UNIVERSITÉ
PARIS 7 - DENIS DIDEROT

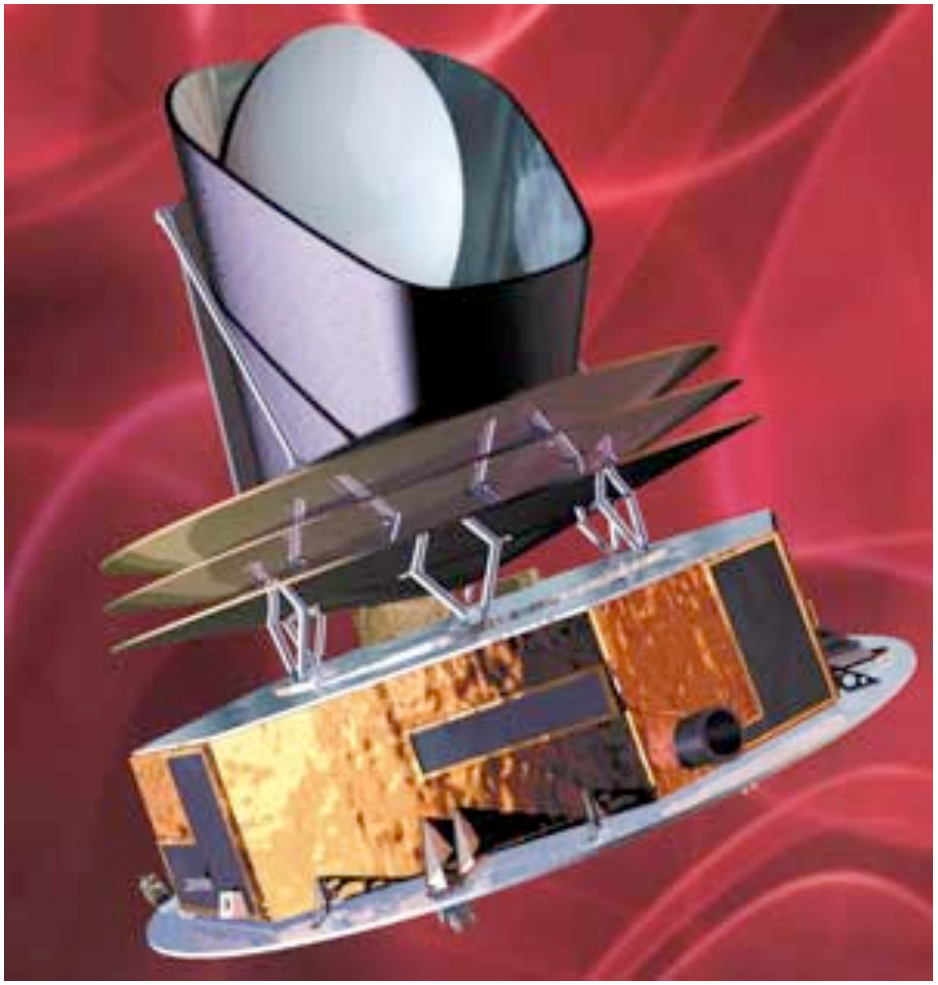
Data analysis of massive data sets a Planck example

Radek Stompor
(APC)

Outline

1. Planck mission;
2. Planck data set;
3. Planck data analysis plan and challenges;
4. Planck vs interferometers;
5. CMB work at APC;
6. Conclusions.

The Planck Satellite



- An ESA mission, with a significant contribution from NASA, due to launch in fall 2007.
- Will carry on board 72 detectors shared between two High (46) and Low Frequency (26) Instruments.
- An 12+ month all-sky survey at 9 microwave frequencies from 30 to 857 GHz.
- $O(10^{12})$ observations;

  $O(10^{8})$ sky pixels;

  $O(10^{3})$ spectral multipoles.

The Planck data set - the sizes

- Raw data
  i. detector readouts;
  ii. pointing information;
  iii. household data;
  all amounting to roughly ~ 2-5 Tbytes of the data.
- Derived products (including mission deliverables)
  i. systematic templates, noise filters, etc
  ii. single frequency maps;
  iii. foreground component maps;
  iv. power spectra and their covariances;
  v. cosmological parameter likelihoods, etc.
- Temporary objects
  i. simulated data corresponding to various stages of the analysis

  all amounting to roughly 100 x raw data ≈ 200-500 Tbytes of the data stored and in a need of frequent processing at a peak of the data analysis activities.

The Planck data set - the characteristics

- Mostly low signal-to-noise data (100 uK √sec ≈ 1500 uK) with a trend set by a cosmological dipole (~3mK) and occasional high signal nearly periodic intervals corresponding to the Galactic plane crossing;
- Properties slowly (i.e., piece-wise stationary) evolving with time and have to be derived directly from the data themselves;
  i. instrumental noise: piece-wise stationary with "1/f" correlations;
  ii. responses, etc.
- Uniform sampling (~ 5 msec for HFI detectors), but cosmic rays hits, planet crossings, other transients and telemetry drop-outs;
- Systematic effects (temperature drifts etc) common to most of the bolometric detectors;
- Two types of the detectors (LFI radiometers and HFI bolometers);
- Simple scanning strategy.

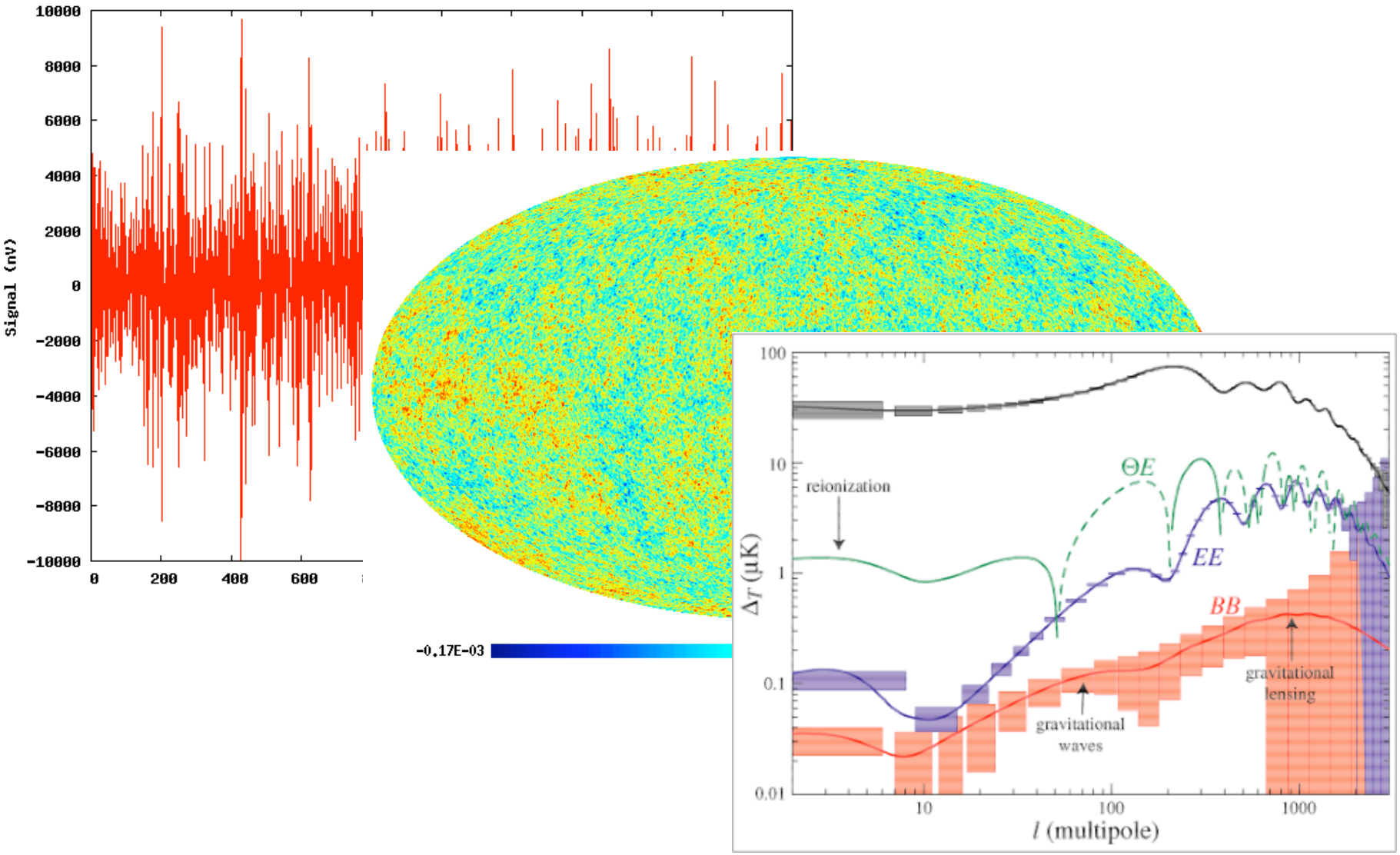
Planck will produce not only large but also complex data set to be analyzed.

The Planck data set - the context

- Probably the largest (or one of) CMB data sets available at the time;
- Almost two orders of magnitude larger than the WMAP data set (~50 Gbytes) and significantly more complex);
- Yet to be superseded soon by next generation of CMB experiments aiming at the detection and and characterization of the B-mode polarization and thus required to produce at least 2-3 orders of more data than Planck.

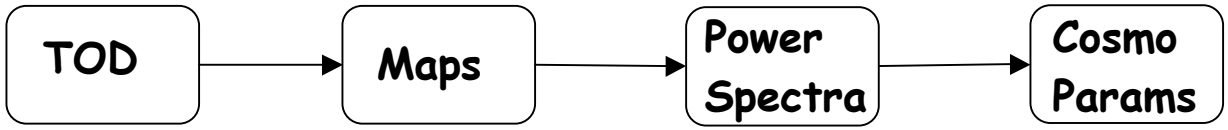
Planck marks a transition in the CMB data analysis
from interactive and subjective
to automatic and objective.

CMB Data Analysis - I



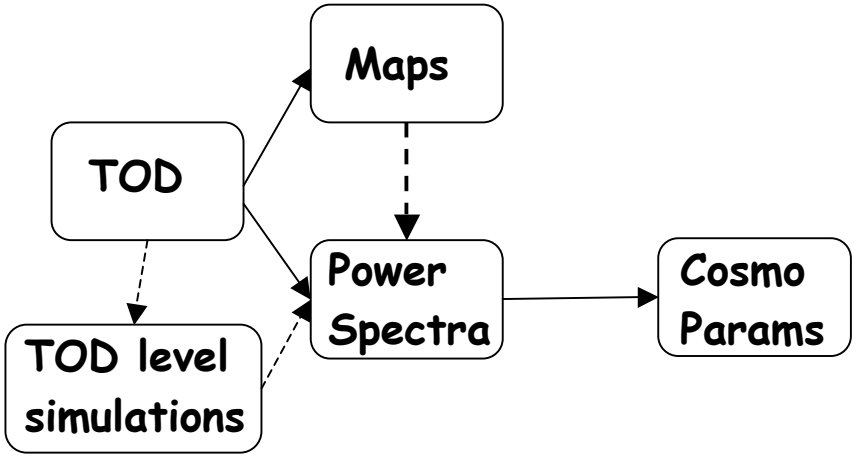
CMB Data Analysis - II

- "standard" way:



+ all the rest ….

- Planck way:



+ all the rest ….

CMB Data Analysis - III

Approximate analysis algorithm kernels are :

i. FFT in time-domain - $N_t \log_2 N_t$

ii. SHT in pixel domain - $N_p^{3/2}$

Implicit multiplications and/or inversions of a rank $O(10^8)$.

i. PCG solvers [$O(10^2)$];

ii. Sparse matrix algebra;

iii. Monte Carlo realizations [$O(10^4)$].

For Planck :

$N_t \sim 5 \times 10^{11}$ (72 detectors, 12 months)

$N_p \sim 3 \times 5 \times 10^7$ (i, q & u maps)

$N_l \sim 6 \times 3 \times 10^3$ (TT, TE, EE, BB, TB & EB spectra)

Planck computational resource forecasting

- O(10-100) Exaflop of total processing capacity (per year);
- O(10) TB of total memory;
- O(10) GB of memory per processor;
- Fast inter-processor and inter-nodes communication (latency less of an issue);
- O(100) TB of archival file storage for primary data and derived data products;
- O(10) TB of scratch file storage for temporary products;
- O(1-10) GB of local tmp storage on each processor or node for out-of-core calculations and internal code products;
- General single file system; fast, scalable I/O suitable for reading and writing of huge data sets, and allowing for a simultaneous disc access of multiple massively parallel I/O-heavy applications;
- Special hardware facilitating fast Fourier and spherical harmonic transforms;
- Fast (possibly Gigabit or better) network connections with DPC computers, and fast connection with other supercomputing centers involved in the Planck effort.

CMB Data Analysis - IV

- Planck data analysis will require $O(10^{19})$ operations
  - 1 ExaFlop (10 million Tflop) - or more !

Planck marks a transition in the CMB data analysis from serial to parallel on basically all levels of the data processing.

Time domain data analysis tasks

Goal: to create and test a data model and use it to estimate the sky signals (maps)

- Pointing reconstruction;
- Noise estimation and characterization;
  - i. Gaussianity/stationarity tests;
  - ii. joint frequency-time analysis, etc …
- Deglitching;
- Instrumental effects (de-)convolution;
- Calibration:
  - i. detector responsivity;
  - ii. bolometric constant determination;
  - ii. beam reconstruction.
- Systematic effects reconstruction and modelling ;
- Map-making and some of component separations (time domain based) techniques;
- Time-domain based point source extraction techniques;
- Map error estimation;
- Time domain based simulations.

Pixel domain tasks

- Image processing/filtering/smoothing;
- Pixel-based component separation (e.g., ICA based methods);
- Pixel domain point source subtraction;
- Non-Gaussianity tests;
- Power spectrum estimation:
  i. monte carlo (pseudo-cls) techniques;
  ii. bayesian techniques:
    a. unwanted template marginalization;
  iii. hybrid approaches.
- .

Planck data analysis highlights

- Computationally very heavy;
- Sophisticated, novel signal, image and data processing techniques;
- Diverse and heterogeneous;
- Parallel, very efficient, high performance numerical tools and implementations.

Total power vs. interferometric experiments

- Radiometer based interferometers mean "simpler" time domain data and processing (at least in CMB, e.g., DASI and CBI);

- The total power techniques (like these developed for Planck) have often straightforward applications for the interferometers, where modes of the (u,v) plane replace the sky pixels, e.g., DASI:

  in particular,

  i. map-making techniques;

  ii. power spectrum estimators (with fancy template marginalizations);

  iii. component separation/point source extraction algorithms;

  iv. non-Gaussianity tools.

- .

Planck/CMB data analysis at APC

- ~10 researchers plus students and (software) engineers;

- with interests and an active involvement in different (and basically all) aspects of the data analysis:

1. infrastructure (data bases, I/O environments, etc);
2. Time domain processing;
3. Map-making, image processing, component separation, etc.
4. Power spectrum estimators and non-Gaussianity tests.

- And their scientific exploitation.

- Involved in a number of past (e.g., Boomerang, MAXIMA, Archeops), current (e.g., QUAD, Bicep, APEX) and future (Planck, Brain, EBEx …) CMB experiments.

Conclusions

- The Planck data analysis will produce a legacy of an entire slew of methods, algorithms, their implementations and high performance software pieces which can be directly for some of the scientific goals of the LOFAR project.

- A 'natural' synergy seems to possible between the data analysis pipeline and software development for the Epoch Of Reionization project of LOFAR and Planck. This work in the context of LOFAR could potentially benefit from the Planck DA experiences.

- Planck will be a "WMAP of the E-mode" pinning down all the details of the E-mode reionization bump down to sampling variance limit.

- A potential synergy between the science goals and science exploitation of the Planck and LOFAR data sets aiming at constraining the nature and details of the cosmological reionization.