HEP HPC Requirements Workshop: Cosmic Microwave Background Data Simulation & Analysis

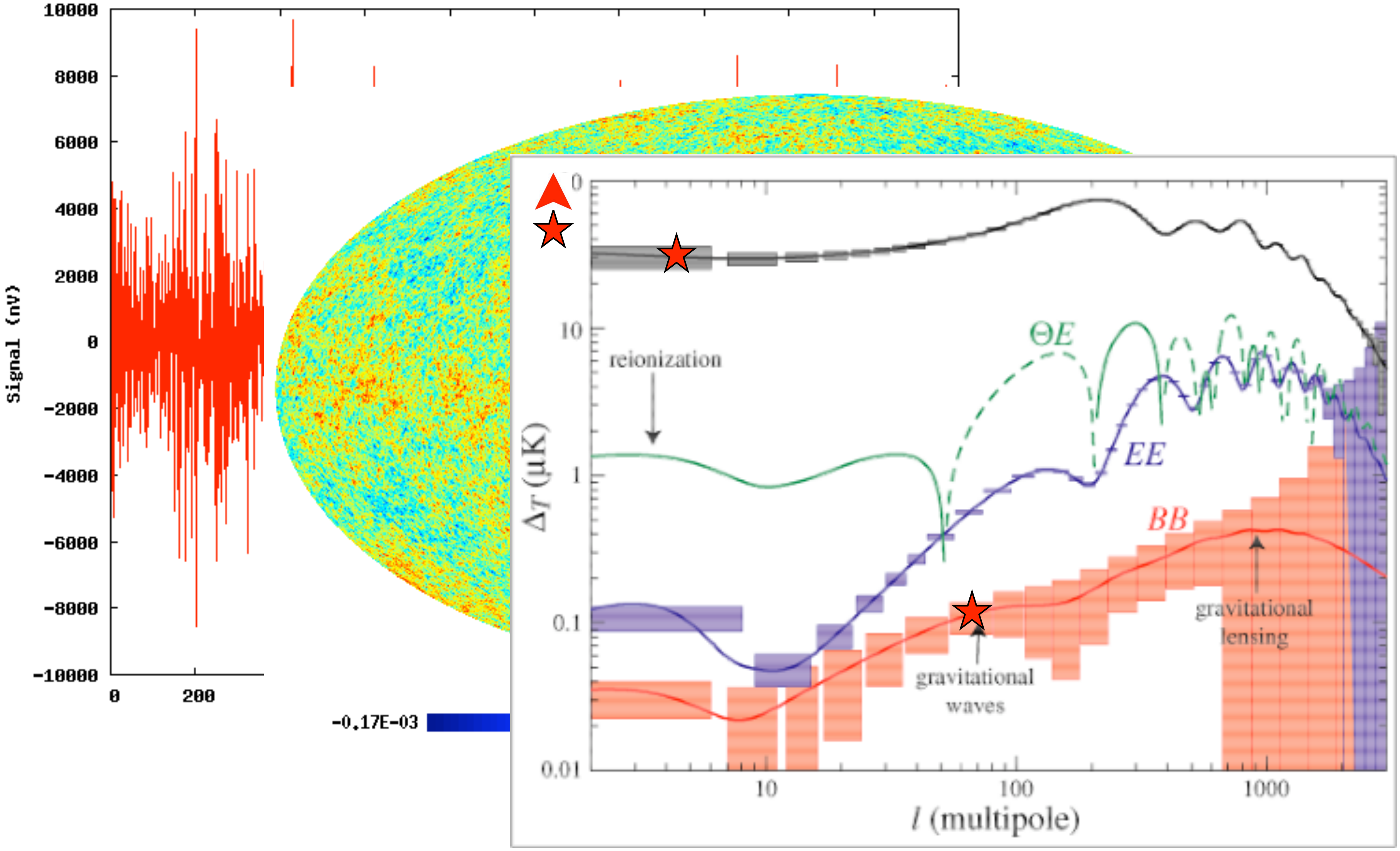
Julian Borrill
Computational Cosmology Center, Berkeley Lab
& Space Sciences Laboratory, UC Berkeley

Overview

- 3 NERSC repositories
  - mp107: O(10) NSF/NASA/DOE/international suborbital experiments
  - planck: current ESA/NASA satellite mission (with DOE IAA)
  - cmbpol: proposed NASA satellite mission
- O(100) users, 4 projects (Planck, EBEX, PolarBear, QUIET), 2 modules (cmb, planck)
- LBNL/UC Berkeley, JPL/Caltech, U Chicago, U Minnesota, etc
- Paris, Rome, Trieste, Helsinki, London, Cambridge, Cardiff, Munich, etc

- Simulation & analysis of CMB data
  - Algorithm validation & verification
  - Implementation efficiency & scaling
  - Mission design & science exploitation
    - Snapshot of the Universe 380,000 years after Big Bang
    - Fundamental parameters of cosmology
      - e.g. Planck results assumed by all Dark Energy missions
    - Highest energy physics
      - Big Bang as ultimate particle accelerator

CMB Data



10000
8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
-10000
Signal (nV)
0
200
-0.17E-03
0
10
1
0.1
0.01
Δ_T (μK)
reionization
ΘE
EE
BB
gravitational waves
gravitational lensing
10
100
1000
l (multipole)

CMB Data Analysis

- In principle very simple
  - Assume Gaussianity 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* $O(N_t)$, 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
  - Approximations => systematic effects from analysis itself.

CMB Data Sets

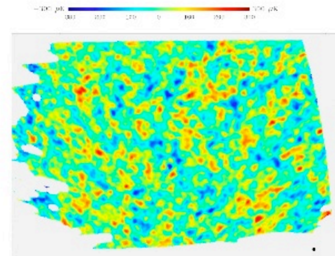
Date	Experiment	Description	Time Samples	Sky Pixels
1990-93	COBE	All-sky, low-res, T	8×10^{8}	3×10^{3}
1998	BOOMERaNG	Cut-sky, mid-res, T	9×10^{8}	3×10^{5}
2001-10	WMAP	All-sky, mid-res, TE	2×10^{11}	6×10^{6}
2009-11	Planck	All-sky, high-res, TE	3×10^{11}	1×10^{8}
2010	EBEX	Cut-sky, high-res, TEB	3×10^{11}	6×10^{5}
2010-12	PolarBeaR	Cut-sky, high-res, TEB	3×10^{13}	1×10^{7}
2011-14	QUIET-II	Cut-sky, high-res, TEB	1×10^{14}	7×10^{5}
~2020	CMBpol	All-sky, high-res, TEB	1×10^{15}	9×10^{8}

Increased resolution & sensitivity needed for evolving science goals requires ever larger data sets to achieve necessary S/N.

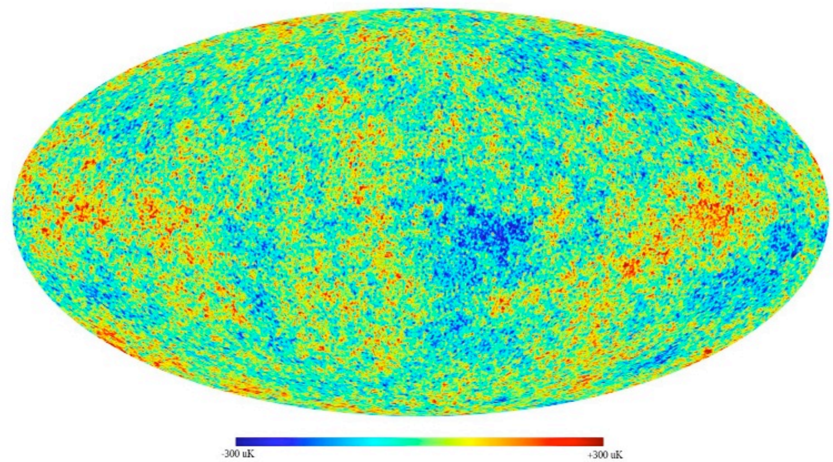
N_t increases as Moore's Law.

Scaling To Date

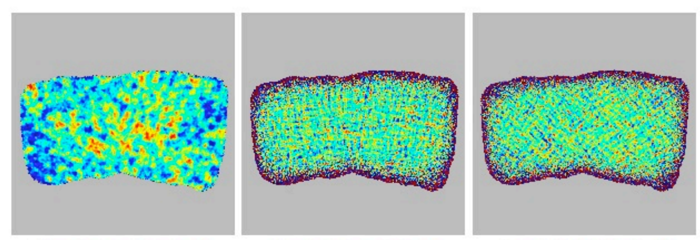
2000: BOOMERanG-98 temperature map (10^8 samples, 10^5 pixels) calculated on 128 Cray T3E processors;



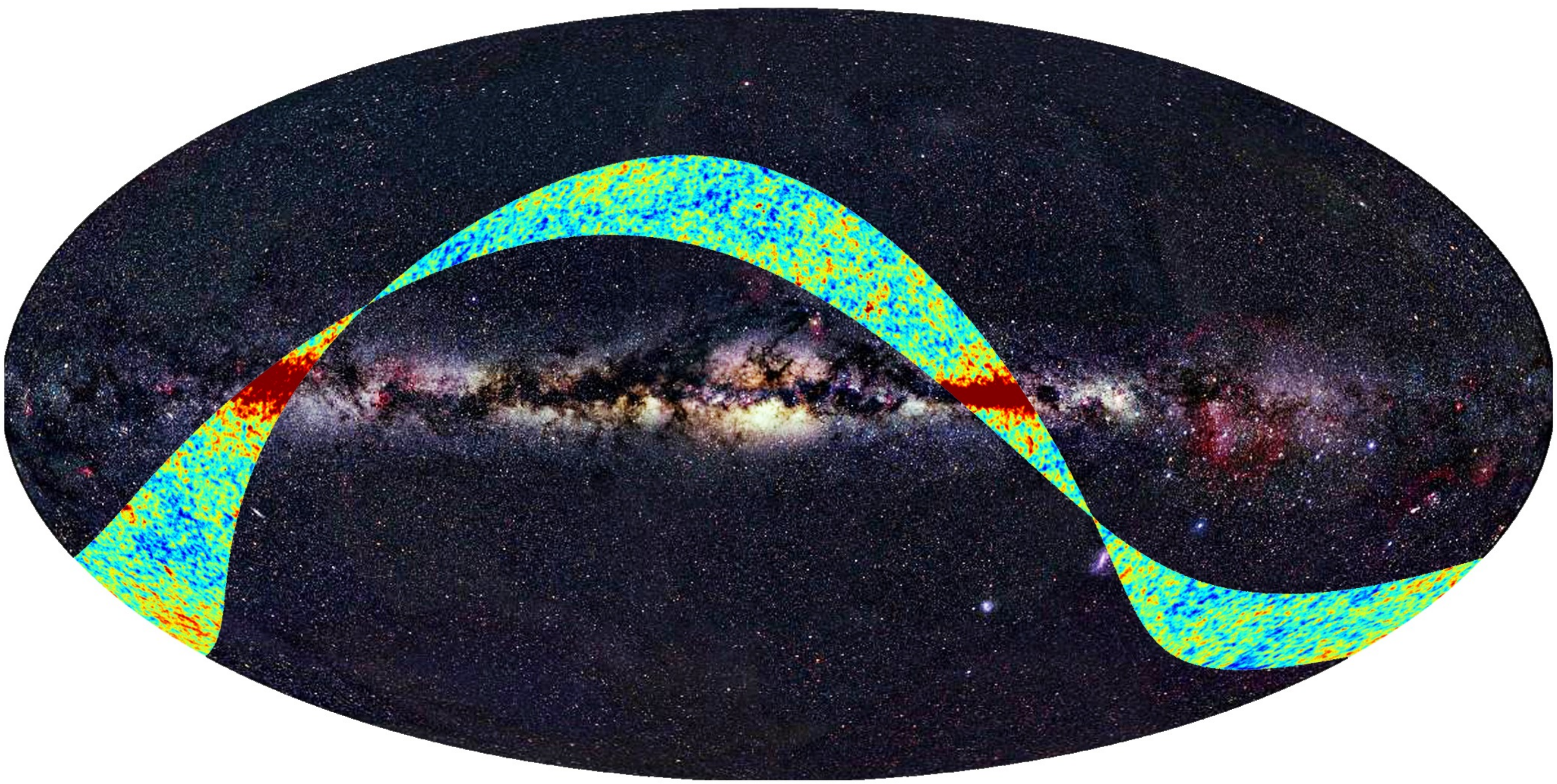
2005: A single-frequency Planck temperature map (10^{10} samples, 10^8 pixels) calculated on 6000 IBM SP3 processors;



2008: EBEX temperature and polarization maps (10^{11} samples, 10^6 pixels) calculated on 15360 Cray XT4 cores.



Planck First Light Survey



Code Profile - I

- Calculations dominated by simulation & mapping of time-ordered data:
  - Iterative analysis cycle: $O(10^2) + 1$ analyses
    - Monte Carlo error analysis: $O(10^2) + O(10^4)$ realizations
      - PCG solver: $O(10^{1-2})$ iterations
        - » FFT/SHT: $O(10^{12-15})$ operations
  - Overall $O(10^{17-21})$ flops (plus prefactor & efficiency).

- Memory dominated by maps:
  - TOD distribution over cores implies pixel distribution
  - Only hold local sub-map on each core
  - Ancillary data growing fast

- Communication dominated by map-reductions:
  - Replace MPI_Allreduce() with TBD (UPC, FastBit, other)

Code Profile - II

- IO dominated by reading TOD (including pointing) & writing maps:
  - M3 data abstraction layer
  - Replace full with compressed pointing, reconstructed on the fly
  - Replace simulate/write => read/map cycle with on-the-fly sim/map

- Storage dominated by TOD:
  - MC map sets significant now but essentially constant
  - Full TOD needs to be spinning/accessible simultaneously
  - CMB rule-of-thumb, need to store 100x TOD & 10000x maps

Current HPC Requirements

- E.g. SimMap 100 realizations of 1-year Planck mission in <1 day wall-clock.

- Franklin & NGF
  - Use destriping map-maker to reduce FLOPs & communication
  - 10,000 cores x 3 hours (noise) + ? (signal)
  - 500-800MB/core
    - Asymmetric beam simulations require 5-10GB/core!
  - Minimal read, 1TB write
    - NGF too slow => use scratch & transfer data post hoc.

- Known bottlenecks:
  1. Flops
  2. Memory
  3. Communication & IO bandwidth
  4. Disk space

- (Partial) software solutions trade memory/bandwidth/disk for flops.

The Next 3-5 Years - Expectations

- Time-ordered data volume increases by at least an order of magnitude:
  - 10x calculations, 10x concurrency
  - Constant memory/core
  - Constant communication volume
  - 10x I volume, ~constant O volume
  - 10x storage

- Planck is special
  - Baseline for sub-orbital experiments ~0.1x

- Anticipated problems at scale
  - Communication & IO don't scale with calculations
  - Delivering data fast enough to exploit capability
    - GPUs, etc
    - Heterogeneous heterogeneity
  - System stability

The Next 3-5 Years - Preparation

- NSF PetaApps project to address scaling
  - Modularize code by HPC component
  - Implement a range of trade-offs between components
  - System-specific (one-time) tuning
  - Analysis-specific (run-time) tuning

- GPU exploration for flops
  - Test-bed systems useful, but where do they lead?
  - Pooling resources, libraries?

Summary

- Recommend:
  - maintaining balance in NERSC systems
    - data delivery is (almost) everything
      - NGF still needs work; what is NERSC committed to?
  - investment in human resources for scaling challenges
  - longer-term allocation commitments to projects

- With 50x resources I could:
  - stop pretending that the resources needed to analyse next-generation suborbital experiments with 10x data will only be 10x Planck
  - start performing simulations for CMBpol to inform its concept/design

- The expanded HPC resources I want:
  - more of everything!

- Additional NERSC services:
  - Enhanced project support; data & job management tools