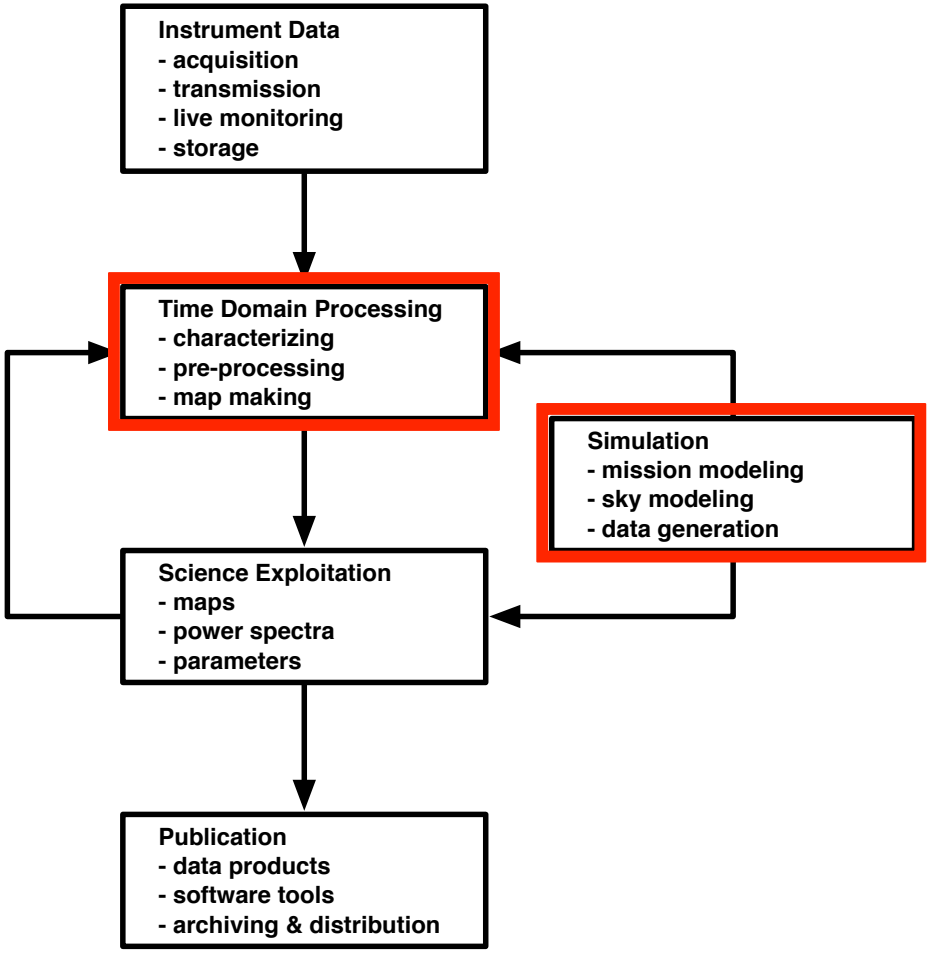
Data Analysis & Infrastructure

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

Data Management Overview

- Subdivide overall data management into 5 areas
  1. Instrument
  2. Time-Domain
  3. Science
  4. Publication
  5. Simulation
- Identify the key tasks & the human and computational resources for each.
- Now in the epoch of formal Data Project: WBS, Manager



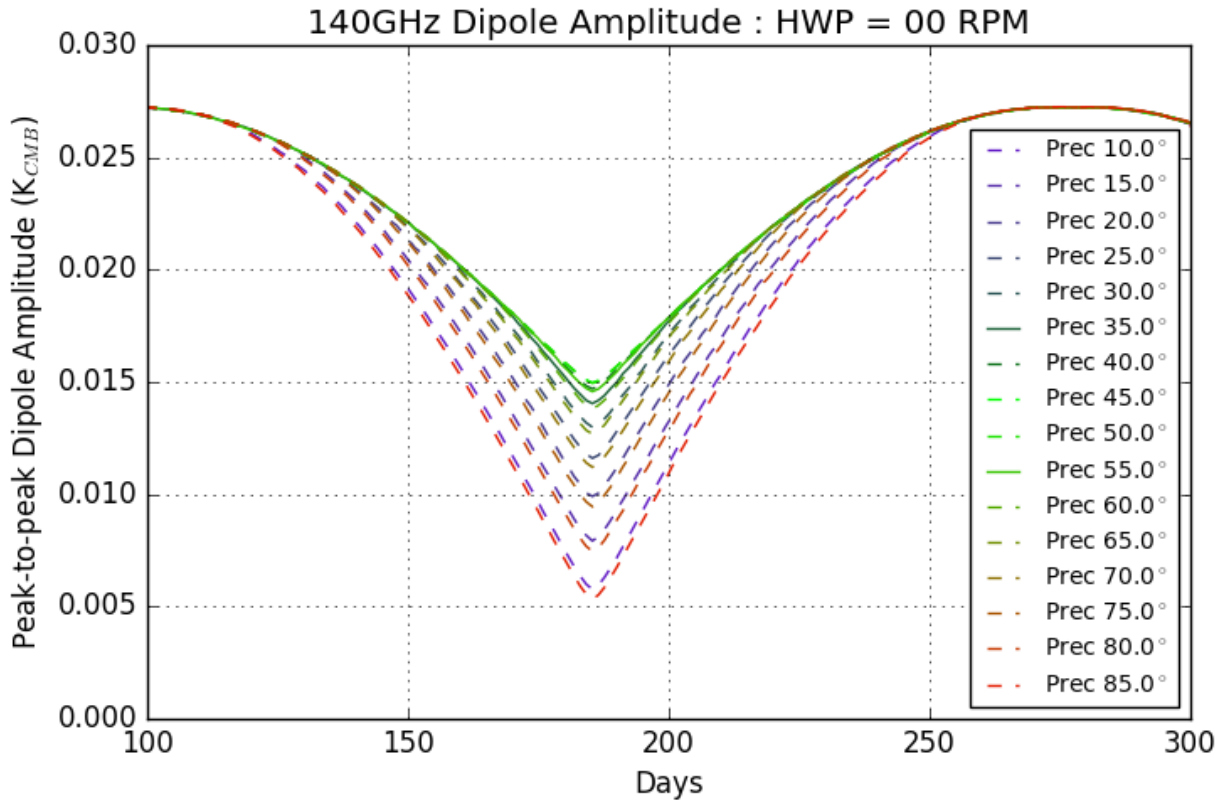
Simulations

Overview

- Given a sky model
  - Scalar, tensor & non-Gaussian CMB
  - Extragalactic foregrounds & lensing
  - Galactic foregrounds
- Given a mission model
  - Instrument
  - Observation
- Apply the mission to the sky to generate a synthetic mission dataset in the time, map, or spectral domain and analyze it
  - Trade cost against realism in selecting domain
  - Define simpler metrics as proxies for final parameter constraints

Example Metric

- Daily peak-to-peak dipole variation as a function of satellite scanning angles (fixed α + β)
  - Defines ability to perform daily dipole calibration



Sky Modeling

- A number of multi-frequency next-generation missions are in design & development
  - Suborbital – Simons Observatory, CCAT-p, CMB-S4
  - Space – LiteBIRD, ~~CORE~~, PICO

- Common sky models
  - Across scales: coherent low- and high-ell science
  - Across missions: comparison and combination
  - Across disciplines: CMB x LSS

Current Status

- With WMAP & Planck, these missions currently span 73 bands
  - Band definitions here

- Modelers are being asked to generate their sky in all bands
  - Tophat & delta bandpassed skies delivered to NERSC.
  - Mostly PySM to date; others for specific subsets.

- Most missions are using NERSC as a central repository
  - /project/projectdirs/{sobs, cmbs4, litebird, core, pico, …}

Wish List

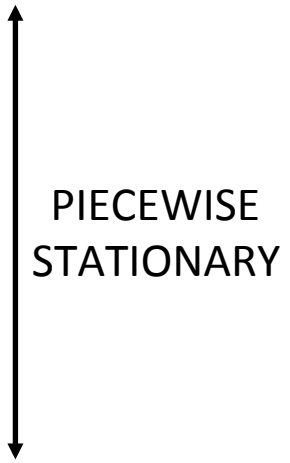
1. **More representative sky models**
   - PICO-sponsored sky modeling workshop: UCSD, 10/30-11/1

2. **Comprehensive sky maps**
   - Goal is (i) full-sky, (ii) IQU, (iii) nside ≥ 8192

3. **Sky models as callable functions of frequency**
   - Needed for on-the-fly band integration, to enable full exploration of band-space including intra-frequency bandpass mismatch.

Mission Modeling

- Instrument
  - Individual detector beams, band-passes
  - Auto- and cross-correlated noise components
    - detector, pixel, wafer, readout, telescope
  - Absolute & relative calibration
  - Polarization modulation systematics
  - Other
- Observation
  - Scanning strategy
  - Environment (atmosphere, ground-pickup; cosmic ray flux, …)



Wish List

- **Define mission parameter *distributions* to sample from.**
  - Required to simulate full missions before exhaustive characterization

- **Define *comprehensive* mission definitions and then derive reduced approximations**
  - Map domain: band-averaged noise, bandpass, symmetric beam
  - Spectral domain: spectra of residuals
  - Required for meaningful comparison of simulations in different domains.

Time Domain

Overview

- Fundamental mission deliverable is the reduction of the time-ordered data to a set of well-characterized frequency maps.

- Despite their expense, time-domain simulations are absolutely necessary
  - Design and development (after optimization in cheaper domains)
  - Pipeline validation & verification
  - Monte Carlo data characterization (10^4 x data!)

- Time domain analyses introduce their own systematic effects.

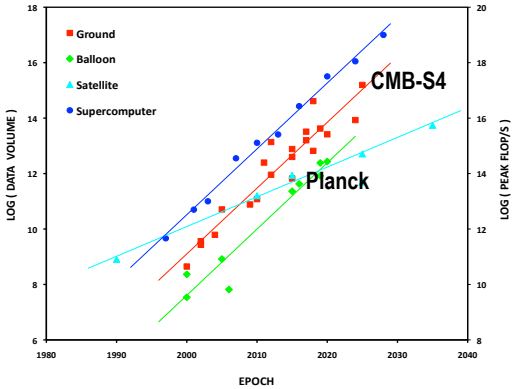
TOD Challenges

REALISM: ALGORITHMS

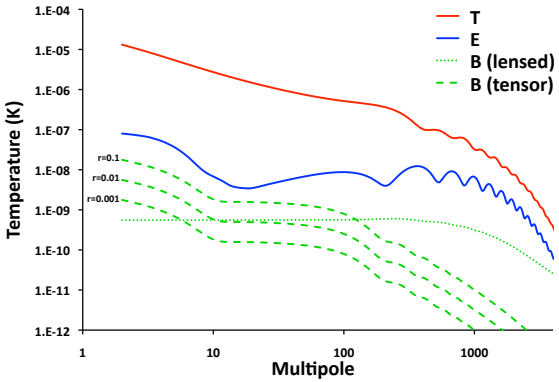
- 1000x systematics sources
  - Atmosphere
  - Ground pickup
  - Advanced instrumentation
  - Cross-correlated noise
  - …

PERFORMANCE: IMPLEMENTATIONS

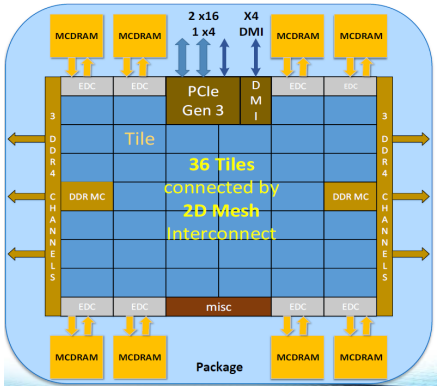
- 1000x data volume



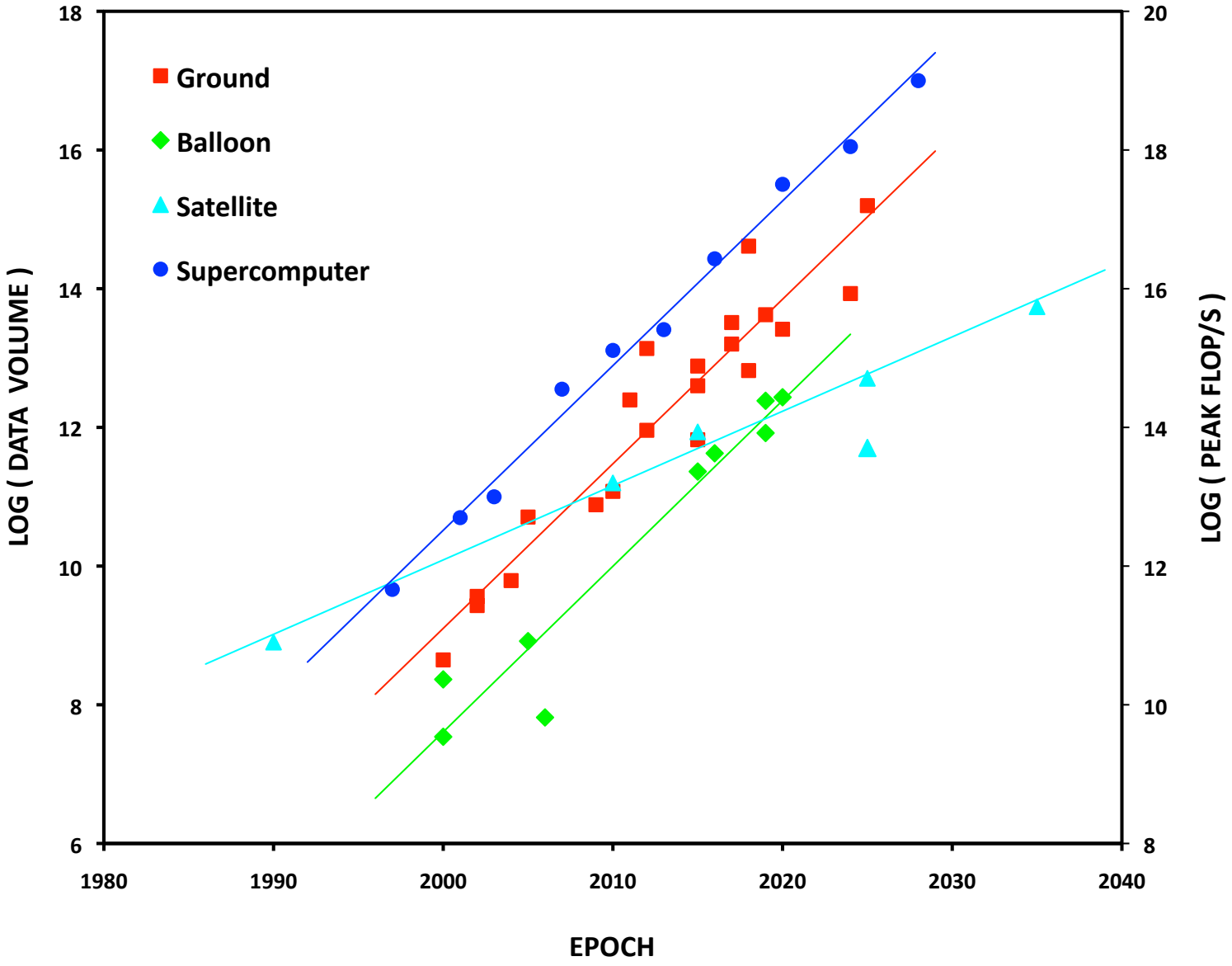
- 100x lower systematics threshold



- 100x fewer watts per FLOP



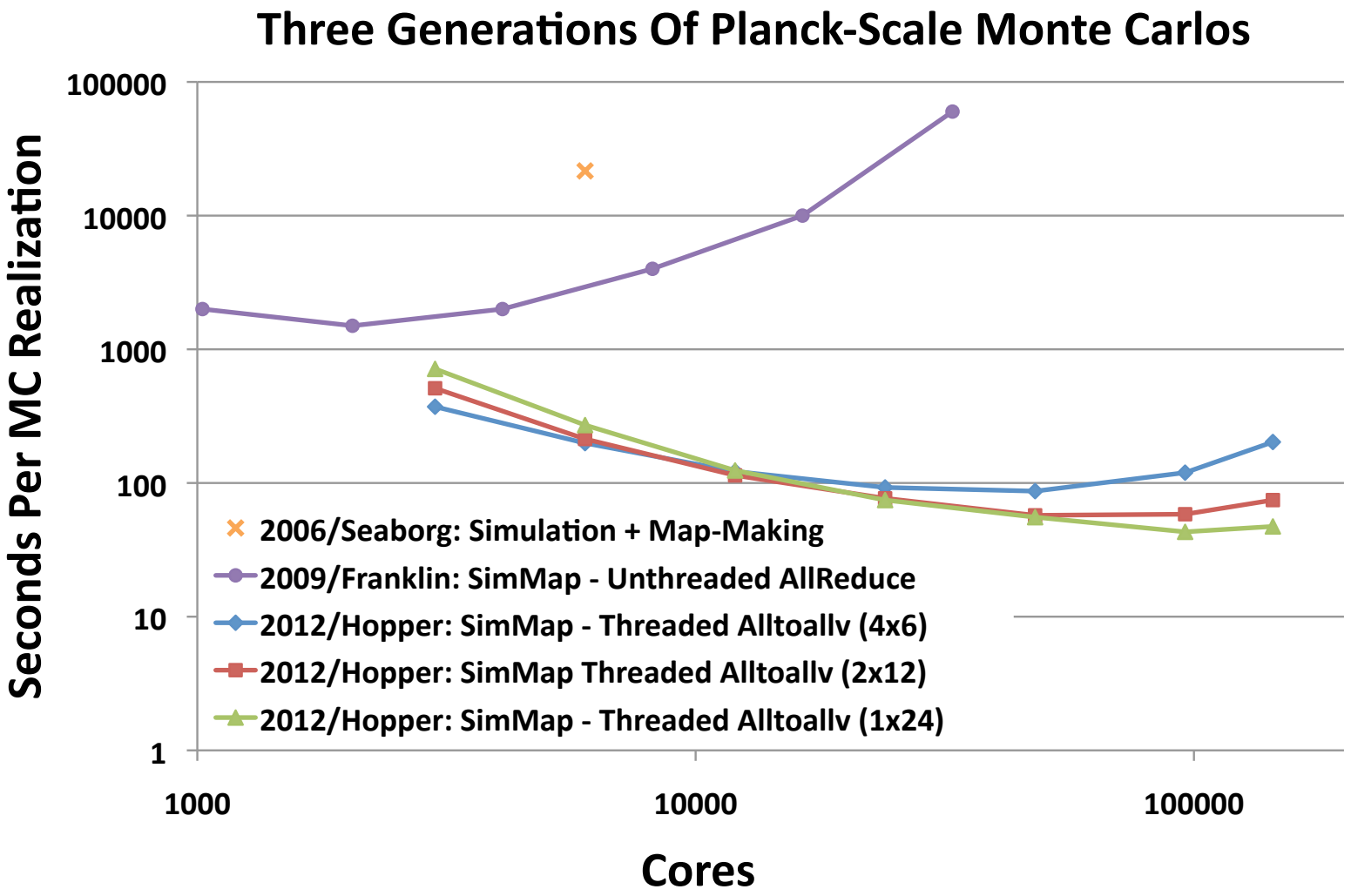
Moore's Law for CMB



BERKELEY LAB

C3 COMPUTATIONAL COSMOLOGY CENTER

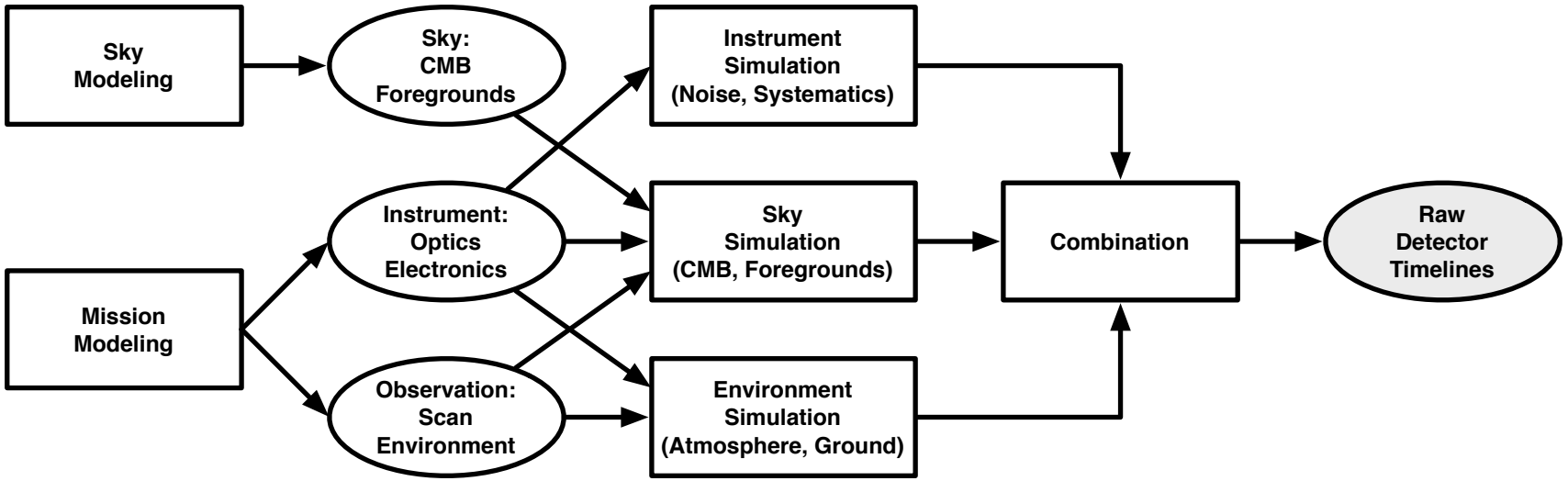
It's Déjà Vu All Over Again



BERKELEY LAB

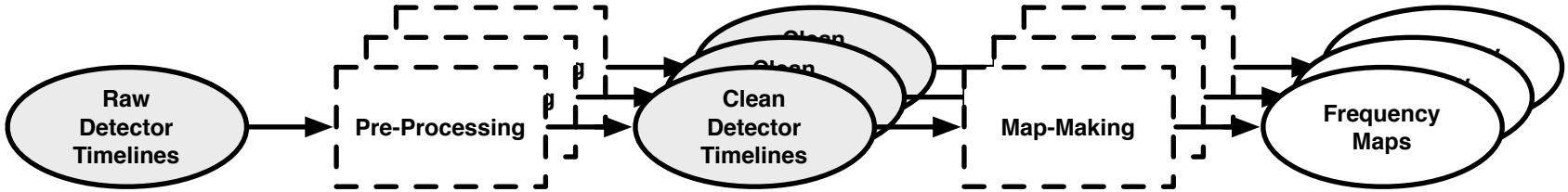
C3 COMPUTATIONAL COSMOLOGY CENTER

TOD Simulation Framework



TOD Reduction Framework

- Feed the total timeline directly into multiple reduction pipelines to amortize the simulation cost – fully on-the-fly framework.
- Each map-making may use the same or different pre-processing.



TOAST Design Drivers

1. Must be easy to prototype algorithms
   - We don't yet know how to mitigate systematics to the level required to recover faintest signals.
2. Must be scalable to massive parallelism (both HPC & HTC)
   - We want to simulate/reduce a huge data volume on the fly.
3. Must be readily available & open to developers
   - We need a coordinated, community-wide effort

$\Rightarrow$ Python-wrapped, MPI/OpenMP hybridized, compiled libraries

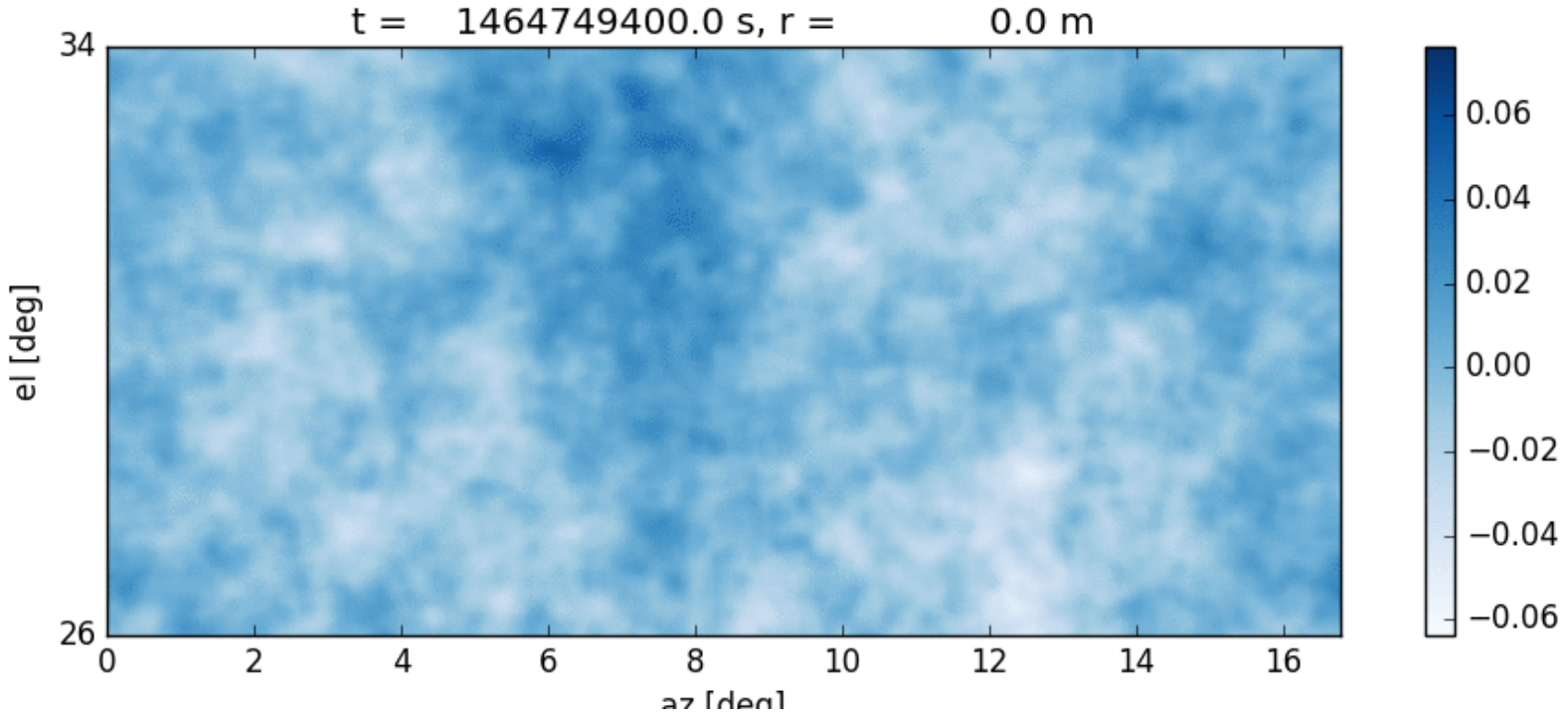
$\Rightarrow$ Code available on github, HPC resources available at NERSC

Proof-Of-Principle Run

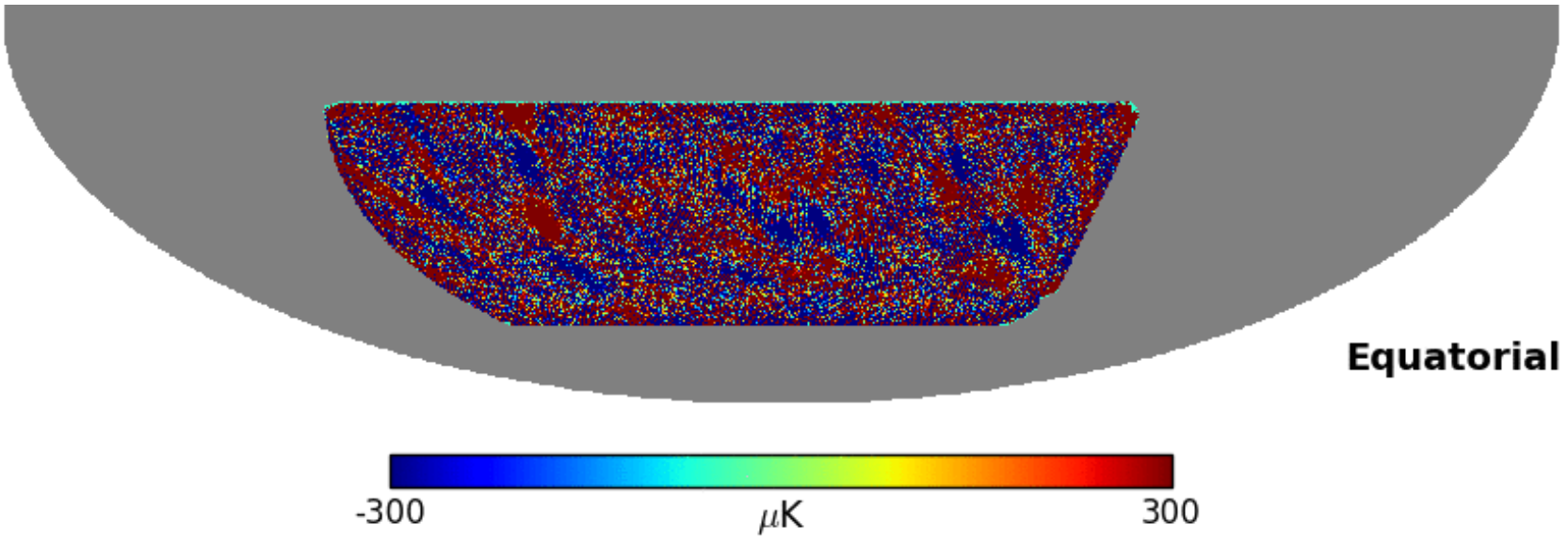
- Simulation:
  - 50K detectors; 7 frequencies; 20% of the sky;1 year: 35X Planck
  - Piecewise stationary atmosphere, noise + sky
  - Per detector calibration uncertainty
- Reductions:
  - 3rd order polynomial filter, bin to daily maps
  - 3rd order polynomial filter, bin to full-season maps
- Run on entire Cori-2 system at NERSC
  - 600,000 cores running 150,000 MPI-tasks each with 4 threads
  - Docker/Shifter to launch Python on 600,000 cores!

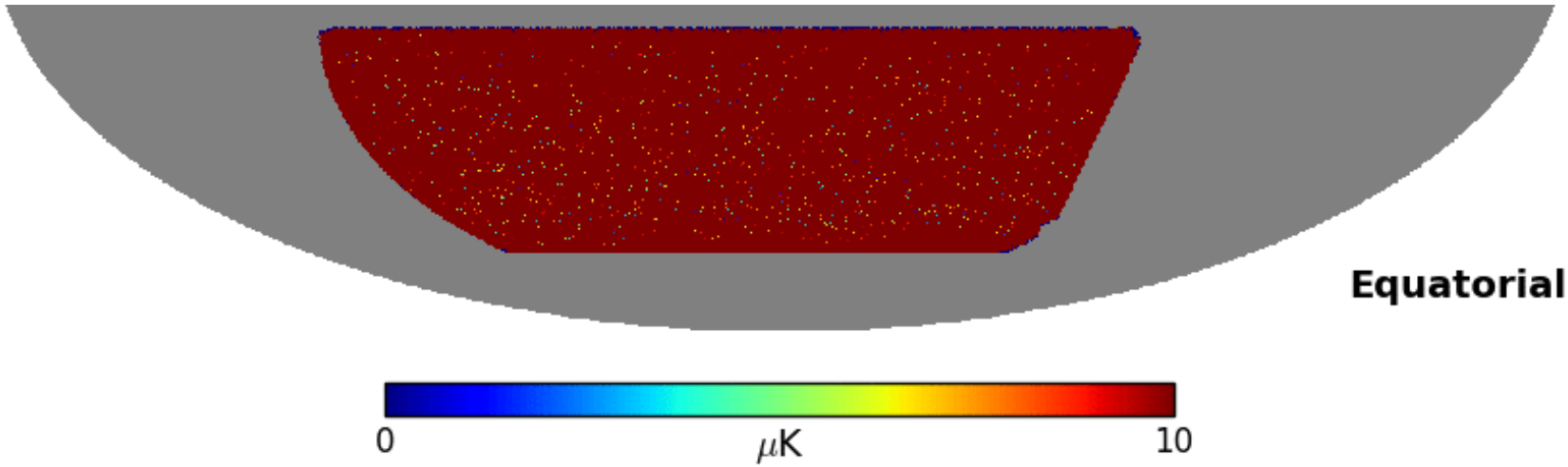
Atmospheric Realization

- Generate bounding volume of atmosphere seen by the telescope in one scan period (2- & 3-D Kolmogorov spectra)
- Perform line-of-sight integral for every observation.

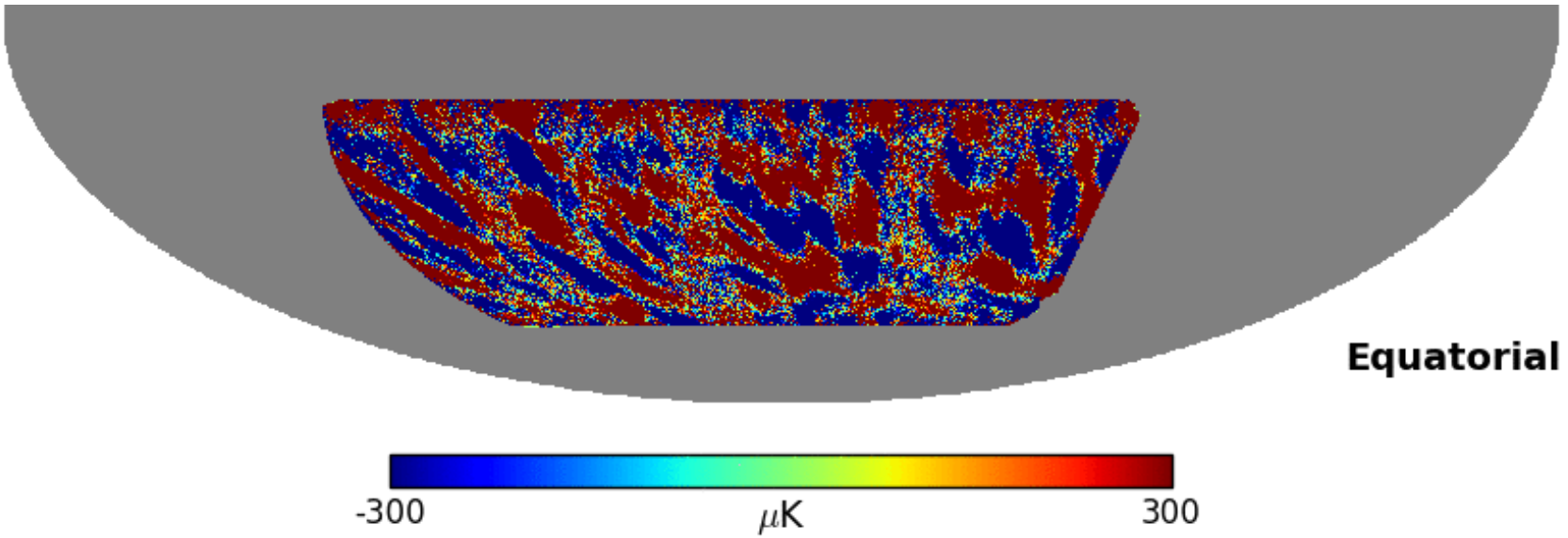


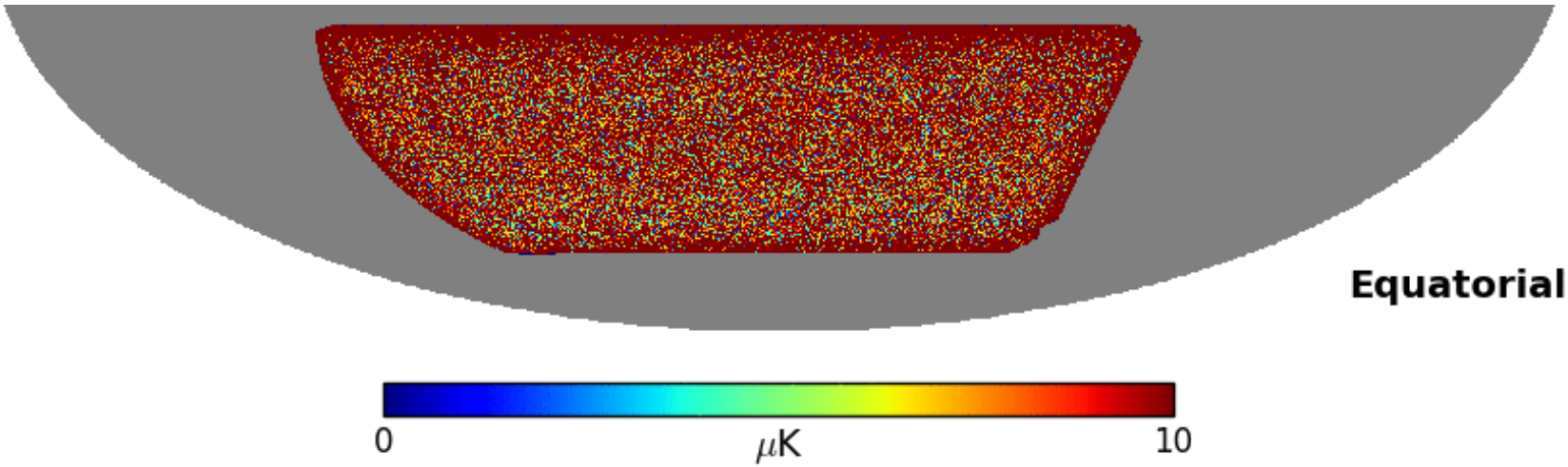
Cumulative T & P Maps: 20 GHz



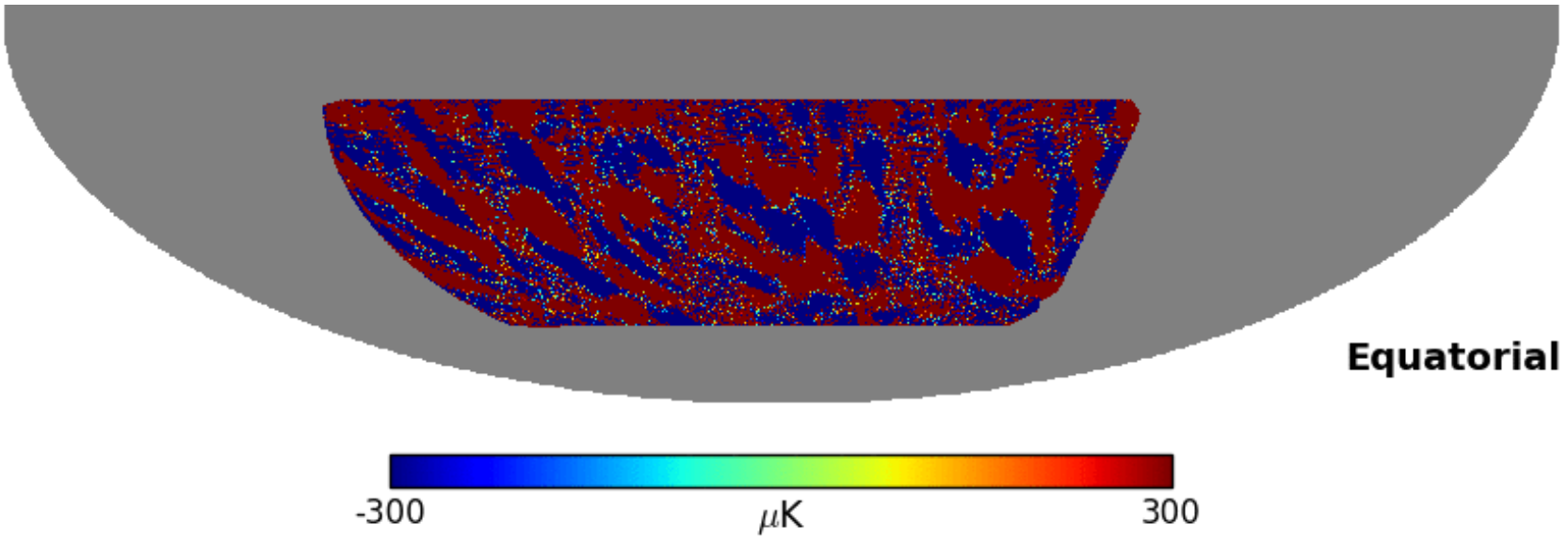


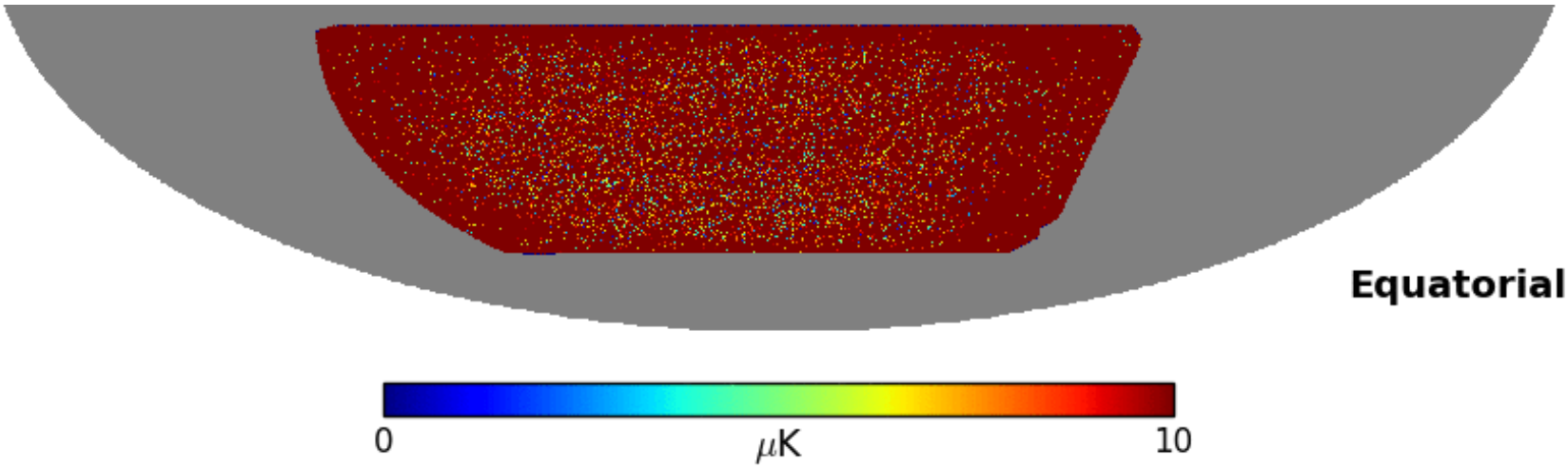
Cumulative T & P Maps: 95 GHz





Cumulative T & P Maps: 270 GHz





Collaboration & Coordination

- We have limited human and computational resources; we must make the most effective use of them across all the proposed missions.
- Multiplicity is valuable, but not at the cost of completeness or consistency.
- Global collaboration on generic issues:
  - Sky modeling
  - Reduced metrics & common requirements
  - TOD framework development & deployment
  - Science analyses
  - HPC, especially Xeon Phi efficiency:
    - NERSC (#6: Cori), ALCF (#16: Theta, Aurora), JCAHPC (#7: Oakforest), CINECA (#14: Marconi), CSC (?)