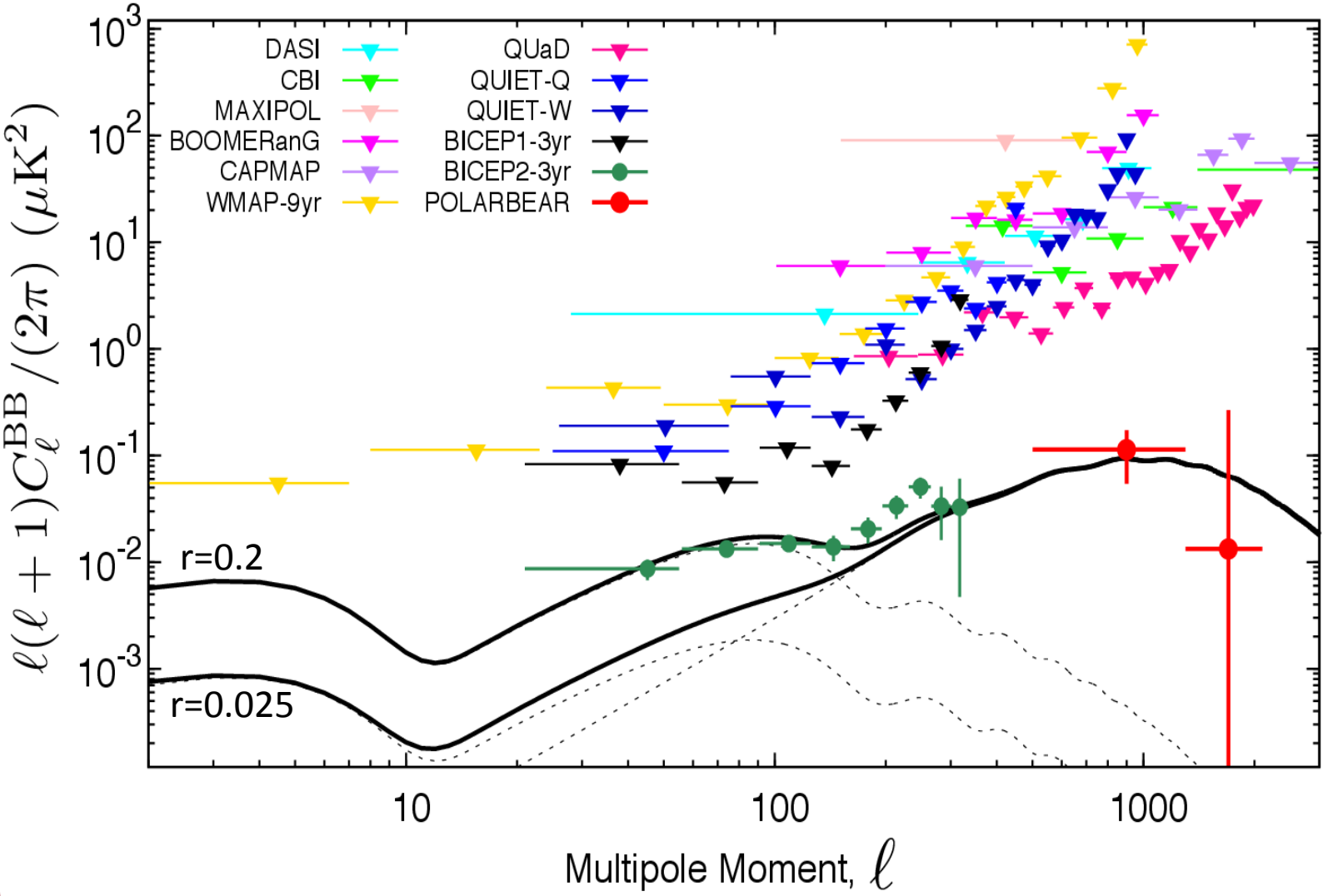
Time Domain Computing: Cosmic Microwave Background Data Analysis

Julian Borrill
Computational Cosmology Center, Berkeley Lab &
Space Sciences Laboratory, UC Berkeley
with
Josquin Errard, Reijo Keskitalo & Theodore Kisner
Planck, POLARBEAR & EBEX

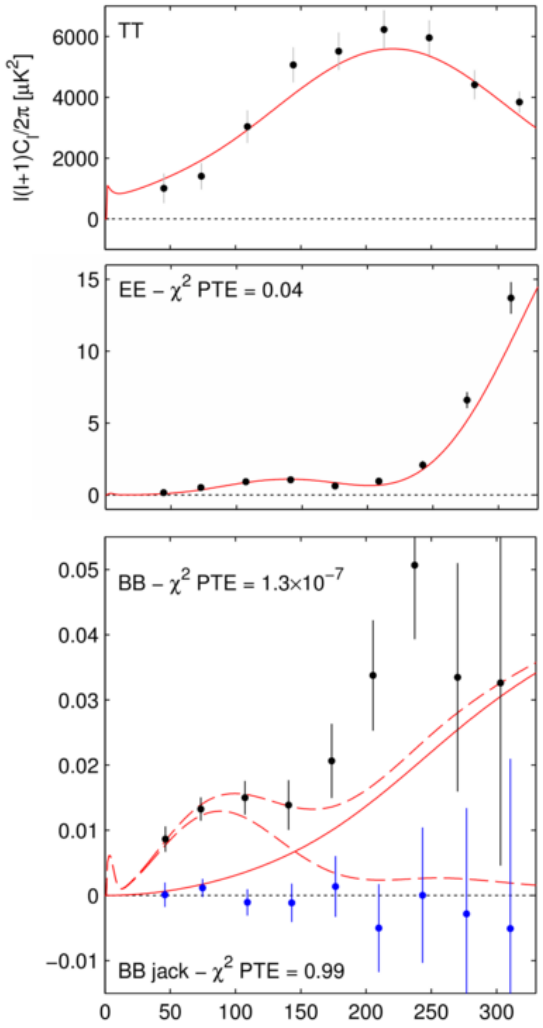
A Great Week For CMB Science!

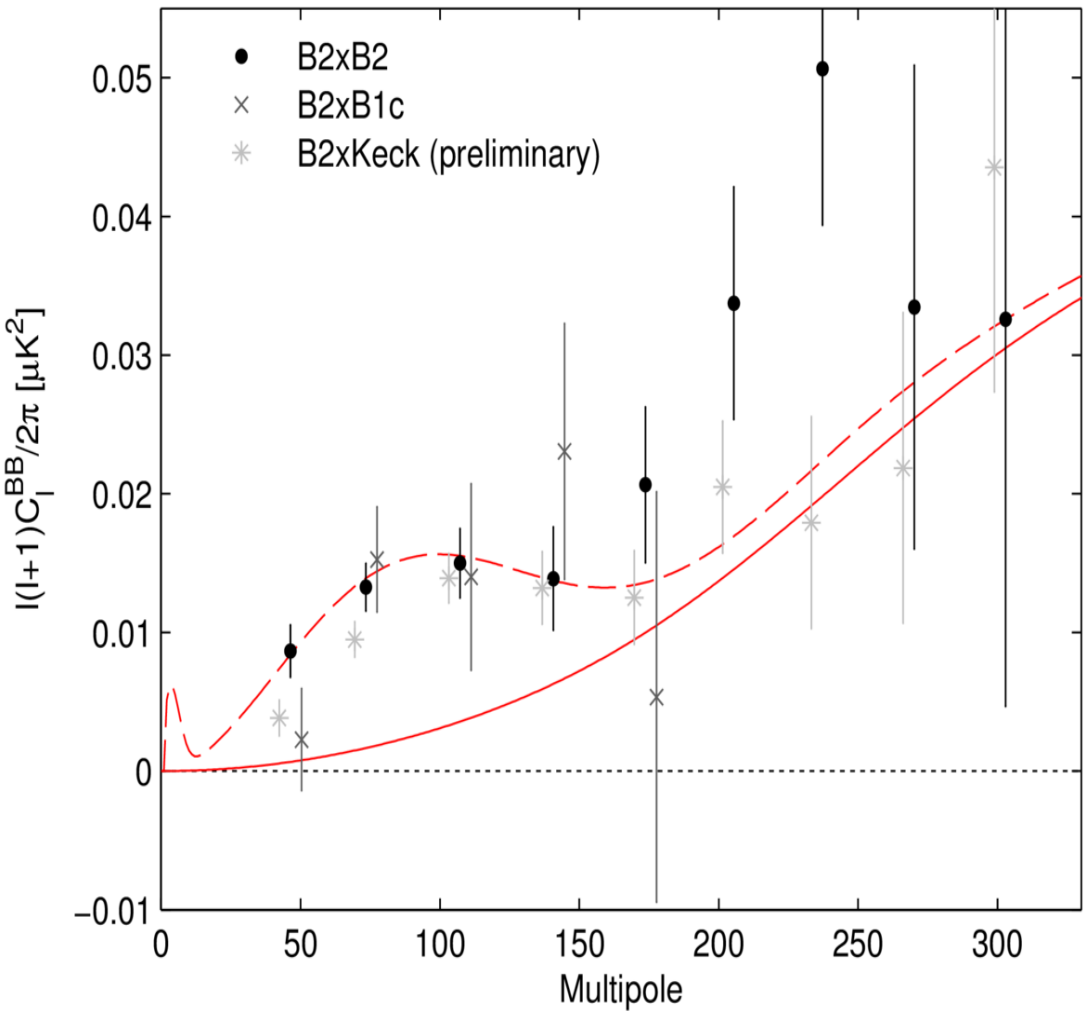


BERKELEY LAB

C3 COMPUTATIONAL COSMOLOGY CENTER

… But There's Still Much To Do





The CMB Data Challenge

Example Experiment	Start Date	Observations	Pixels
COBE	1989	10^9	10^4
BOOMERanG	2000	10^9	10^6
WMAP	2001	10^{10}	10^7
Planck	2009	10^{12}	10^9
PolarBear	2012	10^{13}	10^6
Simons Array	2016	10^{14}	10^7
CMB-S4	2020+	10^{15}	10^{10}

- 1000x increase in data volume every 15 years – Moore's Law!
  - Need linear analysis algorithms & cutting-edge HPC systems.

BERKELEY LAB

C3 COMPUTATIONAL COSMOLOGY CENTER

Forthcoming/Proposed Experiments

- Two more generations of ground-based experiments
  - Atacama Desert & South Pole.
  - Up to 10,000 detectors on multiple telescopes.
  - Eg. PolarBear-2/Simons Array, SPTpol-3G, Keck.
  - O(100x) Planck data volume.
- "Ultimate" CMB polarization experiment
  - CMB-S4: US ground-based with multiple telescopes at multiple locations, observing 50% of the sky with up to 500,000 detectors.
  - CoRE: European space mission observing 100% of the sky with up to 100,000 detectors.
  - O(1000x) Planck data volume.

Computational Challenge - Technical

- Tiny signals
  - Massive data to achieve necessary S/N
    - Growing with Moore's Law for last & next 15 years
    - Approximate methods require Monte Carlos for UQ.
  - Exquisite control of systematics
    - Everything aliases as polarization
    - Instrumental – noise, beams, bandpasses
    - Environmental – atmosphere, foregrounds
    - Mathematical – aliasing from imperfect basis
  - T/E/B hierarchy
    - Danger of leakage
    - Systematics threshold drops

Computational Challenge - Sociological

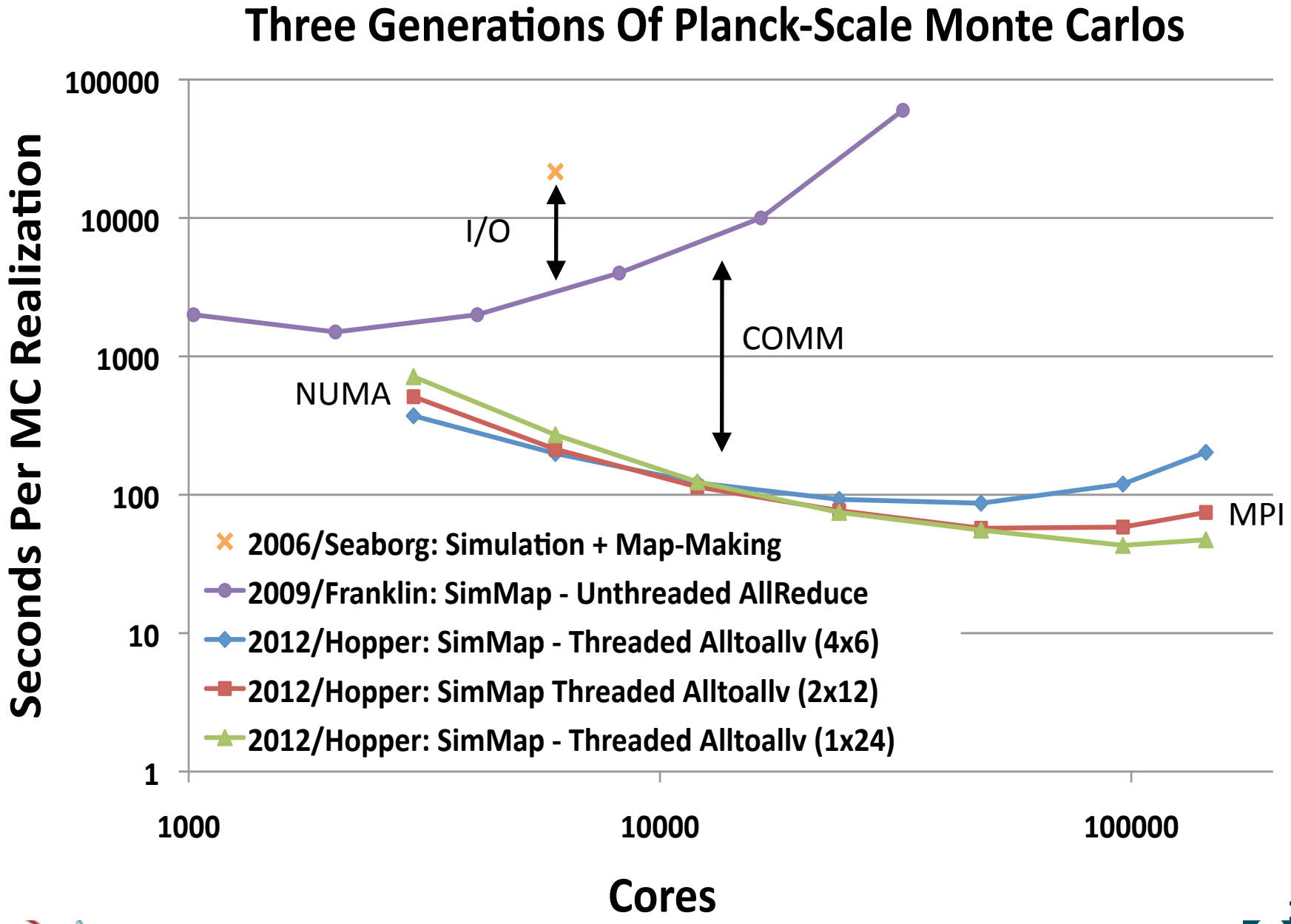
- Data analysis is still the poor relation
  - Never included in operating budgets for suborbital experiments!
- Computational cost claims (ab)used to justify crude "filter-and-fit" approaches:
  - "… it is prohibitively expensive to run a large set of end-to-end simulations that would capture all aspects of the map-making pipeline, and the noise characteristics and correlations in the actual data set." *ACT 3 Year*
  - "it may not be computationally feasible to construct simple timestream templates for some potential systematics. Therefore … we must … estimate the residual contamination and either subtract it or show it to be negligible." *BICEP2*

Computational Resources

- Major NERSC allocations for CMB data analysis for the last 15 years
  - Suborbital: O(100) users from O(10) experiments at any time
  - Planck: unique multi-year allocation + dedicated hardware
- Reliable roadmap to maintain Moore's Law growth
  - Allows us to work to future capability.
  - Planck allocation: 100,000 to 100,000,000 CPU-hrs in 15 years
- Data management
  - Project spaces, pseudo-users, etc
  - Community resources
    - Plan to keep all Planck data (including simulations) spinning for users.

Time-Ordered Astrophysics Scalable Tools

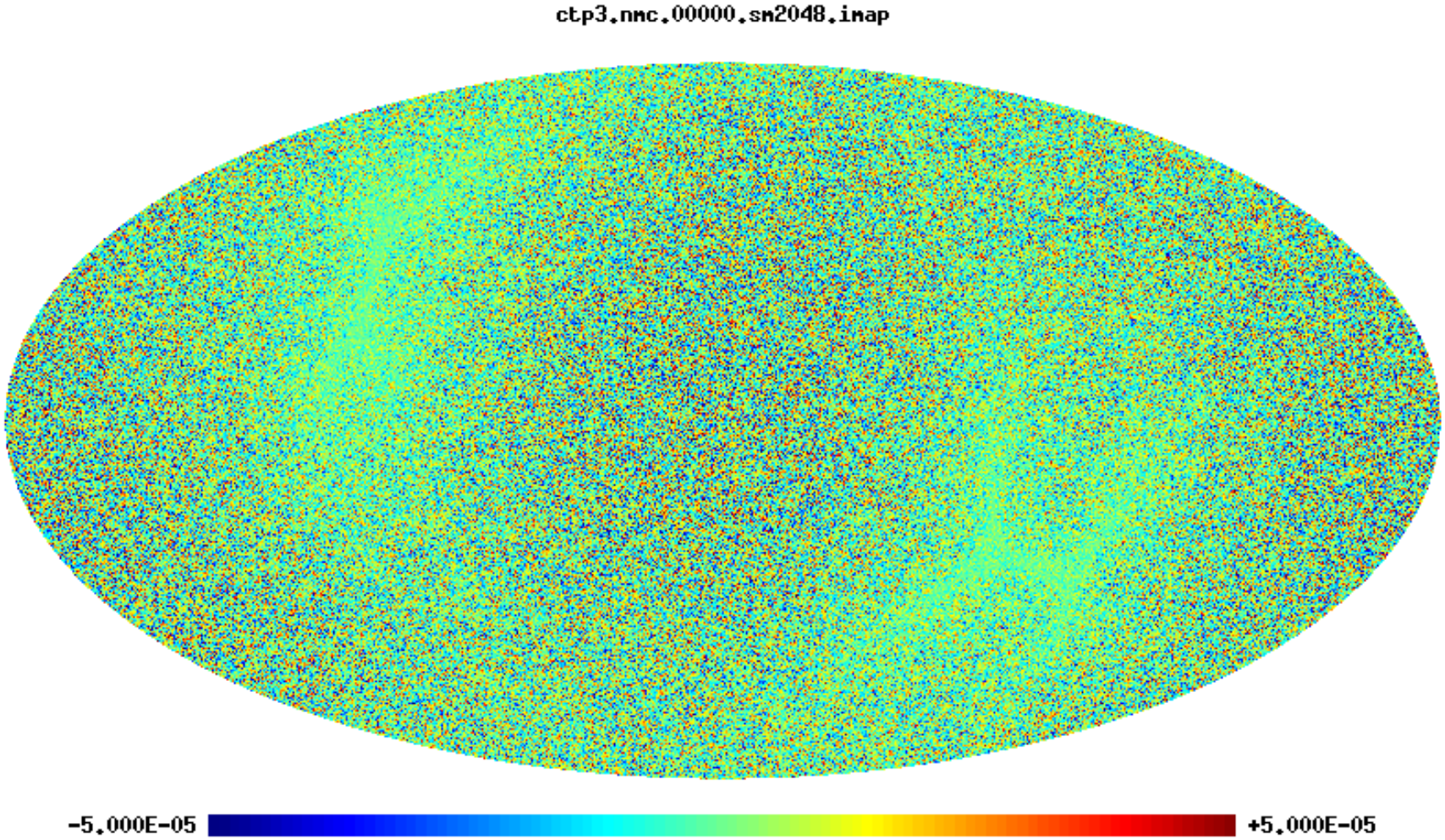
- TOAST addresses the computational challenge of present and future CMB data analysis head-on.
  - If data grow with Moore's Law, analysis capabilities must too.
    - Exploit massive parallelism.
    - Break the data movement bottlenecks.
    - Expect architecture-dependence.
    - Provide data abstraction.
- Used to generate largest CMB Monte Carlo simulation set ever deployed in support of the first Planck results.
  - 1,000 nominal mission realizations reduced to 250,000 maps.
  - Planning 10x increases for 2014/15 releases.



BERKELEY LAB

C3 COMPUTATIONAL COSMOLOGY CENTER

Planck Noise



Conclusions

- CMB data analysis is computationally challenging …
  - Dominated by Monte Carlo time-domain data simulation/reduction
- … but tractable if we stay on the leading edge of HPC
  - Efficient data movement is still the key
  - Heterogeneous power-constrained systems will make this harder
- Community outreach/education/shaming required
  - Role for HiPACC-supported workshop?

Either you get to cite the computational cost of your analysis

Or you get to write your analysis pipeline in Matlab

BUT NOT BOTH!