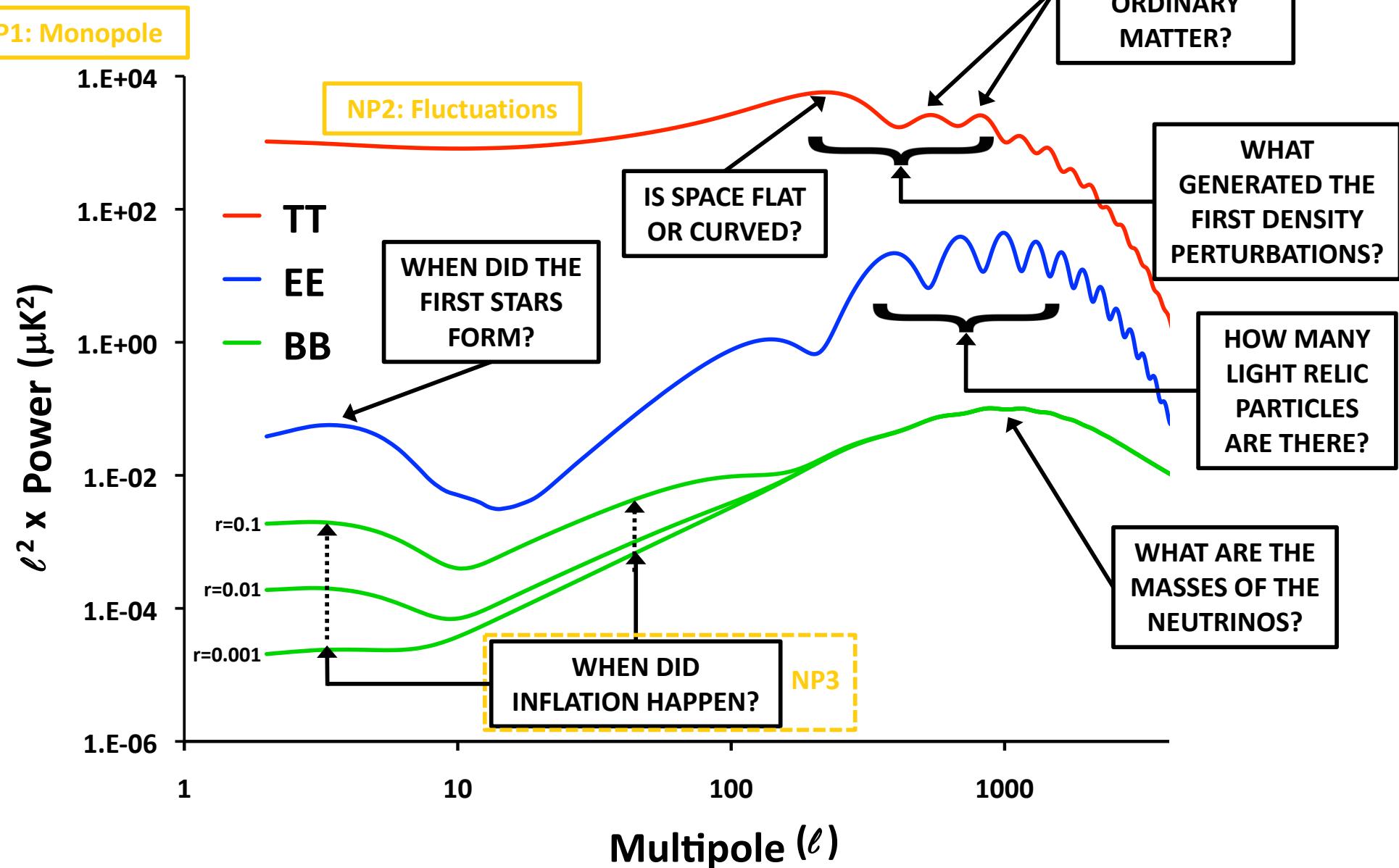


# CMB Fluctuations



- Our view of the CMB sky is only our specific realization.
- To compare it with theory we need a generic statistical characterization.

# CMB Power Spectra

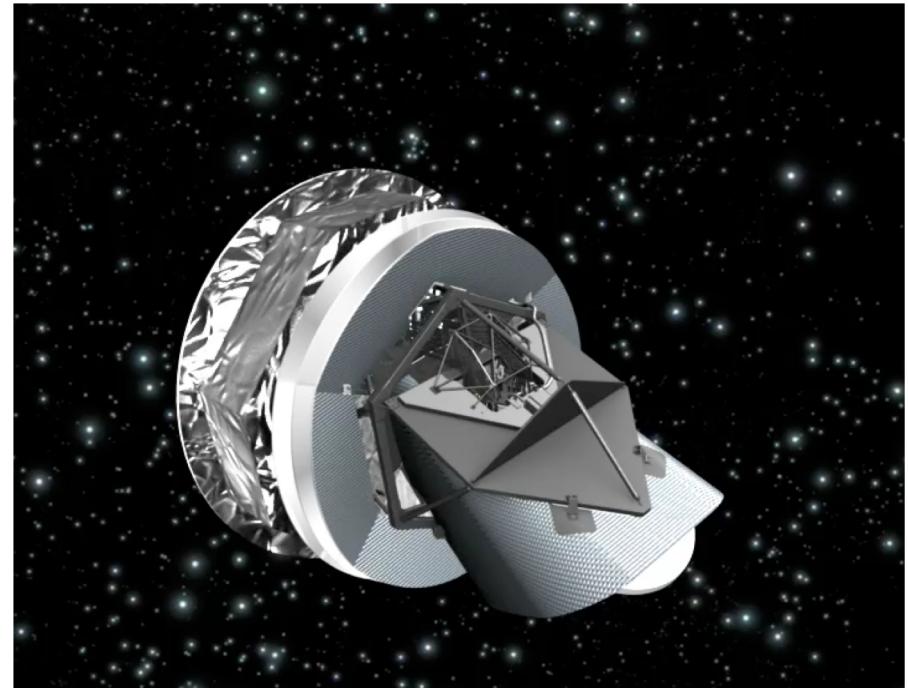


# CMB Signals

COMPONENT	AMPLITUDE (K)	ERA
TT : Monopole	1	1968 (Penzias & Wilson)
TT : Anisotropy	$10^{-5}$	1990 (COBE)
TT : Harmonic Peaks	$10^{-6}$	2000 (BOOMERanG, MAXIMA)
EE : Reionization	$10^{-7}$	2005 (DASI)
BB : Lensing	$10^{-9}$	2015 (SPT, POLARBEAR)
BB : Gravitational Waves	$< 10^{-9}$	2020+

# CMB Observations

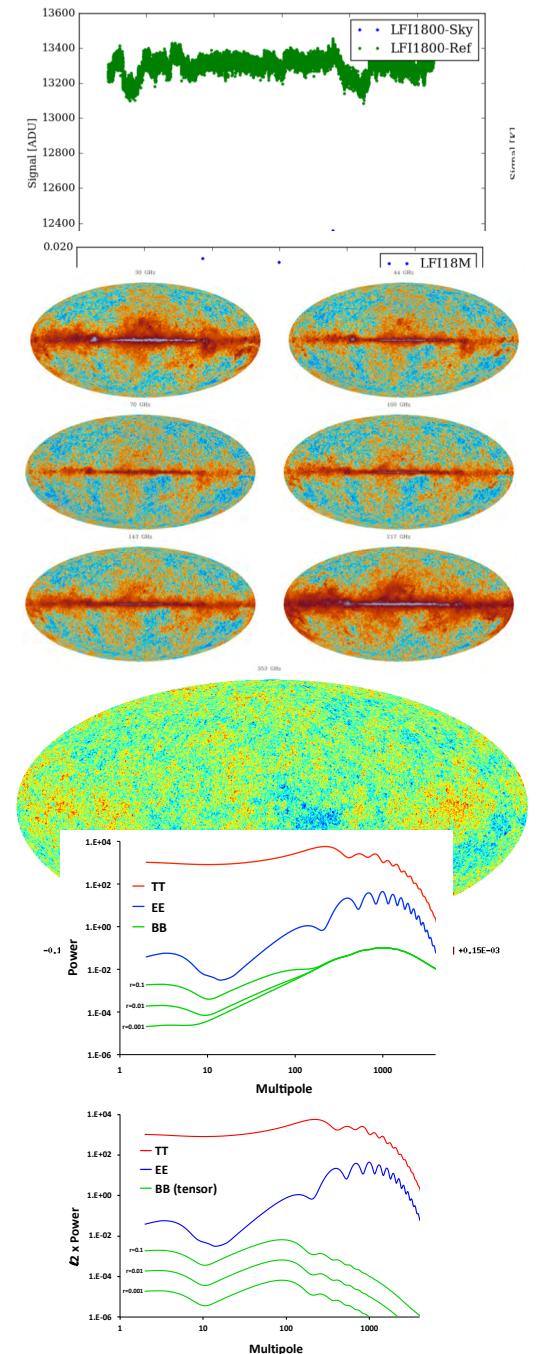
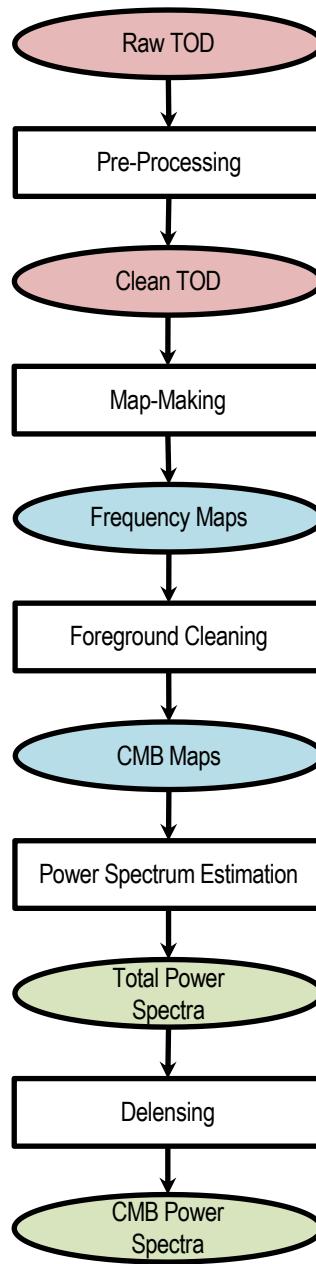
- Searching for micro- to nano-Kelvin fluctuations on a 3 Kelvin background.
- Need very many, very sensitive, very cold, detectors.
- Scan part of the sky from high dry ground or the stratosphere, or all of the sky from space.



# **Cosmic Microwave Background Data Analysis**

# Data Reduction

- An alternating sequence of processes addressing systematic and statistical uncertainties.
- Mitigation within a domain, compression between domains:
  - Time samples
  - Pixels
  - Multipoles
- Must propagate both data *and their covariance* for a sufficient statistic.

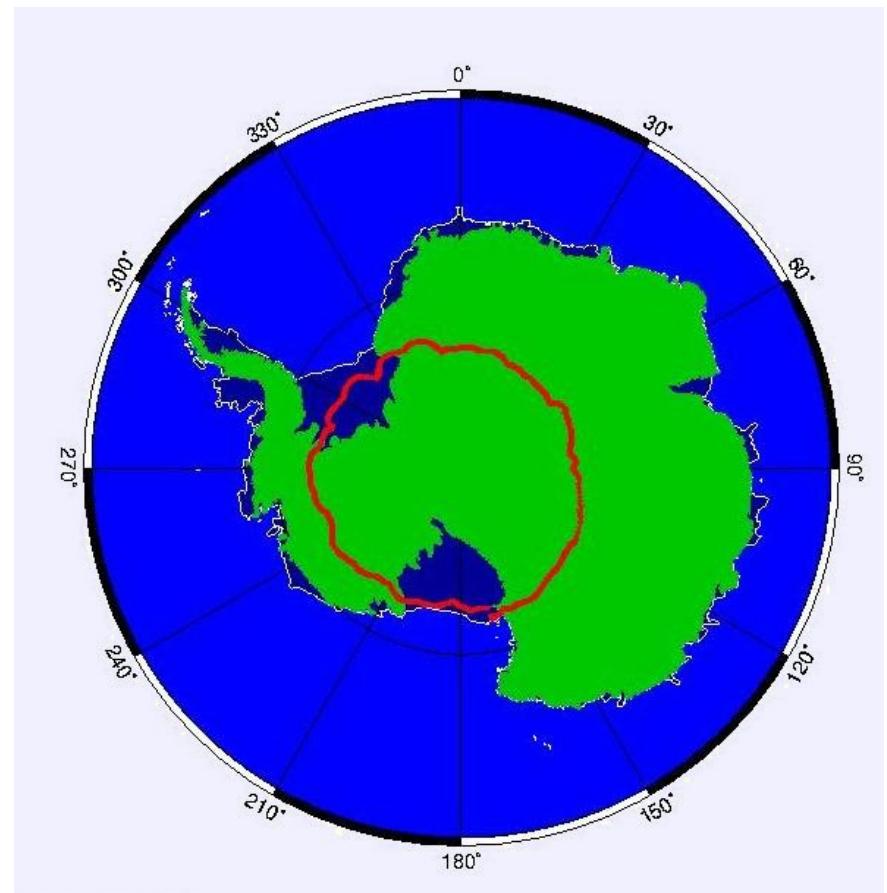


# Data Synthesis

- Synthetic data are required for
  - Design & development of the instrument and observation
    - 10s of realizations during design
  - Validation and verification of the analysis pipeline(s)
    - 100s of realizations during construction
  - Uncertainty quantification & debiasing
    - 1000s of realizations during operations

# Case 1 – BOOMERanG (2000)

- Balloon-borne experiment flown from McMurdo Station.
- Spends 10 days at 35km float, circumnavigating Antarctica
- Gathers temperature data at 4 frequencies: 90 – 400GHz.



# Exact CMB Analysis

- Model data as stationary Gaussian noise and sky-synchronous CMB

$$d_t = n_t + P_{tp} s_p$$

- Estimate the noise correlations from the (noise-dominated) data

$$N_{tt'}^{-1} = f(|t-t'|) \sim \text{invFFT}(1/\text{FFT}(d))$$

- *Analytically* maximize the likelihood of the map given the data and the noise covariance matrix  $N$

$$m_p = (P^T N^{-1} P)^{-1} P^T N^{-1} d$$

- Construct the pixel domain noise covariance matrix

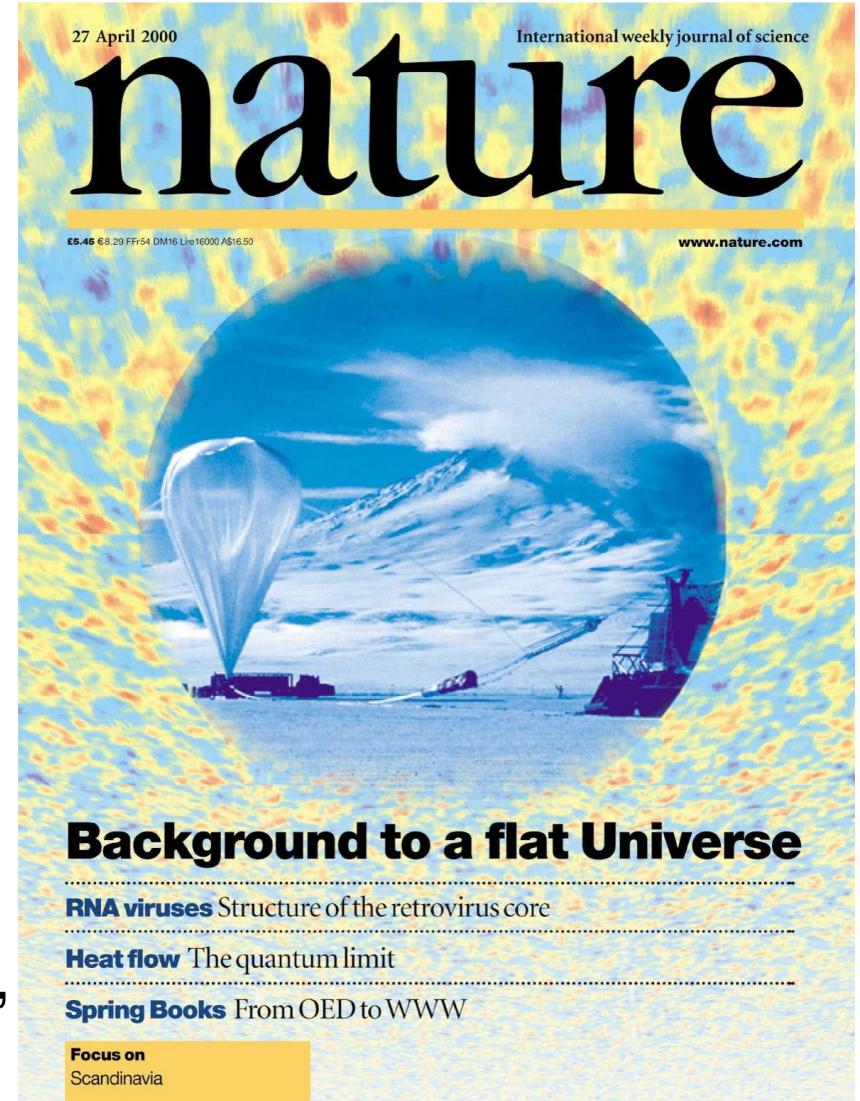
$$N_{pp'} = (P^T N^{-1} P)^{-1}$$

- *Iteratively* maximize the likelihood of the CMB spectra given the map and its covariance matrix  $M = S + N$

$$L(c_l | m) = -\frac{1}{2} (m^T M^{-1} m + \text{Tr}[\log M])$$

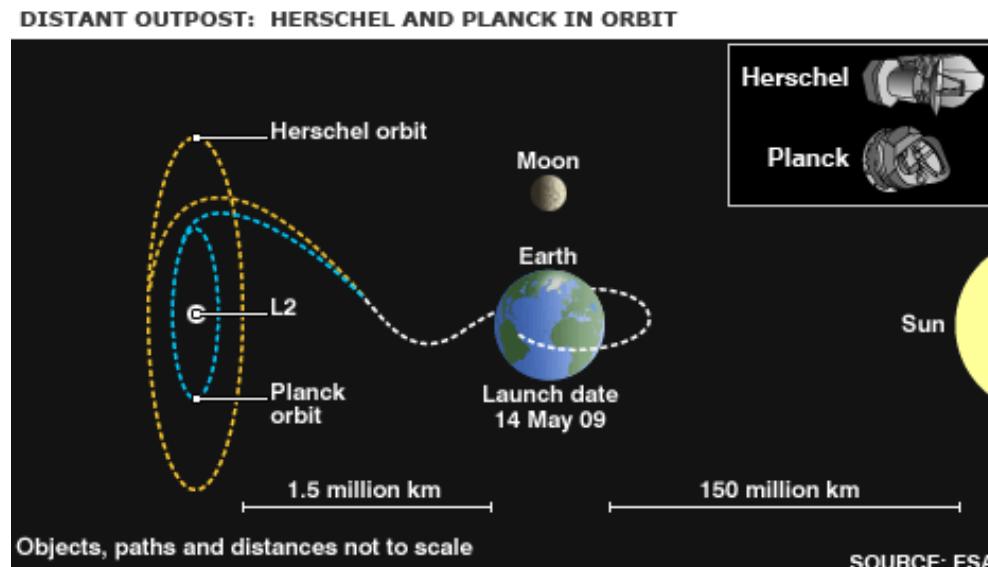
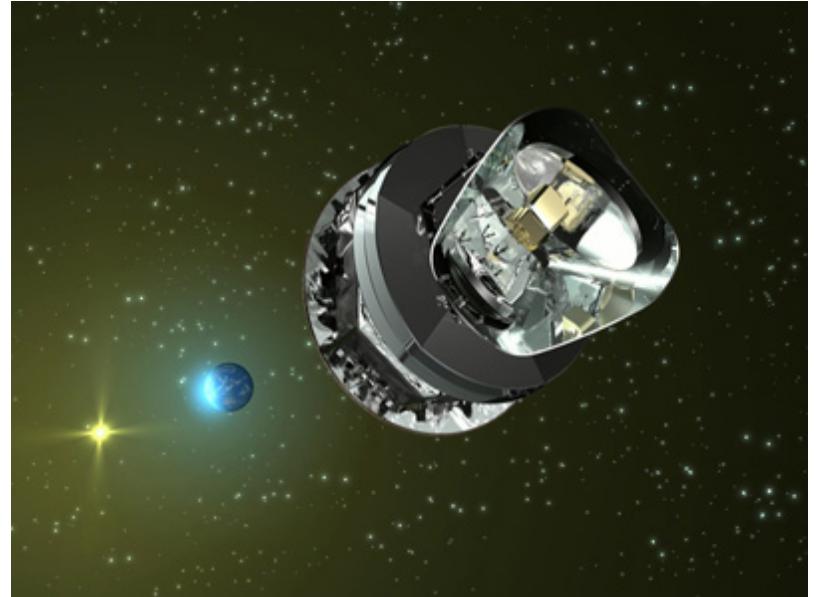
# Algorithms & Implementation

- Dominated by dense pixel-domain matrix operations
  - Inversion in building  $N_{pp}$ ,
  - Multiplication in estimating  $c_l$
- MADCAP CMB software built on ScaLAPACK tools, Level 3 BLAS
  - scales as  $\mathcal{N}_p^3$
- Execution on NERSC's 600-core Cray T3E achieves ~90% theoretical peak performance.
- Spawns MADbench benchmarking tool, used in NERSC procurements.



# Case 2 – Planck (2015)

- European Space Agency satellite mission, with NASA roles in detectors and data analysis.
- Spends 4 years at L2.
- Gathers temperature and polarization data at 9 frequencies: 30 – 857GHz



# The Exact Analysis Challenge

	BOOMERanG	Planck
Sky fraction	5%	100%
Resolution	20'	5'
Frequencies	1	9
Components	1	3
Pixels	$O(10^5)$	$O(10^9)$
Operations	$O(10^{15})$	$O(10^{27})$

- Science goals drive us to observe more sky, at higher resolution, at more frequencies, in temperature and polarization.
- Exact methods are no longer computationally tractable.

# Approximate CMB Analysis

- Map-making
  - No explicit noise covariance calculation possible
  - Use PCG instead:  $(P^T N^{-1} P) m = P^T N^{-1} d$
- Power-spectrum estimation
  - No explicit data covariance matrix available
  - Use pseudo-spectral methods instead:
    - Take spherical harmonic transform of map, simply ignoring inhomogeneous coverage of incomplete sky!
    - Use Monte Carlo methods to estimate uncertainties and remove bias.
- Dominant cost is synthesizing & mapping time-domain data for Monte Carlo realizations:  $O(\mathcal{N}_{mc}\mathcal{N}_t)$

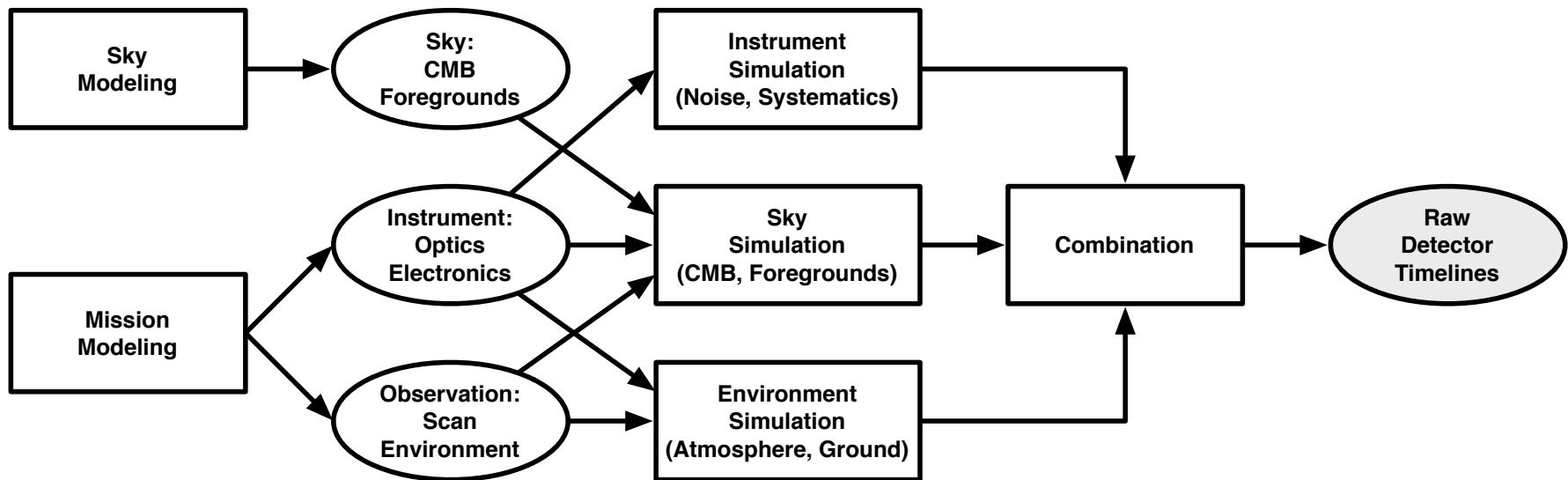
# Synthesis & Mapping: Algorithms

Given the instrument noise statistics & beams, a scanning strategy, and a sky:

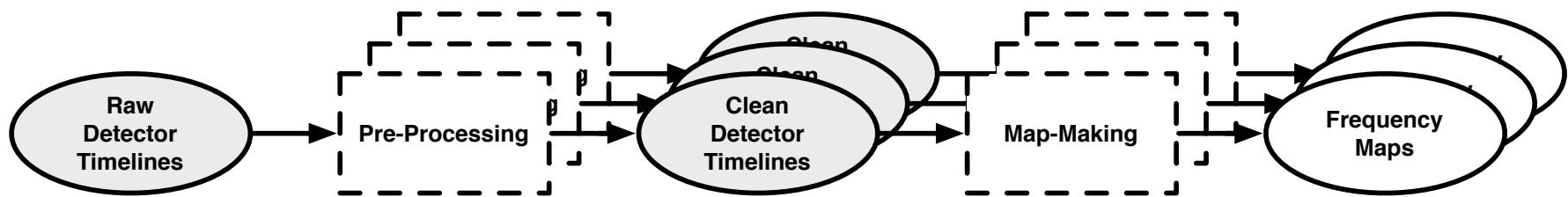
- 1) SYNTHESIS:  $d_t = n_t + s_t = n_t + P_{tp} s_p$ 
  - A realization of the piecewise stationary noise time-stream:
    - Pseudo-random number generation & FFT
  - A signal time-stream scanned & from the beam-convolved sky:
    - SHT
- 2) MAPPING:  $(P^T N^{-1} P) d_p = P^T N^{-1} d_t \quad (A x = b)$ 
  - Build the RHS
    - FFT & sparse matrix-vector multiply
  - Solve for the map
    - PCG over FFT & sparse matrix-vector multiply

# A Synthesis & Reduction Framework

- Synthesis



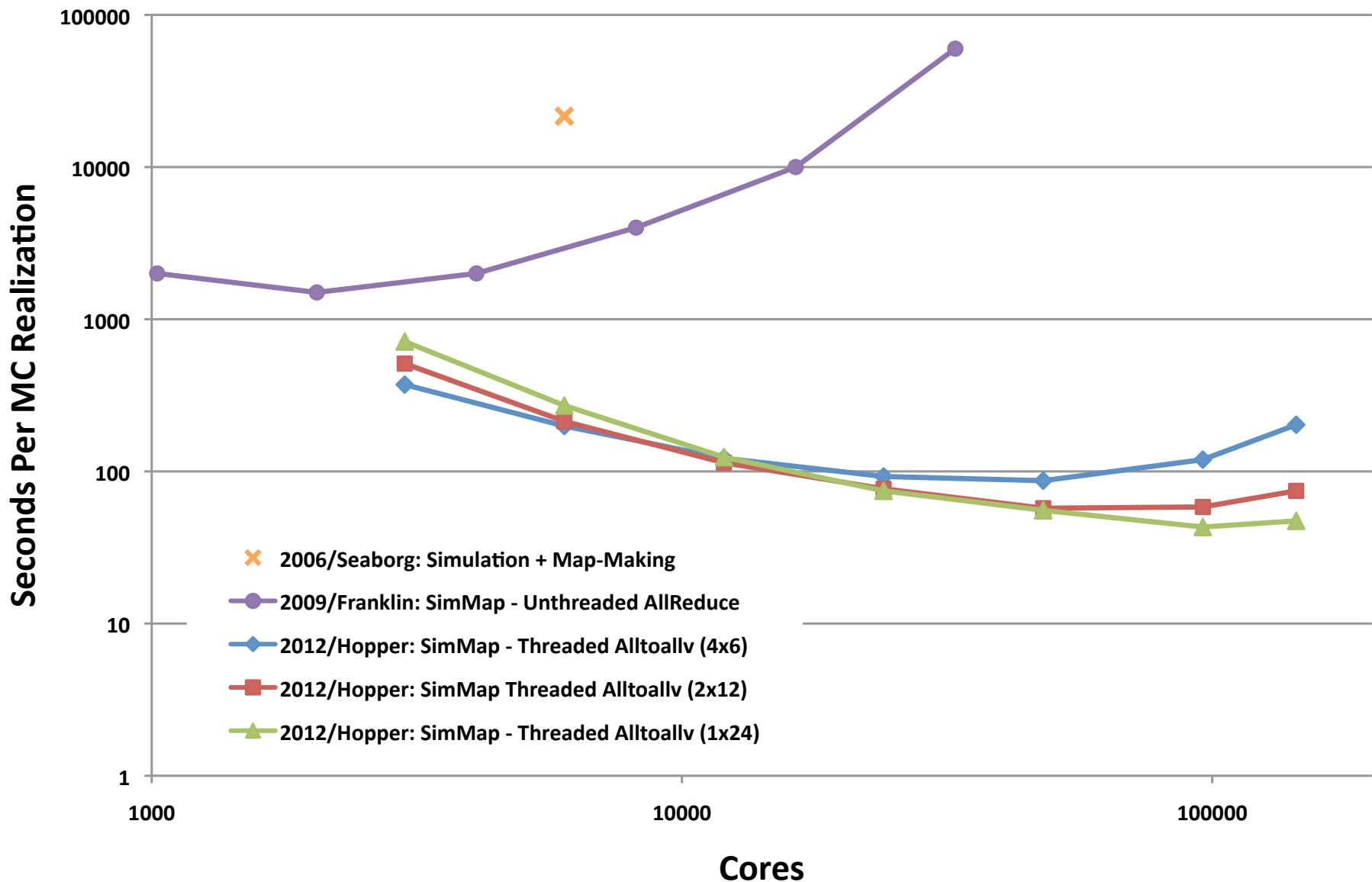
- Reduction



# Synthesis & Mapping: Implementation

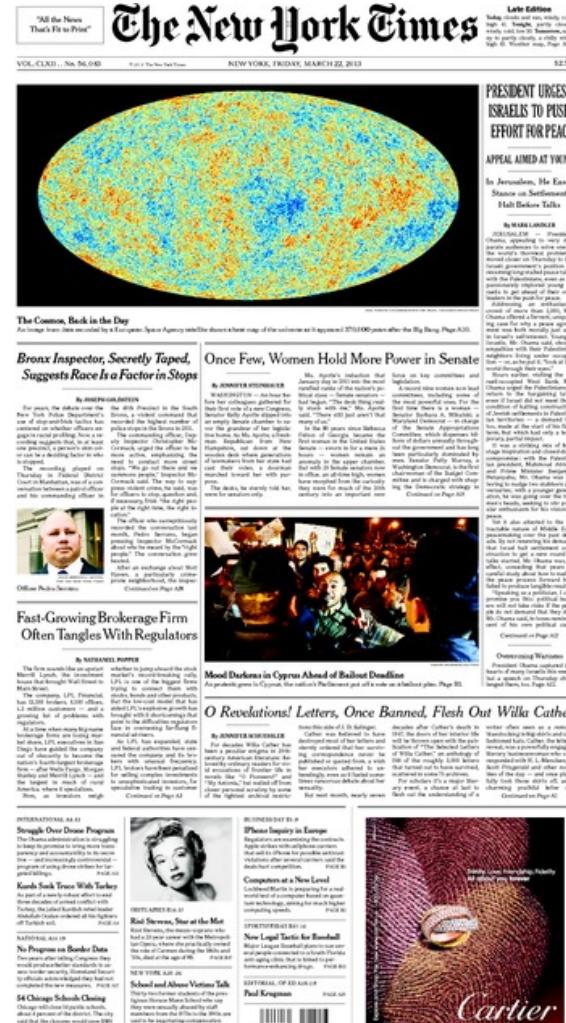
- Linear algorithms reduce calculation costs ...  
... but I/O & communication costs become more significant
- Input/Output
  - On-the-fly synthesis removes redundant write/read
  - Caching common data improves Monte Carlo efficiency
- Communication
  - Hybridization reduces number of MPI tasks
  - Point-to-point communication removes redundancy of all-to-all

# Implementation/Architecture Evolution



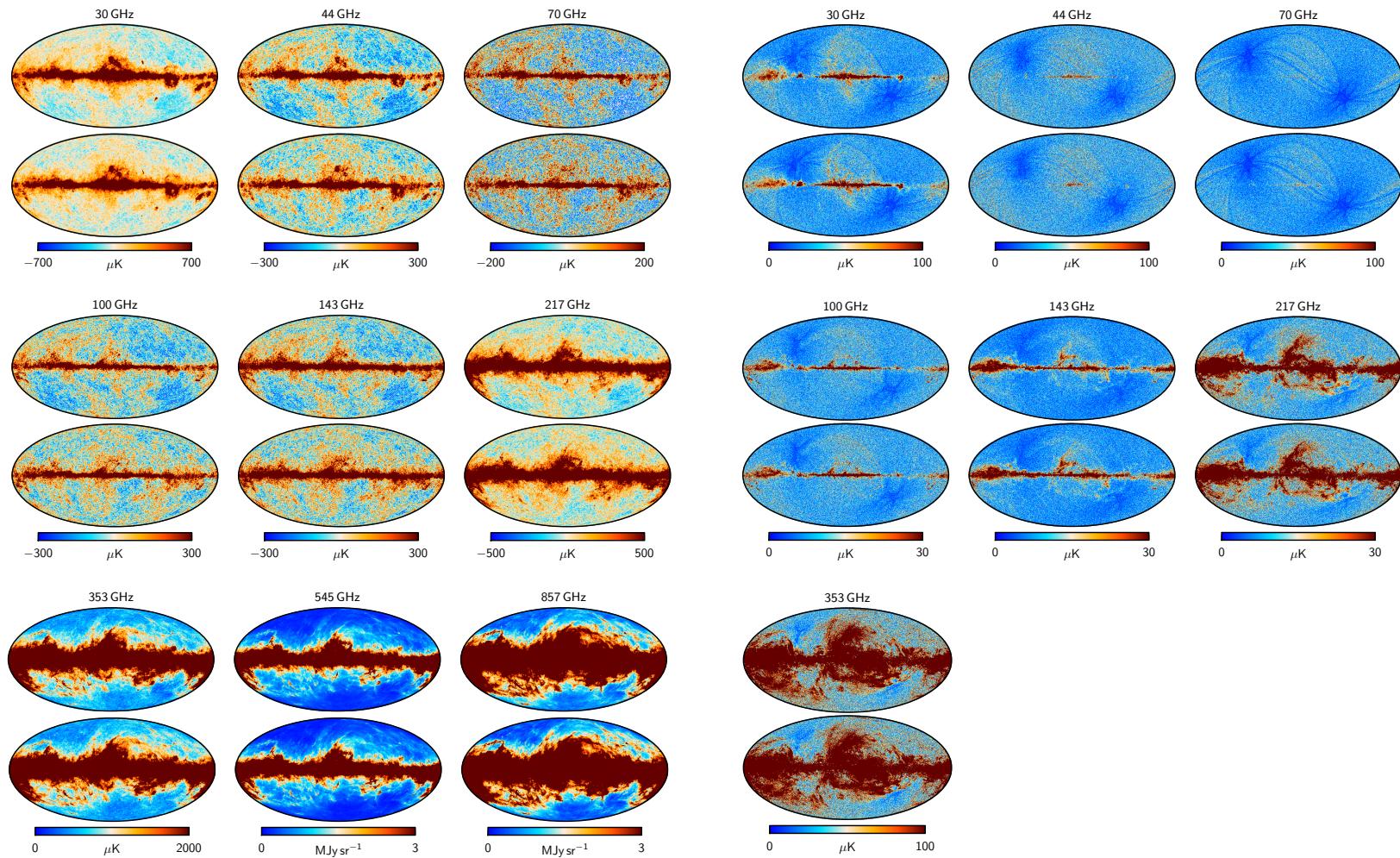
# Results: Full Focal Plane 6 (2013)

- Synthetic data including
  - CMB, foregrounds, detector noise
  - Detailed instrument model
- Fiducial realization for validation and verification of analysis algorithms and implementations.
- $10^3$  Monte Carlo realizations for uncertainty quantification and de-biasing.
- Unanticipated multiplicity of maps
  - 1000 different data cuts per realization!
  - New challenge to on-the-fly simulation.



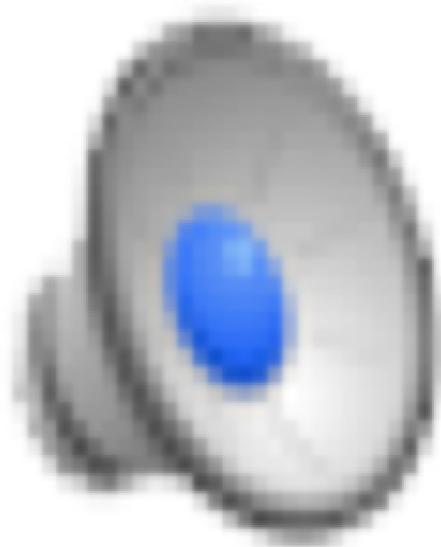
# Results: Full Focal Plane 8 (2015)

- Fiducial realization in temperature and polarization



# Results: Planck Full Focal Plane 8

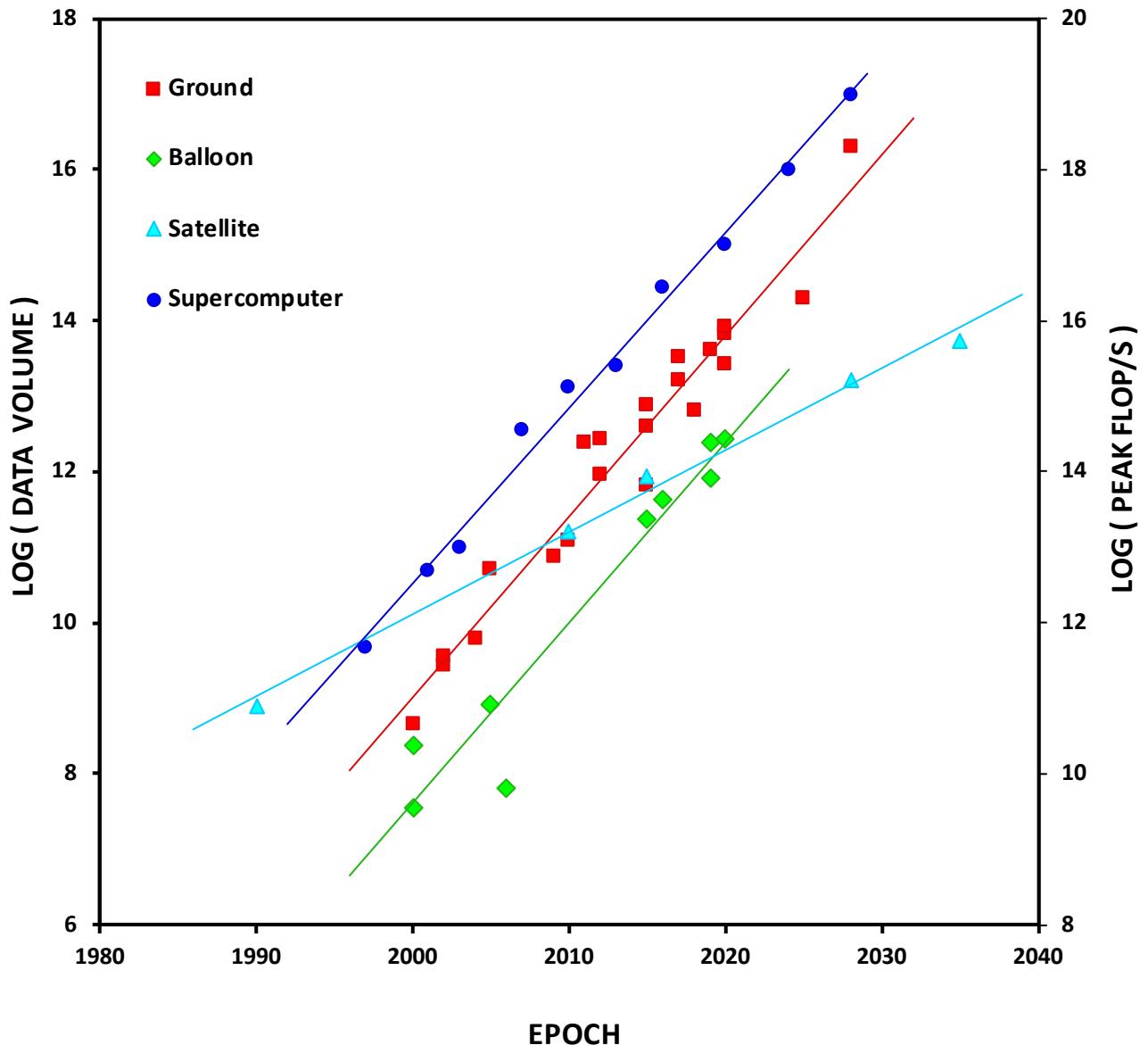
- $10^4$  Monte Carlo realizations reduced to  $10^6$  maps
  - multiple maps made per simulation



# Exponential CMB Data Growth

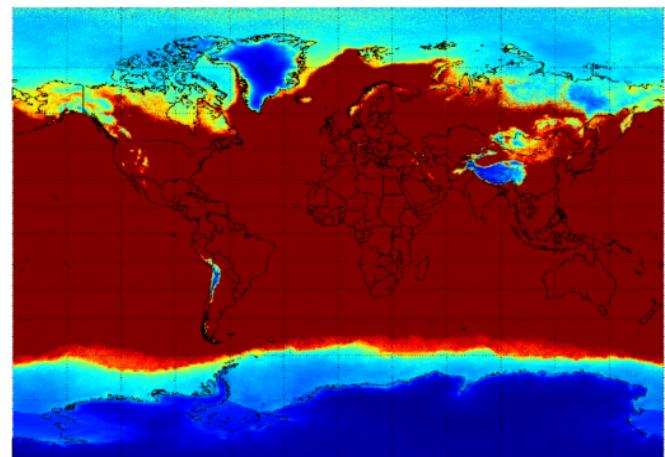
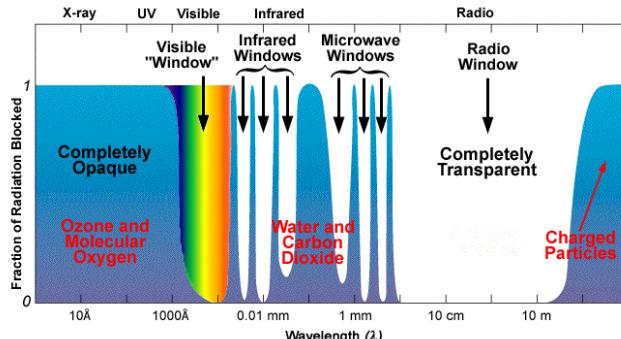
Ever fainter signals require ever larger datasets.

CMB datasets track Moore's Law, requiring us to stay on the leading edge of HPC.



# Case 3: CMB-S4 (2027)

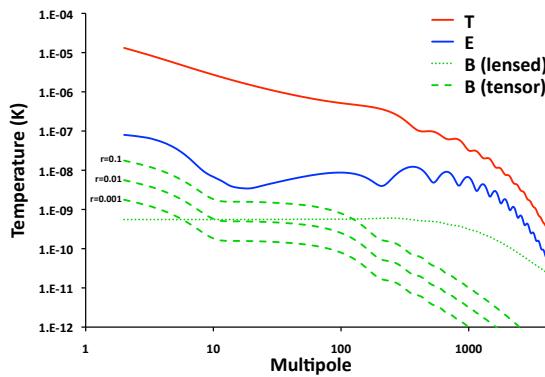
- The ultimate ground-based experiment, observing large and small scales from multiple high, dry, sites
- O(500,000) detectors observing 70% of the sky for 5-10 years through 3 microwave atmospheric windows.



# Data Challenges: Planck to CMB-S4

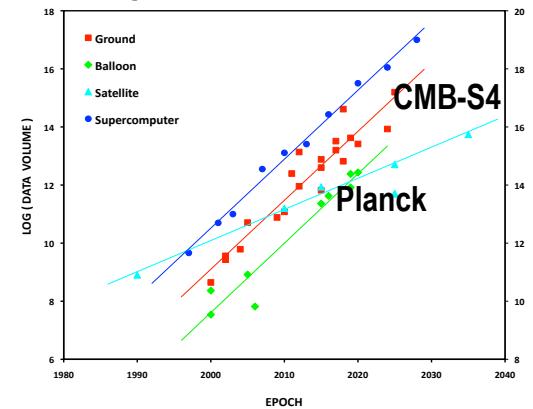
## PRECISION – ALGORITHMS

- 1000x systematics sources
  - Atmosphere
  - Ground pickup
  - Polarization modulator
  - Cross-correlated noise
  - Foregrounds
- 100x lower systematics threshold

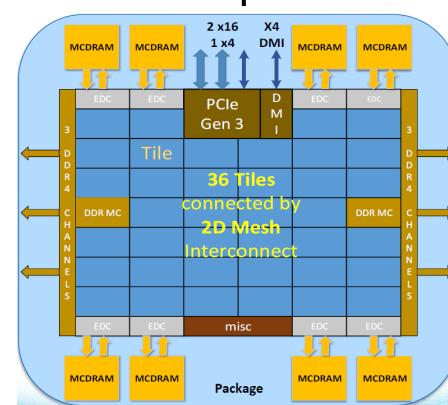


## PERFORMANCE – IMPLEMENTATIONS

- 1000x larger data volume



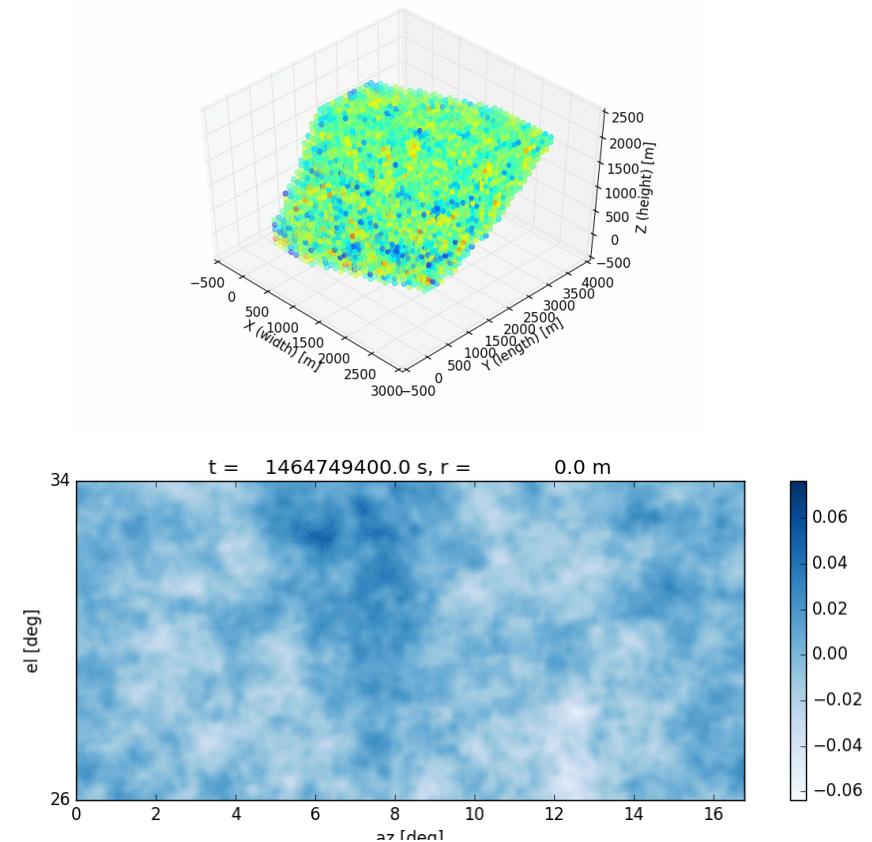
- 100x fewer watts per FLOP



Requires a  $10^{10}$  fold improvement in our computational efficiency!

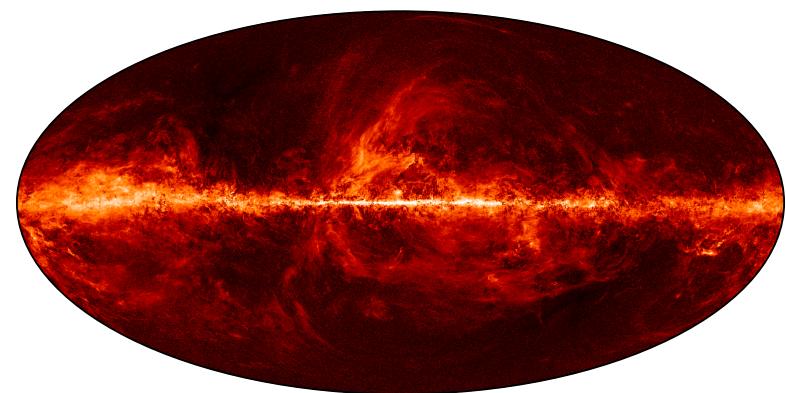
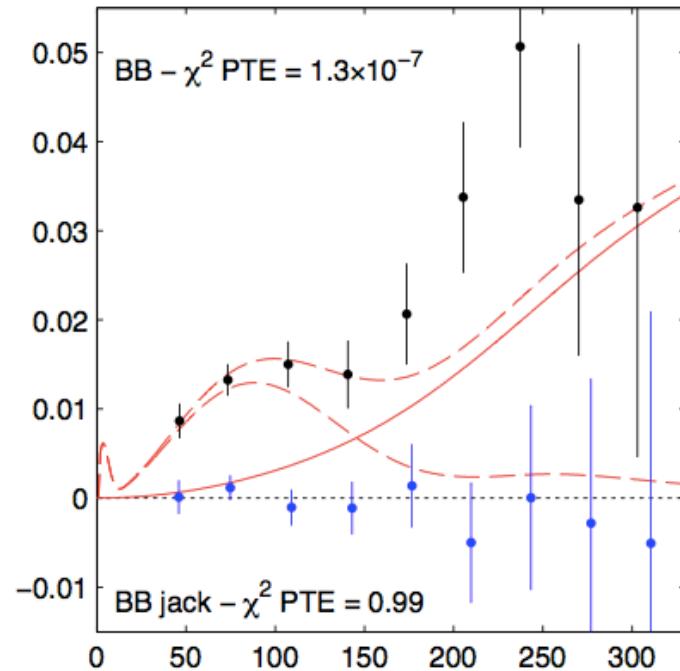
# Example: Atmosphere Simulation

- From the ground, atmosphere is a large, correlated, time-dependent contaminant.
- To reach CMB-S4 sensitivity we must validate and verify mitigation algorithms.
- 3-step algorithm:
  - Calculate bounding box based on scan & wind speed.
  - Generate atmosphere realization based on 2- & 3-D turbulent Kolmogorov spectra.
  - Perform line integral through box for each sample for each detector.



# Example: Residual Systematics

- In March 2014 the BICEP team announces a detection of the B-mode signal from inflation.
- They had done a spectacular job of controlling their instrumental systematics, but only had observations at one frequency.
- Using Planck's 9-frequency coverage we were able to show that their *tiny* signal was actually due to spinning dust grains in the Galaxy.

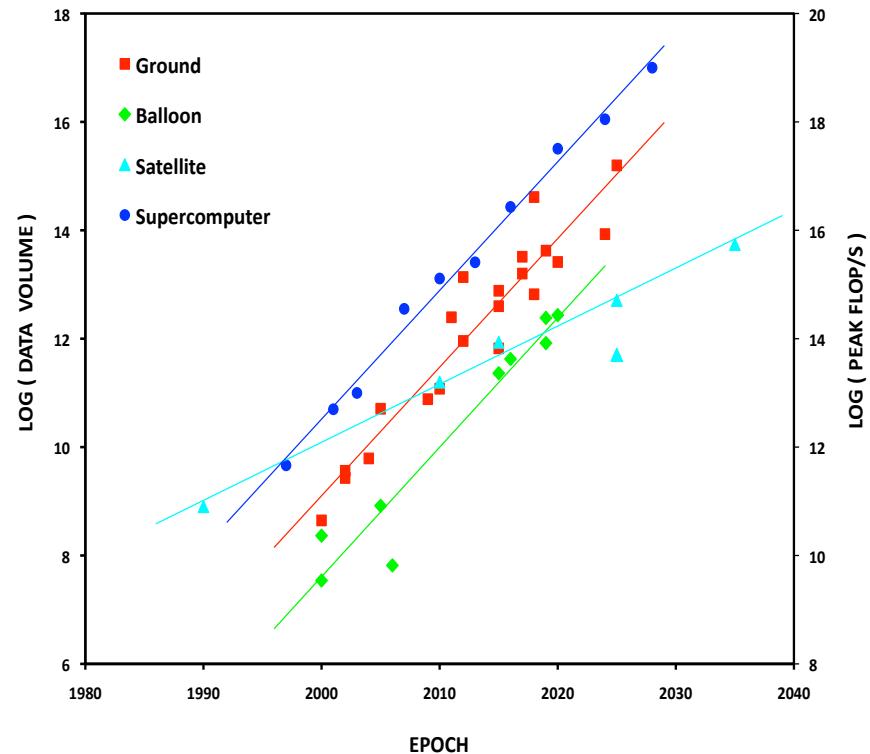


# Example: Data Volume

**Satellite vs  
suborbital  
experiments**

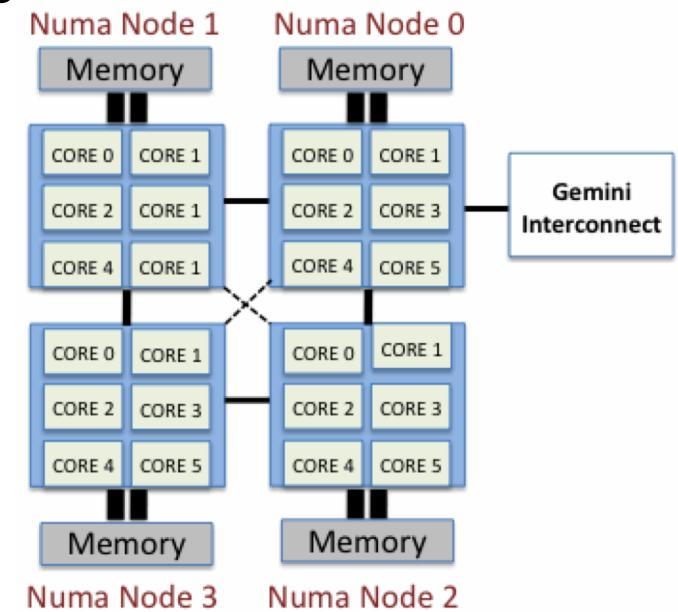
PRO	CON
Environment	Cost
Scanning strategy	Weight/size limits
Hardware quality	Inaccessibility

- We now add the computational tractability of a smaller data volume to the PRO column
  - More precise simulations
  - Larger MC realization sets
- Both clearly seen in Planck compared with Stage 2/3 expts.

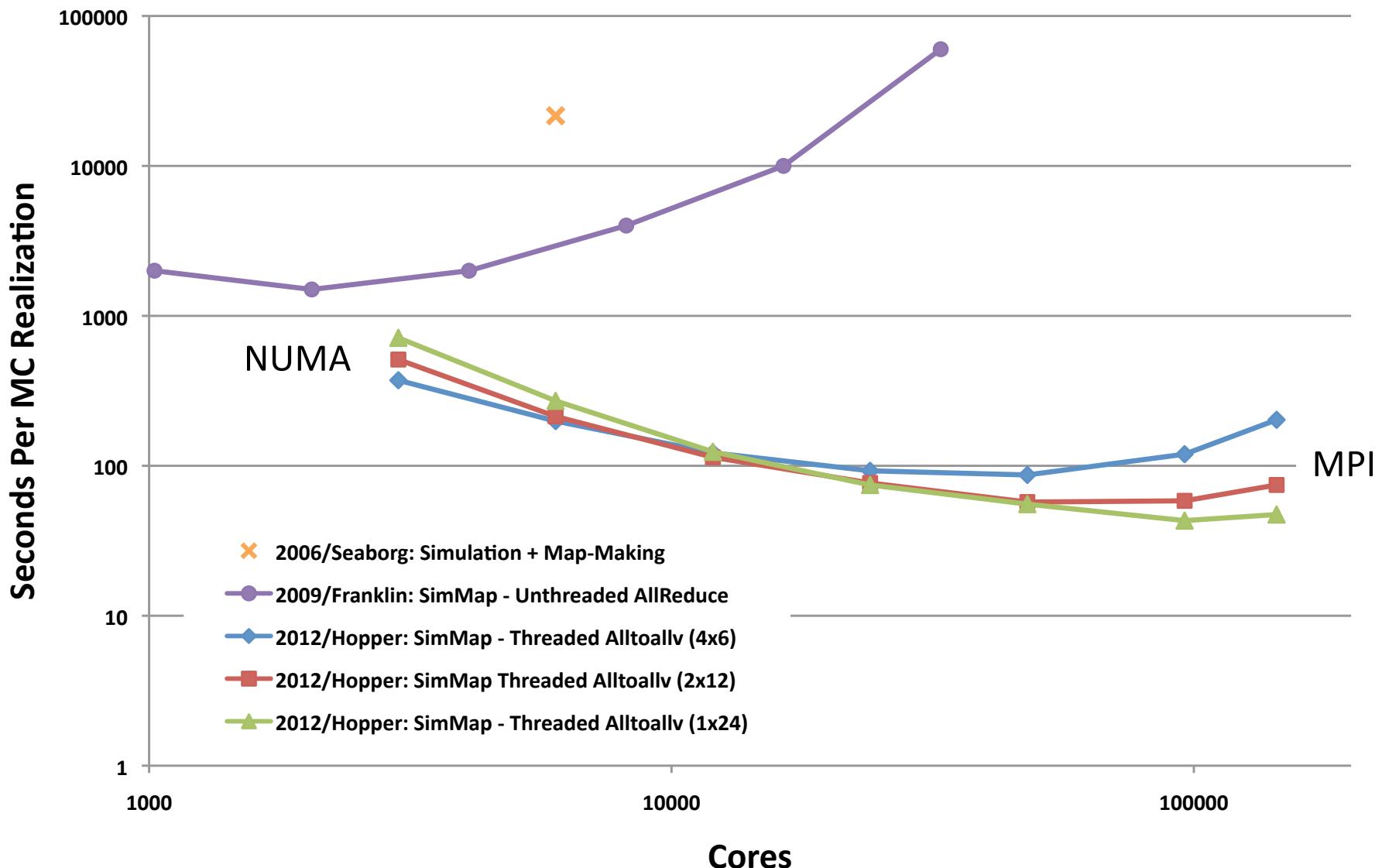


# Example: HPC Architecture

- Clock speed is no longer able to maintain Moore's Law.
- Many-core and GPU are two major approaches.
- Both of these will require
  - significant code development
  - performance experiments & auto-tuning
- Eg. NERSC's Cray XE6 system *Hopper*
  - 6384 nodes
  - 2 sockets per node
  - 2 NUMA nodes per socket
  - 6 cores per NUMA node
- What is the best way to run hybrid code on such a system?

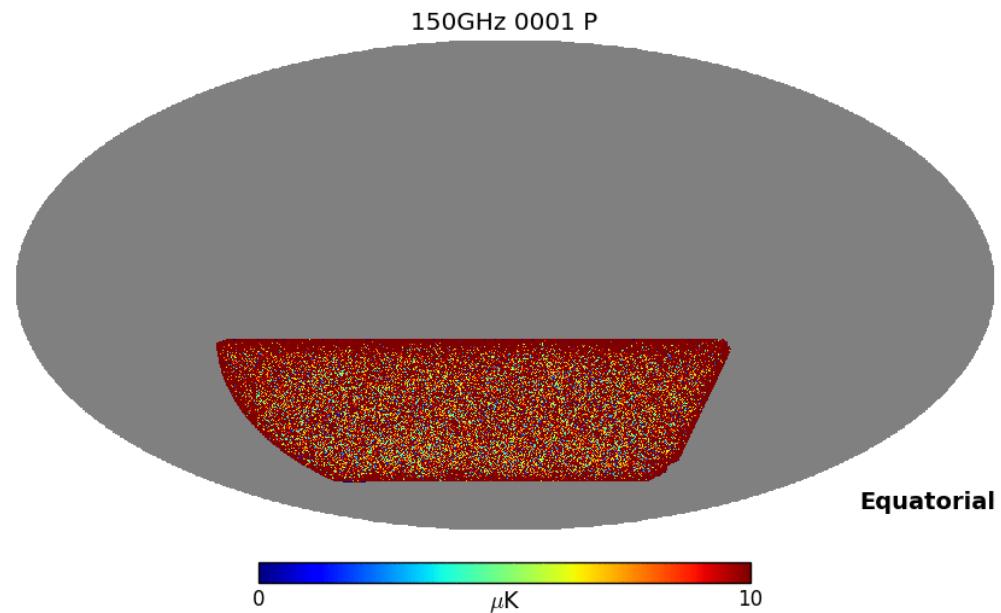
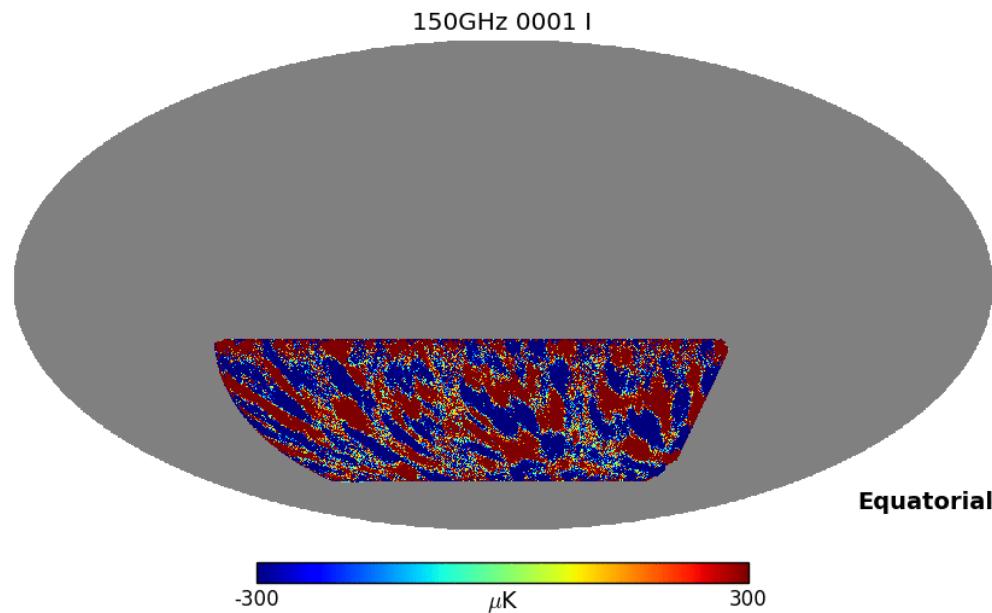


# Configuration With Concurrency



# Current State Of The Art

- Used all of NERSC's *Cori* supercomputer to simulate 50,000 detectors over 7 frequencies observing a 20% sky patch for 1 year
  - 30 trillion samples (35x Planck mission, 1/10th of CMB-S4)
  - atmosphere, instrument noise & sky signal
- Eg. cumulative daily temperature & polarization maps at 150GHz:



# Conclusions

- The Cosmic Microwave Background radiation provides a unique probe of the entire history of the Universe.
- Our quest for fainter and fainter signals requires
  - bigger and bigger data volumes, and
  - tighter and tighter control of systematics.
- Exponential data growth and increasingly complex analyses compels us to stay on the leading edge of high performance computing.
- Our analysis methods, algorithms and implementations necessarily evolve with both the data volumes and HPC architectures.
- CMB-S4 and power-constrained HPC pose the most challenging data/architecture combination we have yet faced.