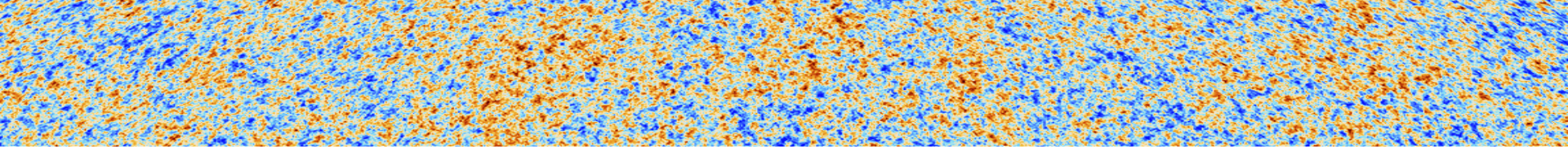


End-to-end simulations

Paolo Natoli

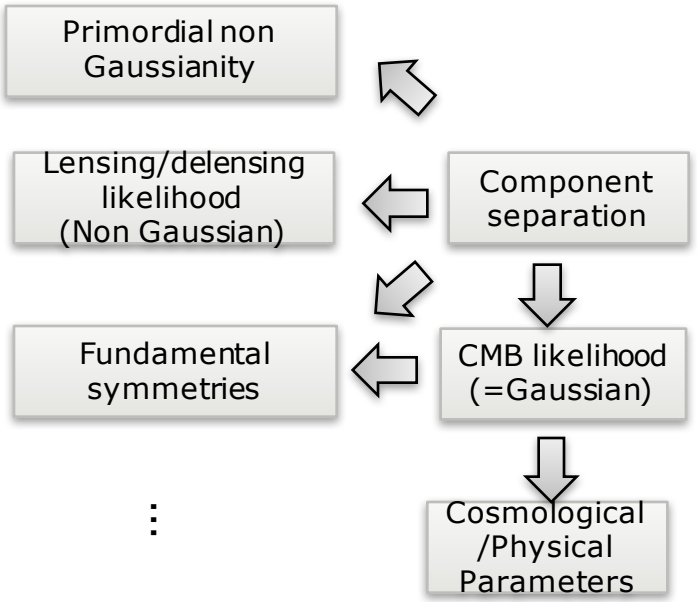
Università di Ferrara and INFN

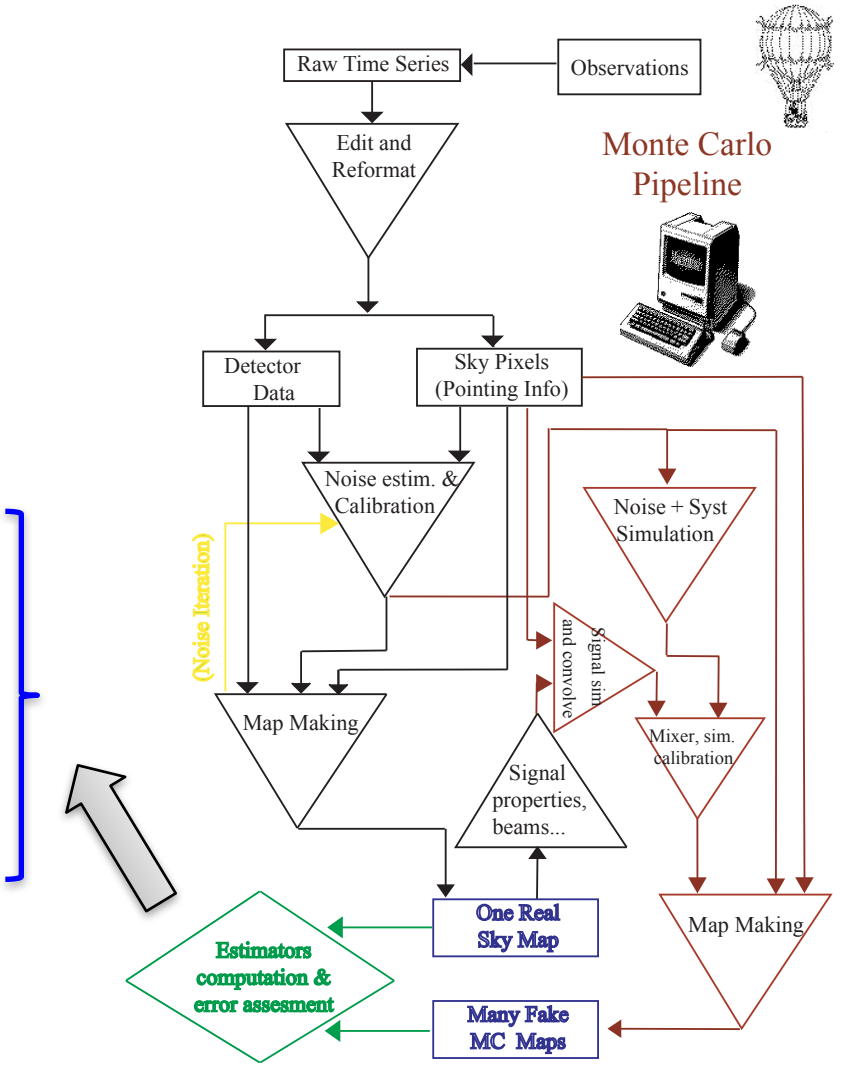
First COSMOS meeting

Bologna, 20 January 2017

CMB data analysis pipeline

- Complex pipeline
- **Relies on simulated data**



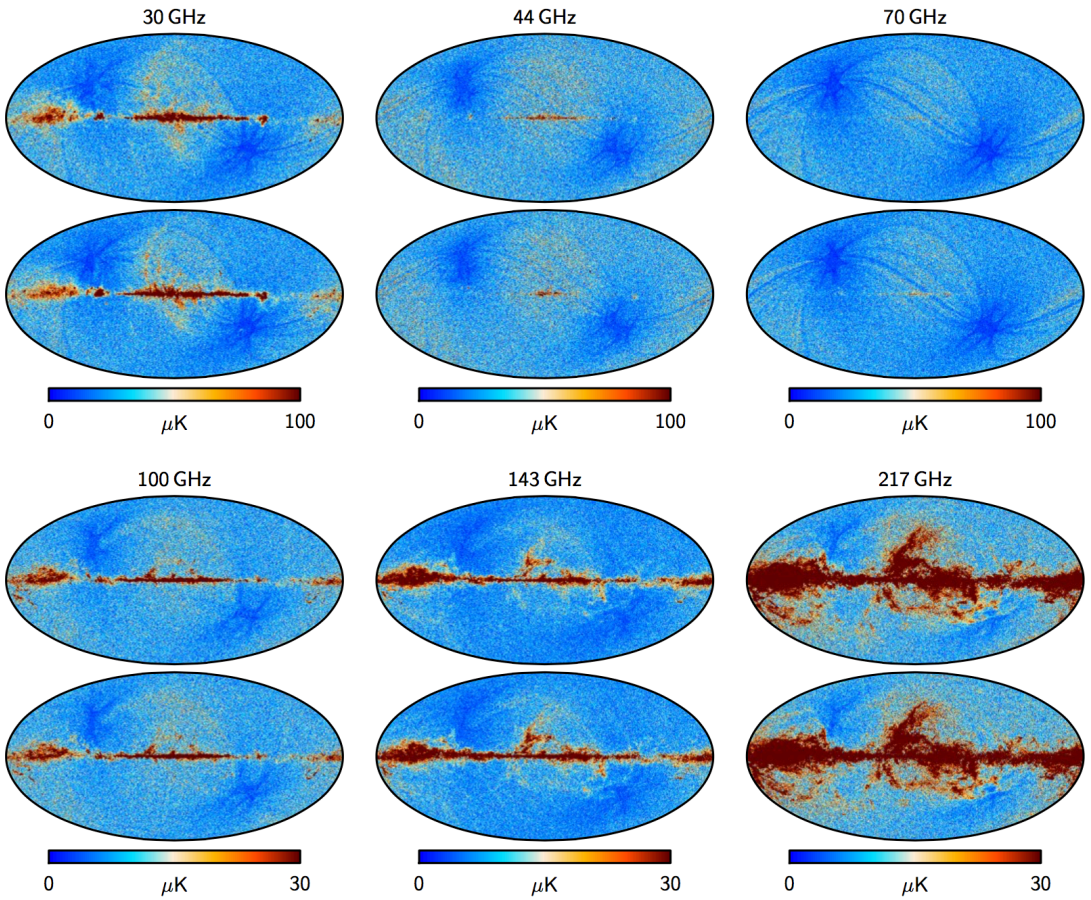


Use of simulations

1. Pre launch mission definition:
    a. To assess the impact of systematic effects (If I do not see it why worry?)
    b. To validate algorithms and tools
2. Actual data analysis:
    a. To perform debiasing of estimators from spurious contributions (systematics and noise which cannot be modeled otherwise)
    b. To assess uncertainties, including covariances that are notoriously tricky.

Planck Full Focal Plane simulations

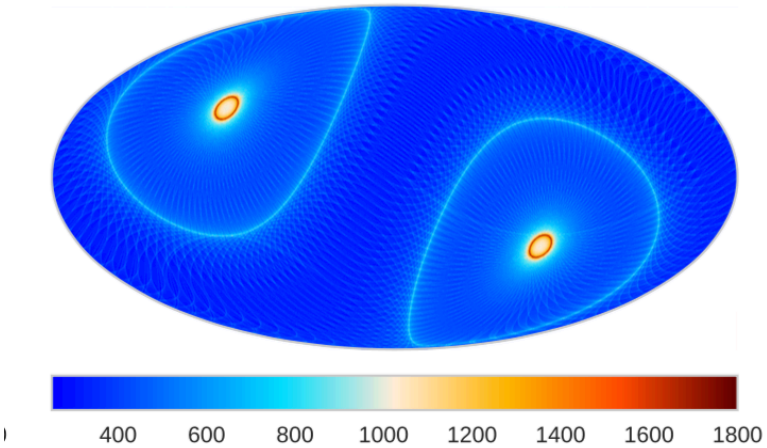
1. End to end effort for all Planck Channels [arXiv:1509.06348]
2. Major computational burden was set of 10^4 Monte Carlo maps: 1 million CPU-days on world class super computer (NERSC and CSC)
3. Supported Planck cosmological analysis



Scaling this up: Planck/FPP8 comparison to CORE

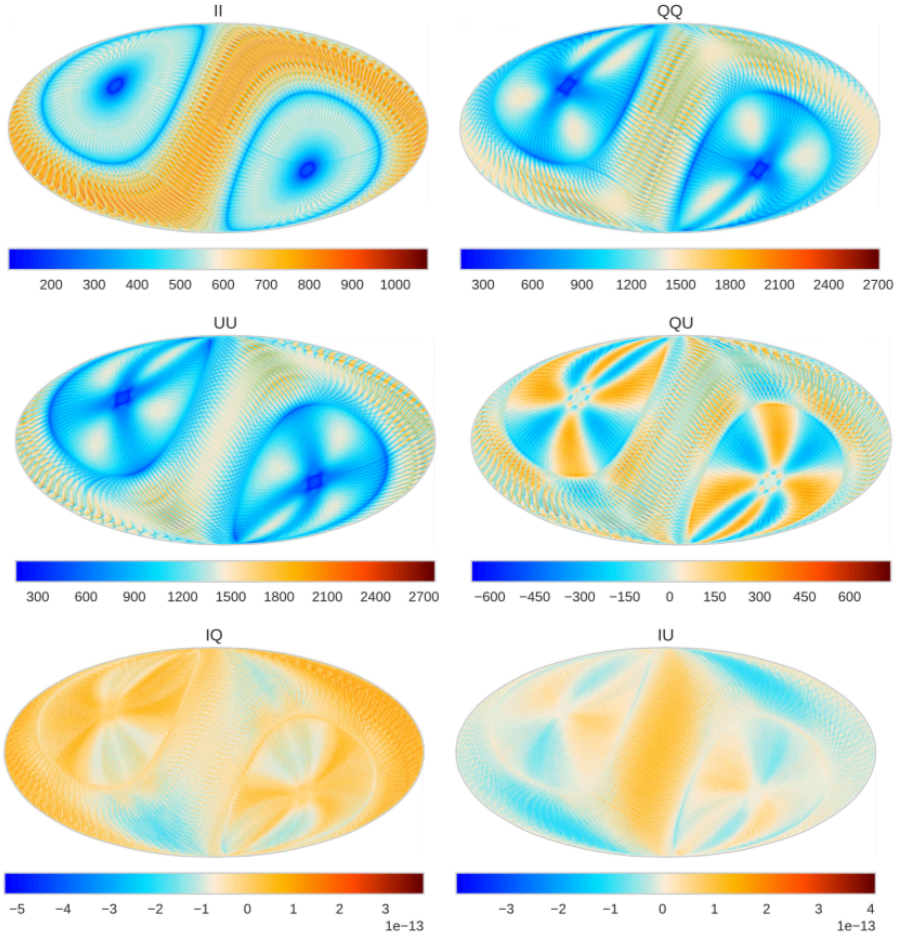
1. Noise (timeline to map) Monte Carlo
    a. 850 CPU-hours for single map set realization of the 52 detectors of the Planck HFI dataset (5 surveys).
    b. For the minimal Core+ configuration, this would translate into about 40000 CPU-hours per map set (assuming 33% increase in sampling rate and same timeline length)**, on existing hardware.**
    c. These are affordable numbers for forecasted HPC allocations even without advocating for aggressive versions of Moore's law
2. CMB Monte Carlo
    a. Dominated by beam convolutions
    b. High number of hit per pixel for Core+ makes pixel space convolution via effective beam (a la FeBECOP) over appealing compared to harmonic methods
    c. Monte Carlo generation for these methods scales with number of pixels in the beam and the map, not with timeline length (once precomputations are done).
    d. More detectors means more beams, resolution (perhaps) smaller
    e. The Planck FFP8 computational requirement is then in the ballpark of what we need: about 200 K Cpu-h for each 10000 MC set.

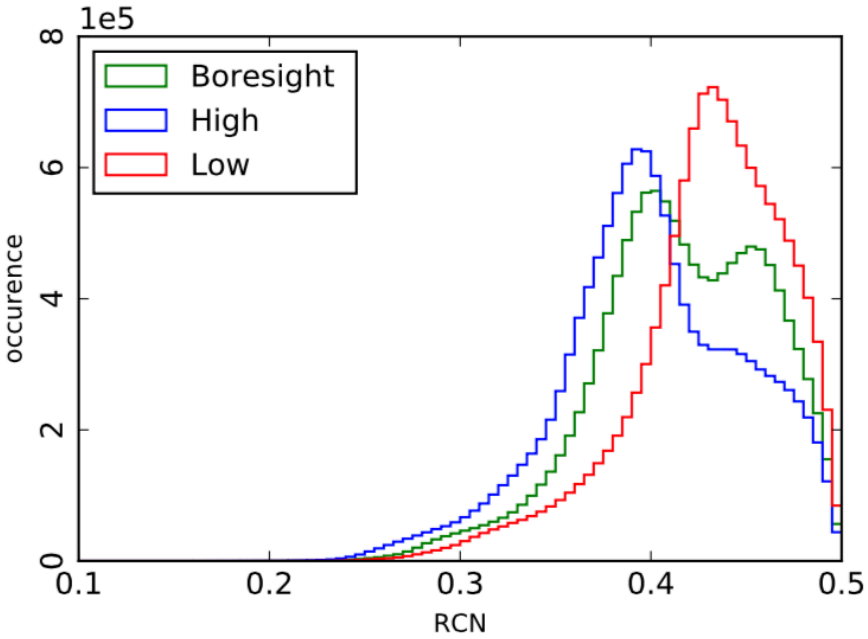
Application to CORE



1. ECO CORE paper on simulations and systematics in preparation (to be submitted in early February). Coordinated by PN and Mark Ashdown (Cambridge)
2. Heavily based on timeline-to-map propagation
3. Scope is to support the case of CORE during selection process. Provides state of the art assessment of where we stand in terms of capabilities
4. Largely based on a quasi-public engine for simulations:
    a. TOAST (http://github.com/hpc4cmb/toast)
    b. madam, flexible and accurate map making (https://github.com/hpc4cmb/libmadam)

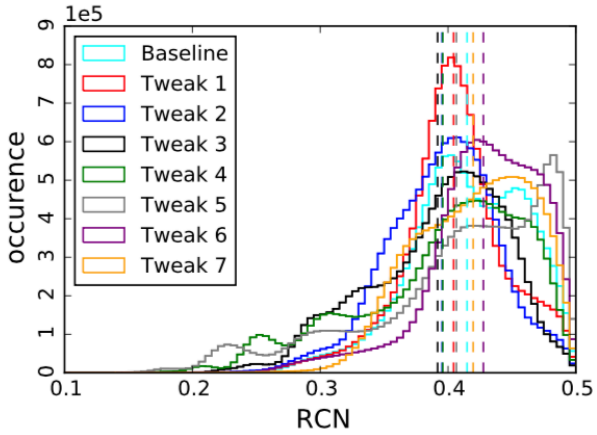
Polarization sensitivity and purity

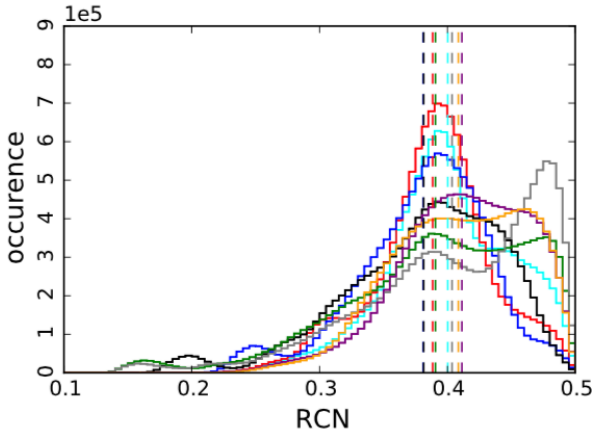


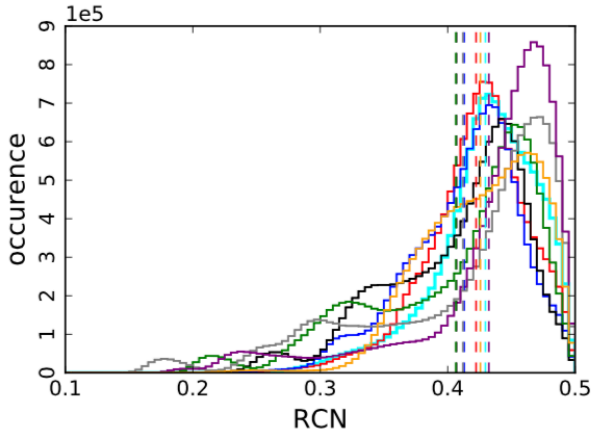


Linda Polastri

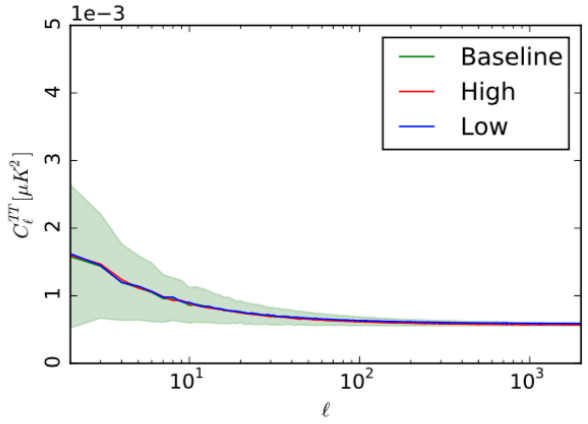
Optimization of the scanning strategy

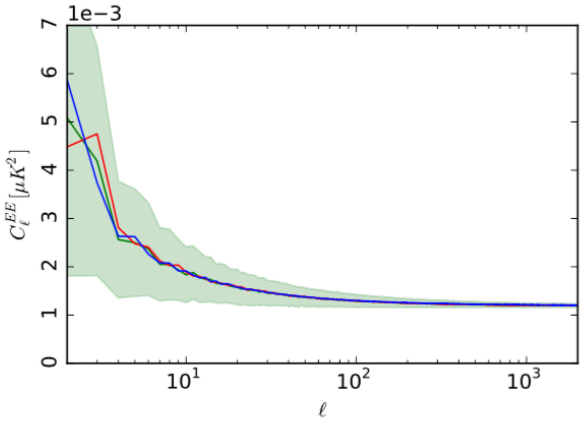


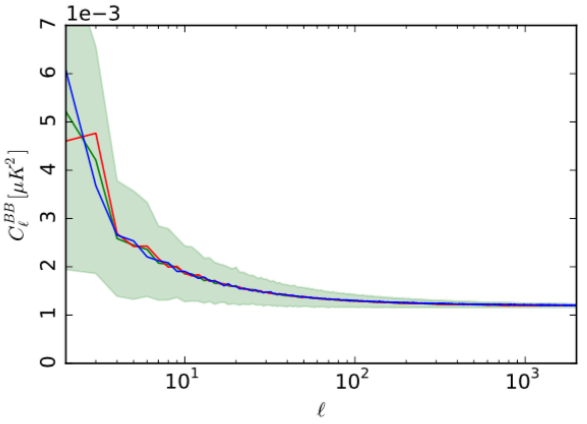




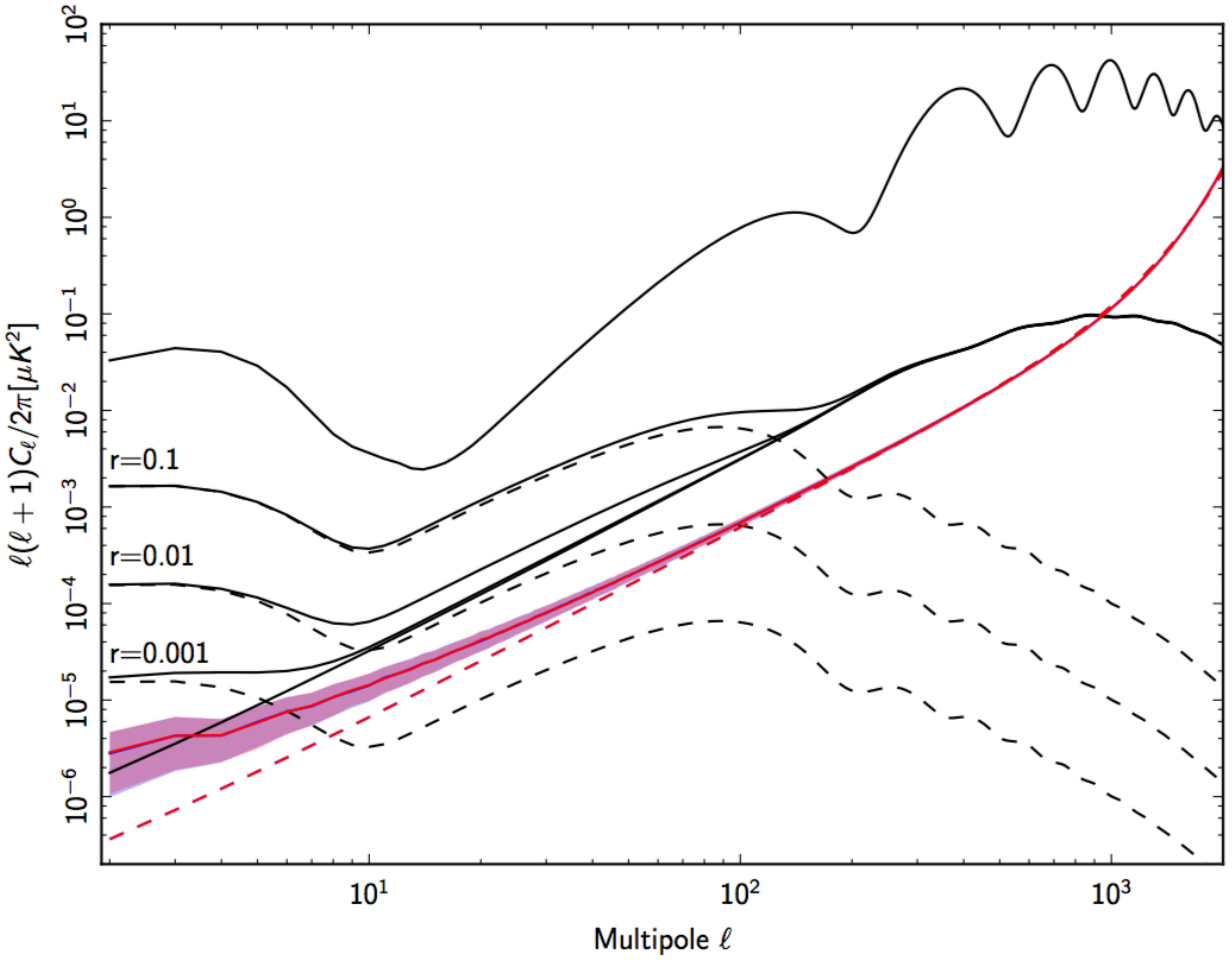
Linda Polastri, Diego Molinari





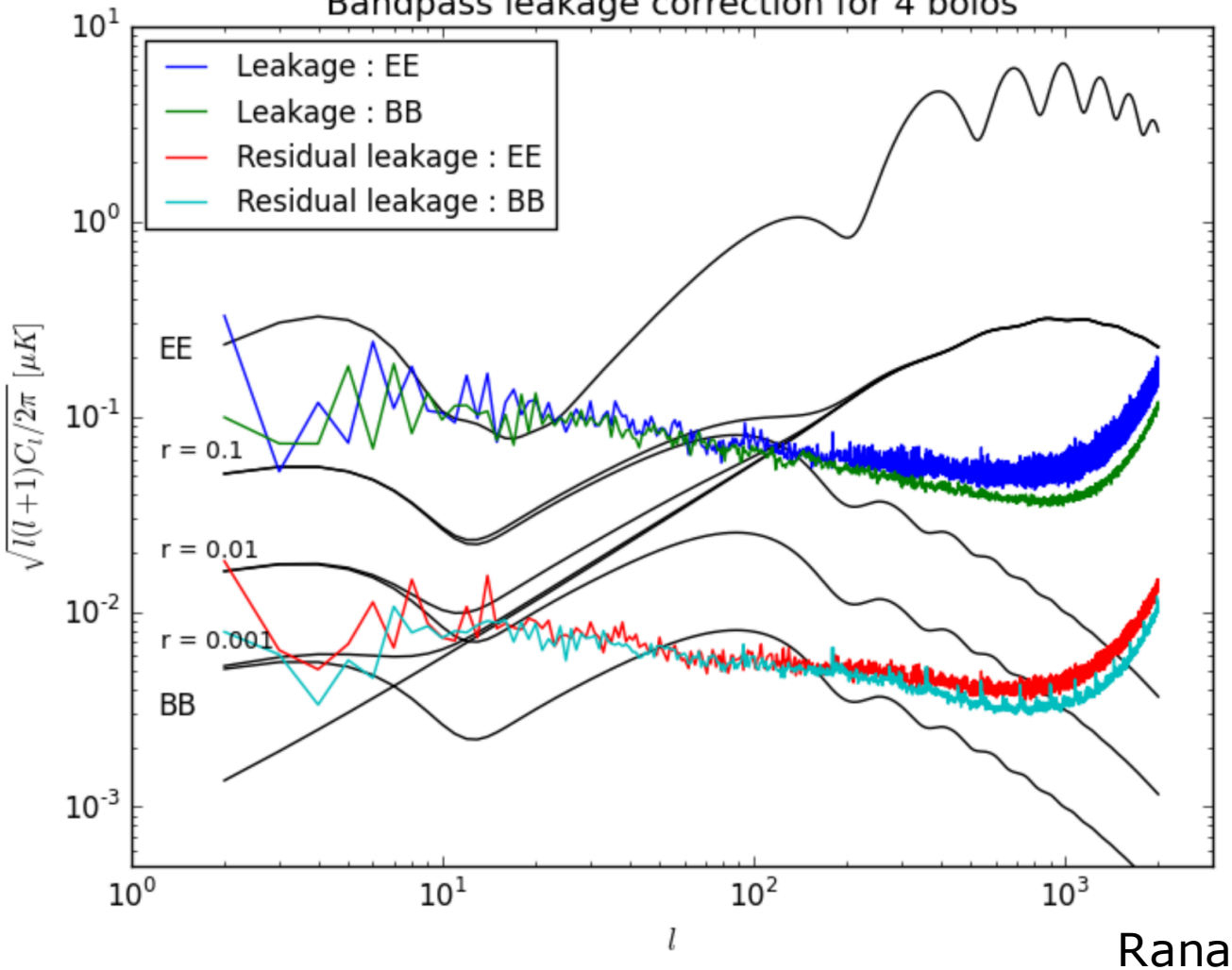


130-220 GHz noise projection for CORE



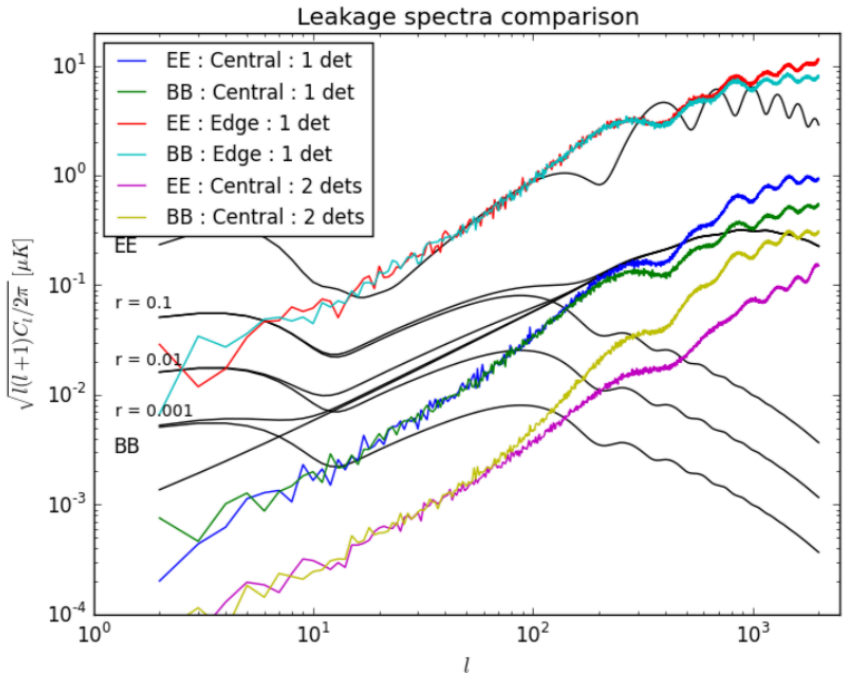
Diego Molinari

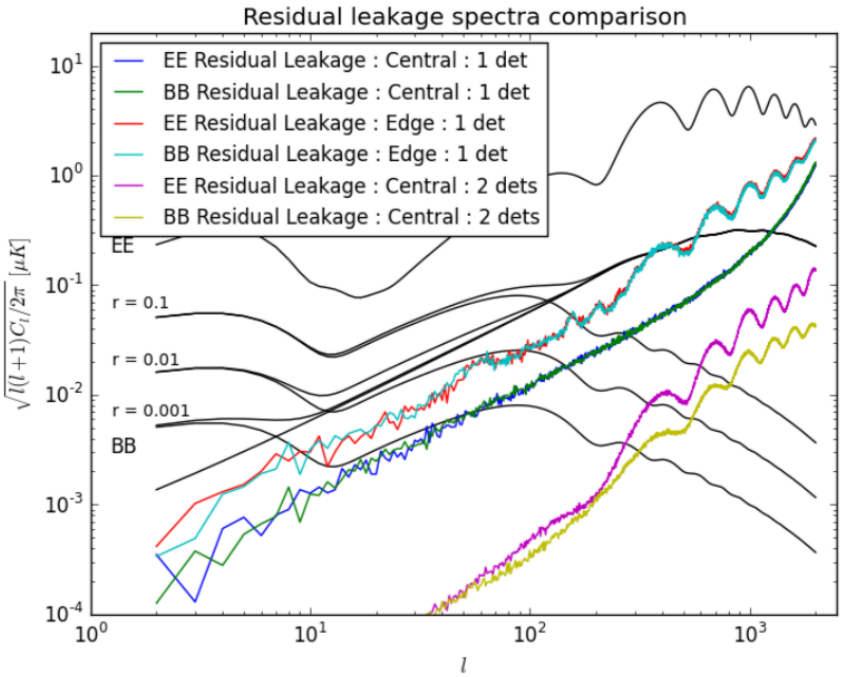
Band pass mismatch correction



Ranajoy Banerji

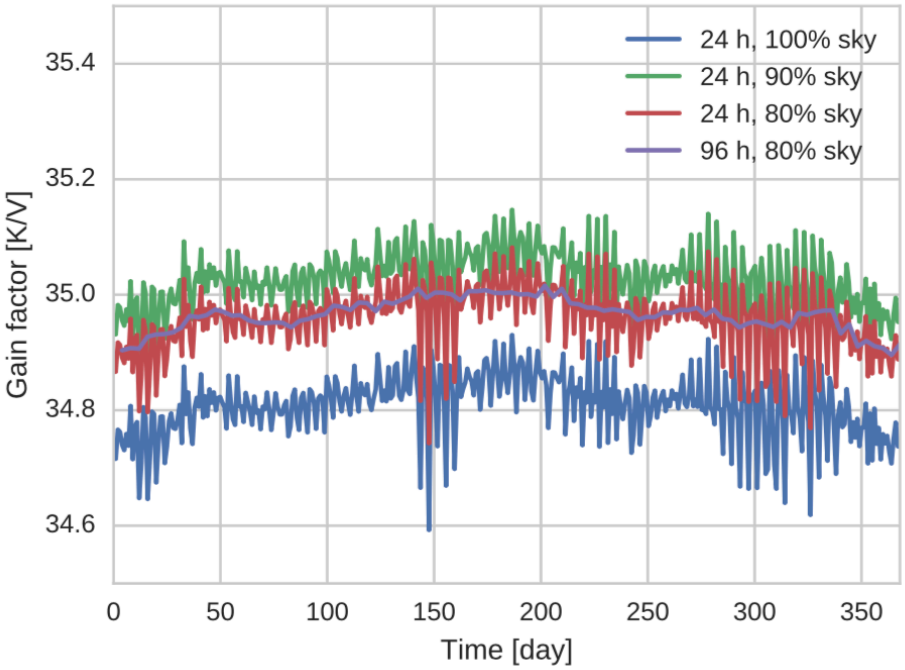
Leakage due to main beam asymmetry

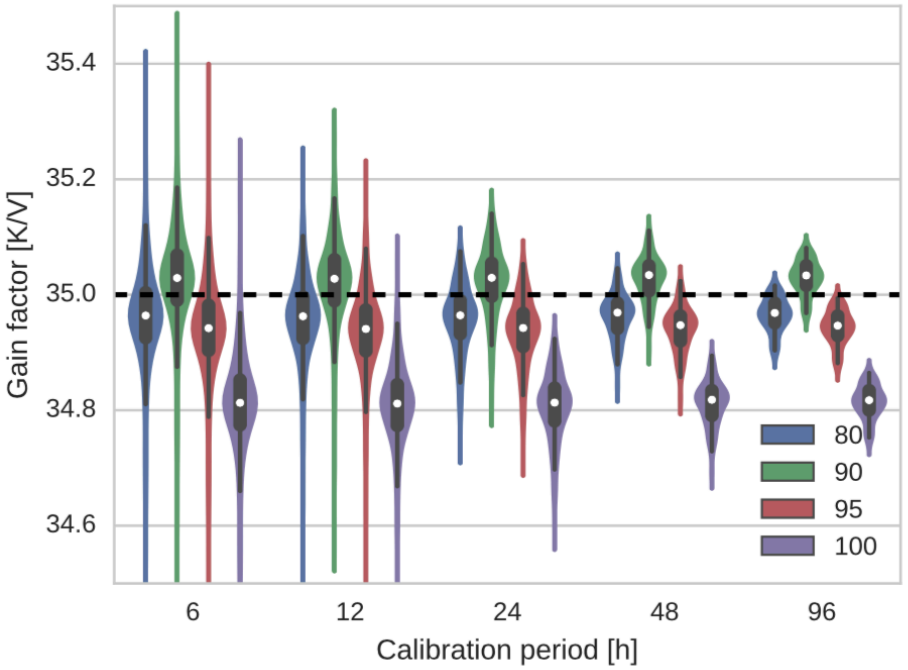




Ranajoy Banerji

Calibration

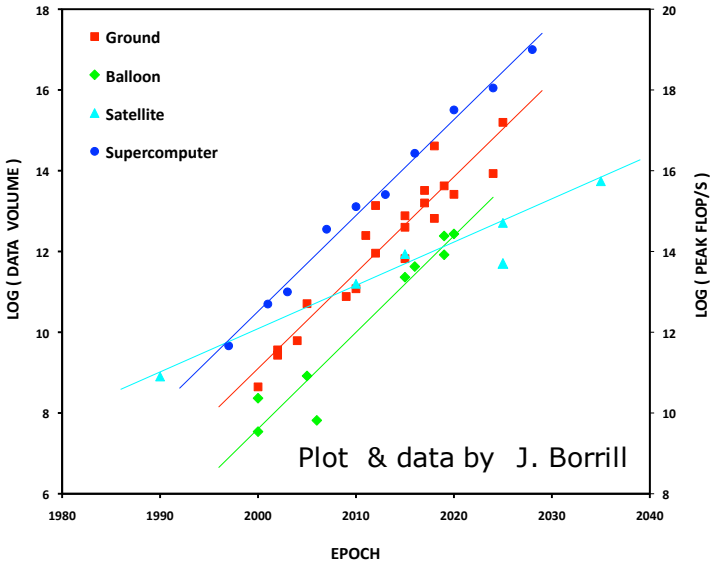




Maurizio Tomasi

More detectors, higher computational needs

1. Projected scaling up of computing power (based on some version of Moore's law) allows in principle to scale up to cover forthcoming ground based experiments…



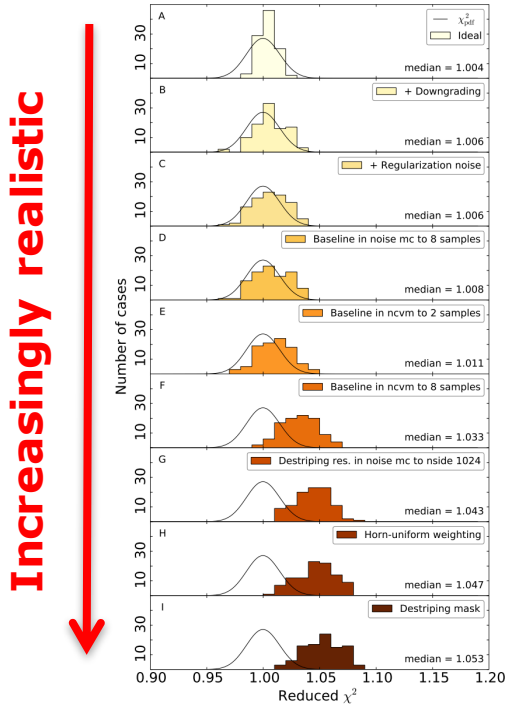
2. All feasible? Beware of several things…

Issues and worries: computing power

1. Difficult to forecast the evolution of supercomputing resources in the next decade, at a time when large system are already becoming severely energy constrained. Many experts speak openly of Moore's law coming to an end. We go into uncharted territory.
2. Even ignoring the above, exploitation of available resources (when available) are limited by user concurrency and cost of flop unit.
    a. *We are not the only community in need of significant computing power. We are already competing for resources*
    b. *Must find a balance between cheap flops offered on clogged computers and costly dedicated service. Can European coordination play a role here?*
3. Sheer size of data limits human direct intervention. Automatization is a must and complicates business.

Issues and worries: accuracy

1. Accuracy needs are already scaling up simulation volume. Analytic shortcuts are not always easy or found.



- Example: Planck LFI map making (arXiv:1502.01585)
- χ^2 for pixel-pixel noise covariance matrix (a fairly large guy!)
- No instrumental nor sky driven systematic
- Pure data analysis effect: the algorithm to build this matrix is just an approximation of the real stuff.
- Would be worse for bolometers/colored noise
- Exact treatment cannot be achieved analytically. Simulation driven correction is an option, but requires large number of realizations

An historical touch

The Challenge Of Data Analysis For Future CMB Observations

J. Borrill, circa 1999

FUTURE PROSPECTS

We have seen that existing algorithms are capable of dealing with CMB datasets with at most 10^5 pixels. Over the next 10 years a range of observations are expected to produce datasets of 5×10^5 (BOOMERanG LDB), 10^6

TABLE 3. The computational requirements for one iteration of the Newton-Raphson algorithm to extract a 20-bin power spectrum for MAXIMA and BOOMERanG

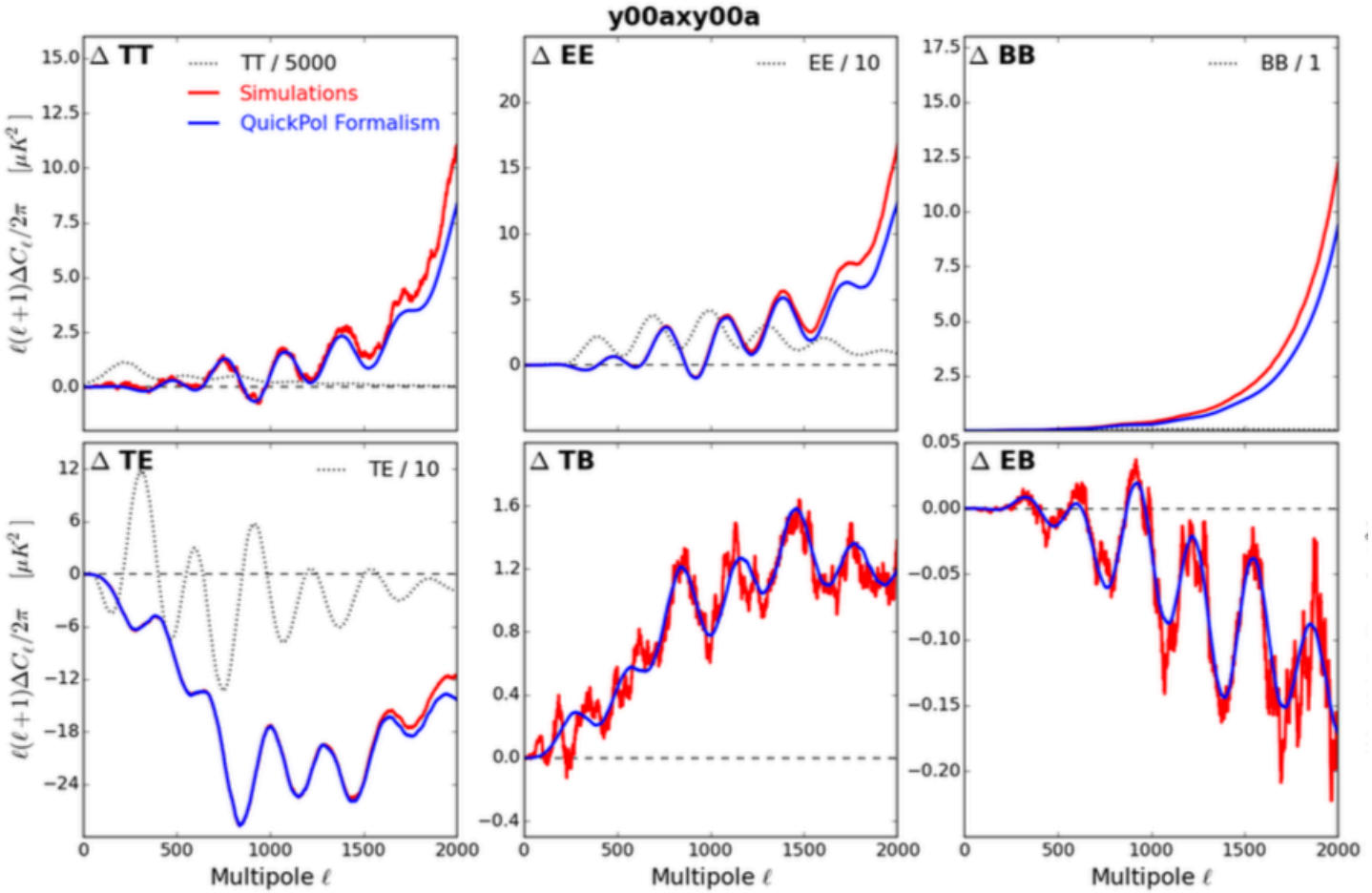
Flight	$\mathcal{N}_p$	Disc	RAM	Operations	Serial Time	Cray T3E Time
BOOMERanG NA	26,000	110 Gb	11 Gb	7.1×10^{14}	14 days	5 hours (64 PE)
MAXIMA 1	32,000	170 Gb	17 Gb	1.3×10^{15}	25 days	9 hours (64 PE)
MAXIMA 2	80,000	1 Tb	100 Gb	2.1×10^{16}	13 months	18 hours (512 PE)
BOOMERanG LDB	450,000	30 Tb	3 Tb	3.7×10^{18}	196 years	140 days (512 PE)

(MAP) and 10^7 (PLANCK) pixels that will necessarily require new techniques. This is an ongoing area of research in which some progress has been made in particular special cases.

A way out

1. It wasn't computers that saved us but rather the development of fast and accurate data analysis tools:
    a. *Fast transforms on the sphere (Muciaccia, PN & Vittorio 1996)*
    b. *Fast pixelization scheme (HealPIX, Gorski et al 1998+)*
    c. *Fast and accurate component separation (Gispert and Bouchet, 1997, …)*
    d. *Fast and accurate map making (PN et al., 2001, Dorè et al 2001, …)*
    e. *Fast power spectrum estimation (Hivon et al 2002, …)*
    f. *….*
2. These efforts have been severely driven by the *push* to analyze real data. We must be able to keep the trend alive. Live data is only way, cannot afford a 10 year gap.

Analytical modelling of effects



Eric Hivon

Conclusions

1. Tenuous CMB signal targets imply larger, more complex datasets. Analysis requirements will scale accordingly.
    a. The need to accurately exploit dataset combination will surely boost requirements up. Forecasting how much is hard.
2. Efficient access to supercomputing power is a critical urgency for our community
3. Cannot underestimate the power of end-to-end simulations but don't want to overestimate at the same time. With the systematic error budget dominating analysis, desire for advancement in methodology is growing