UNIVERSITÀ DEGLI STUDI DI FERRARA
- EX LABORE FRUCTUS -

HPC case studies for the CMB

Luca Pagano
University of Ferrara & INFN

October 13th, 2022

CMB Science: Overview

- Earliest EM image of the Universe
- Tons of information on both early and late universe
- Main effort on studying temperature anisotropies and polarisation
- Information from temperature at large and intermediate scales exhausted
- Current focus on:
  - Lensing
  - Polarisation at large and intermediate scales
  - Small scales in temperature
- Huge experimental effort, most promising:
  - Simons Observatory (Ground, 2023)
  - LiteBIRD (Satellite, 2028)
  - CMB-S4 (Ground, 2027-2035)

CMB Science: tiny signal

Signal to study decreased dramatically over the decades:

~ 1K in the ’60s

~ 1μK early 2000

< nK in the current decade

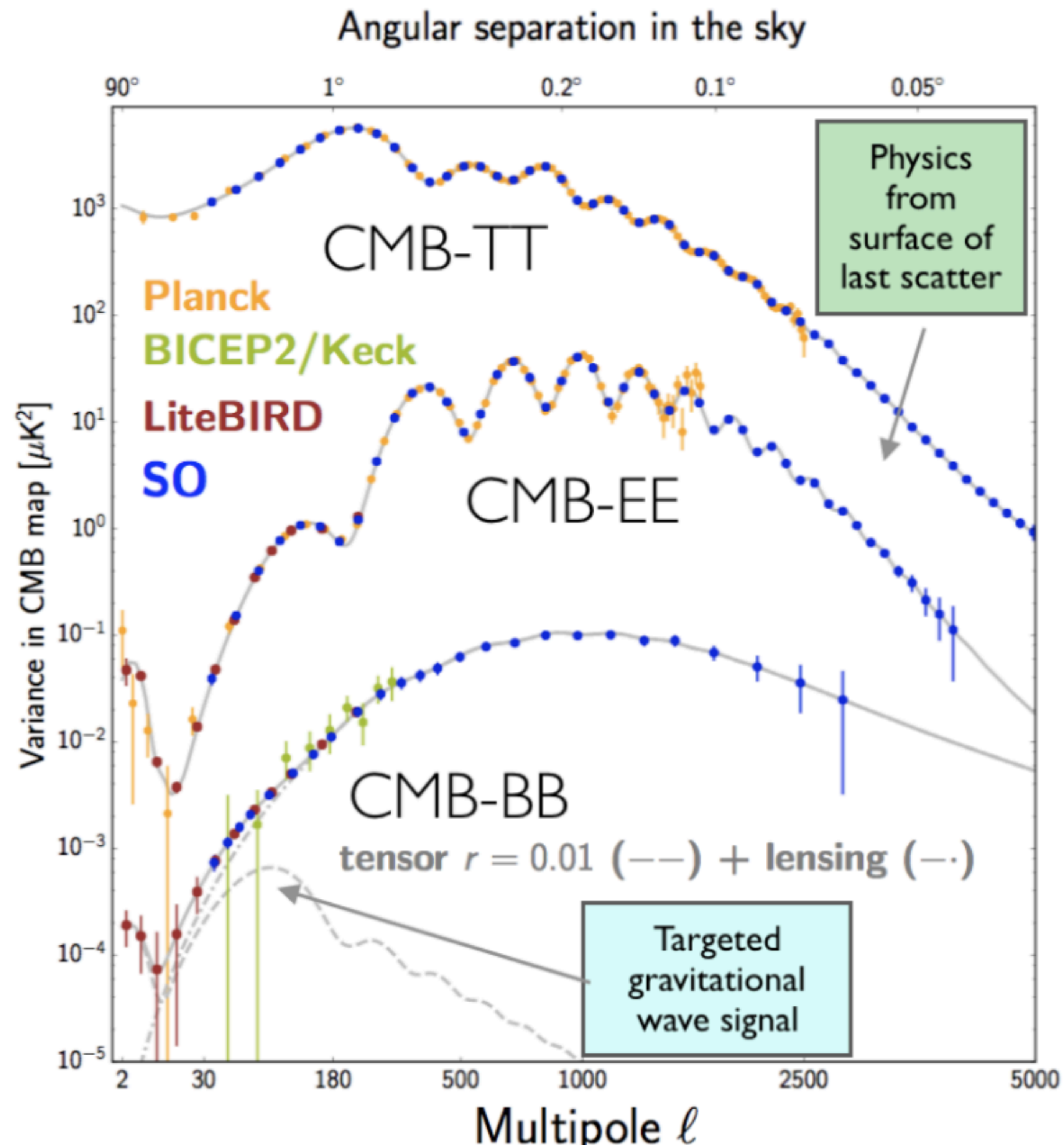
Sensitivity, decades, was reached improving the quality of the detectors

In the last two decades only increasing the number of detectors

~ 1 in the ’60s

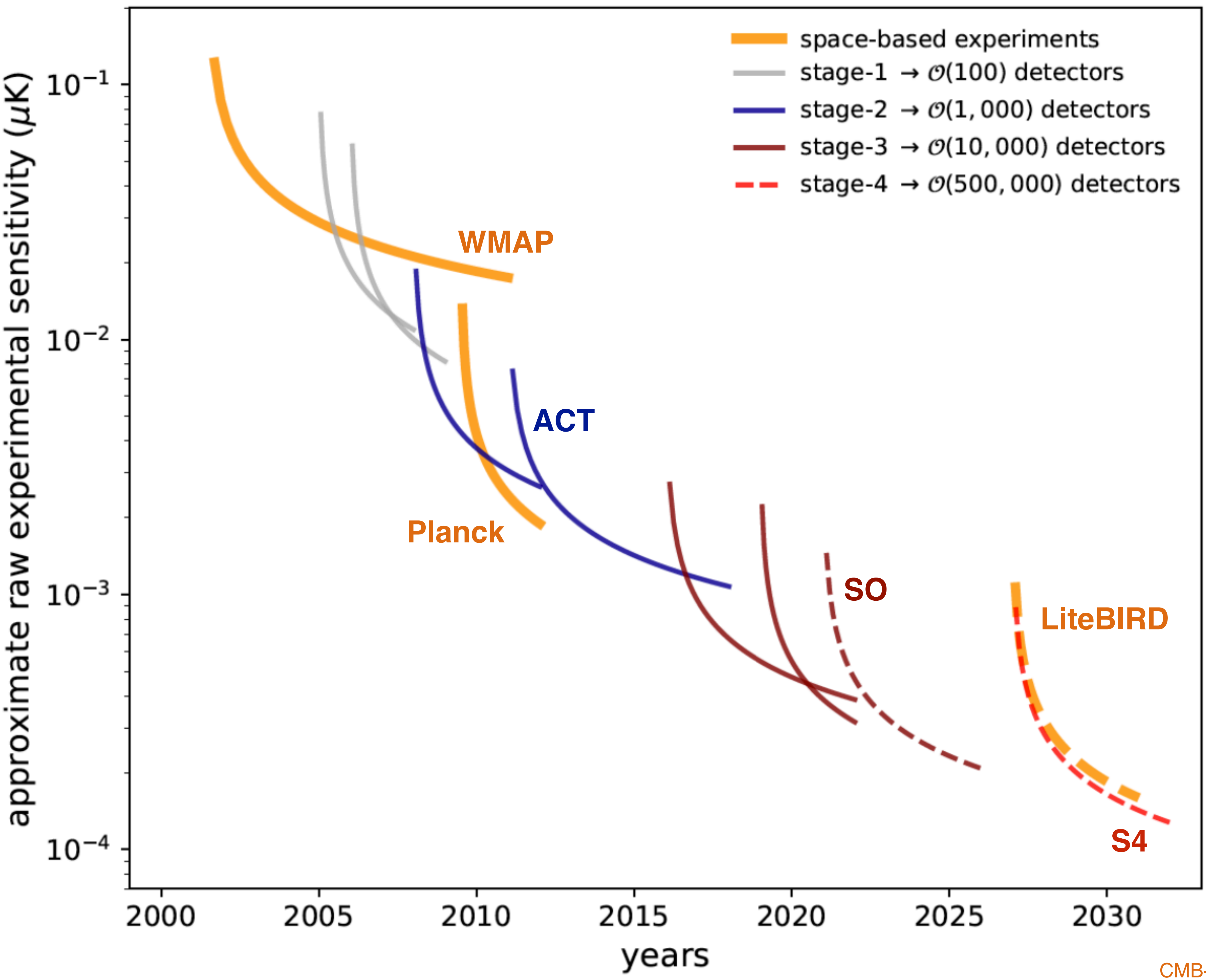
~ 10 - 100 early 2000

~ 1’000 - 10’000 in the ‘20s



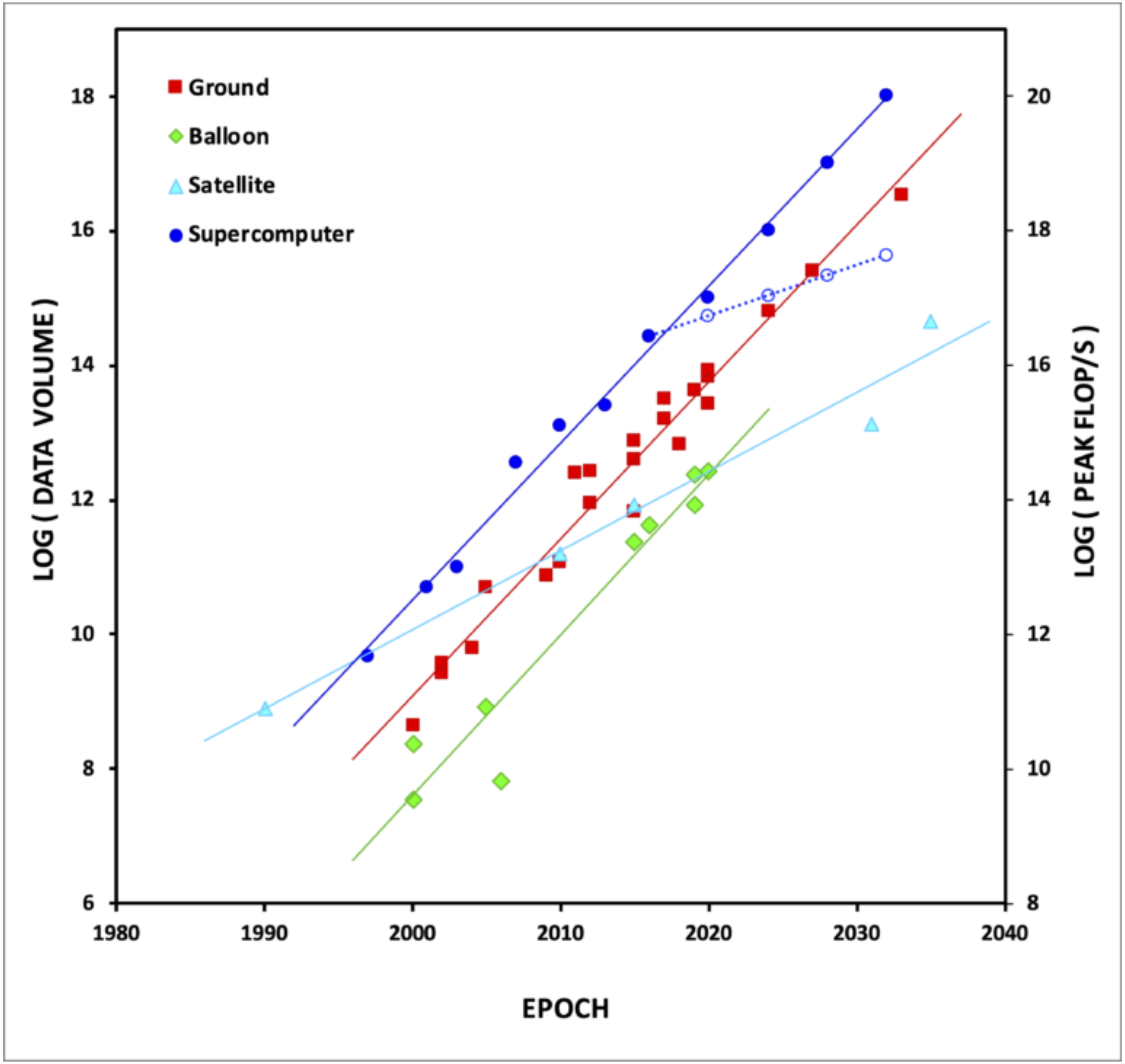
SO collaboration

CMB and HPC: Numerology



CMB-S4 collaboration

CMB and HPC



J. Borrill

CMB data load: some numbers

Data analysis pipelines flow through several domain: time and frequency to pixel and harmonic. Data compression:

- Time domain: $N_t \sim N_{det}$ x Observation time x sampling rate
- Pixel domain: $N_p \sim 3 \times 10^9$ x sky fraction x $(1 / \text{beam in arcmin})^2$
- Harmonic domain: $N_l \sim n_{spectra} \times l_{max}$

Let's give some numbers:

- LiteBIRD: $N_t \sim 10^{11}$ $N_p \sim 3 \times 10^6$ $N_l \sim n_{spectra} \times 10^3$
- SO: $N_t \sim 10^{15}$ $N_p \sim 10^9$ $N_l \sim n_{spectra} \times 10^4$
- S4: $N_t \sim 10^{16}$ $N_p \sim 10^9$ $N_l \sim n_{spectra} \times 10^4$

Analysis scaling dominated by:

- N_p^3 for analytical methods
- MC-size x N_t for approximate methods that use MC of uncertainties

CMB data analysis

Analytic analysis:

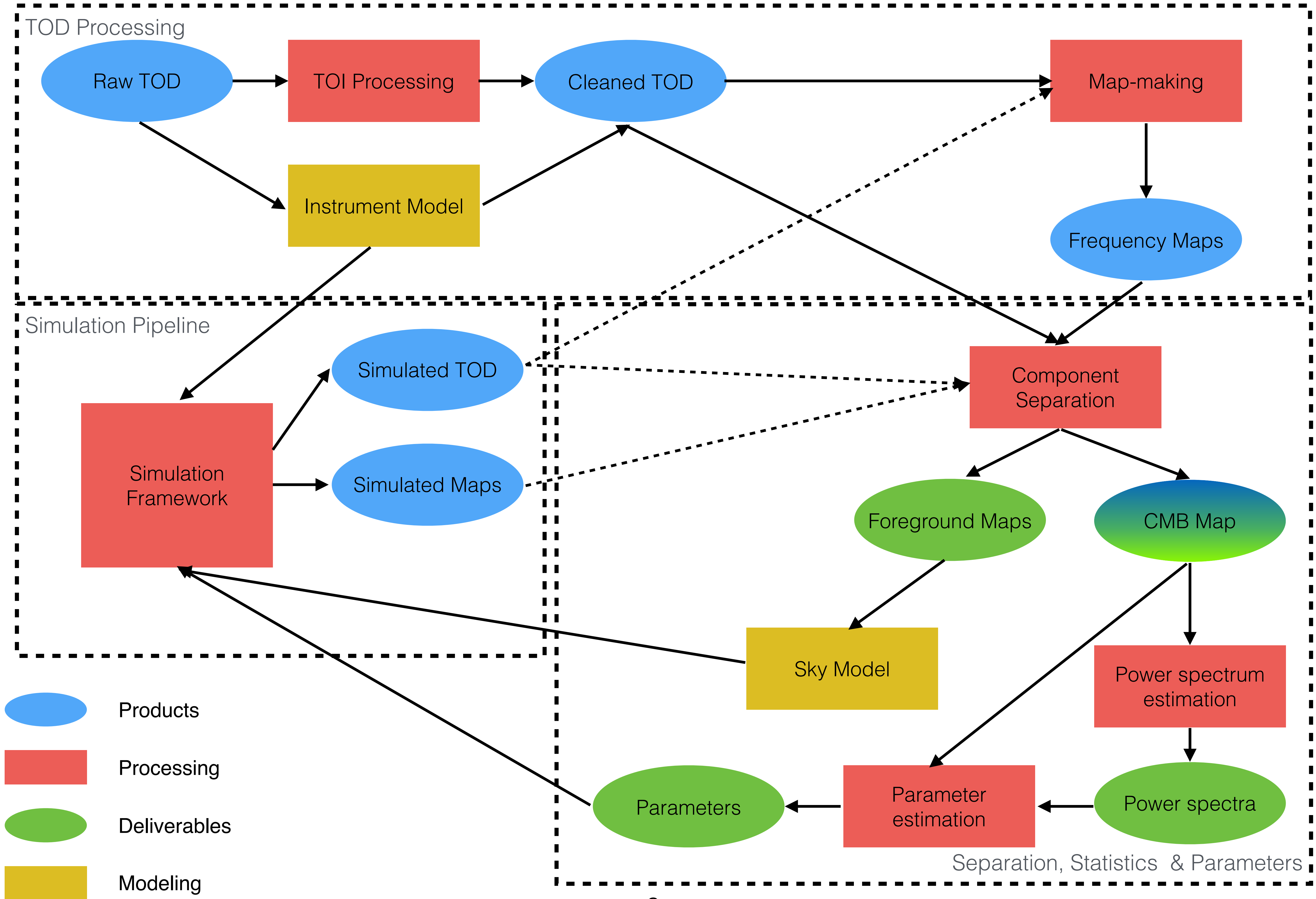
- Maximise Gaussian likelihoods
- High efficiency, close to 100%
- Dominated by matrix inversions, operations scale as $O(N_p^3)$
- Only viable for small patch/low-resolution data
- Totally unfeasible for current generation in full scale
- Unreliable in case of systematic contamination
- Error budget dominated by systematics already for present day experiments. Modulators might help, but to be dominated by gaussian noise is unrealistic
- Foreground residuals difficult to threat

CMB data analysis

MC analysis:

- Monte Carlo replaces explicit covariance
- Low efficiency, less then 1%
- Computational cost dominated by simulation/map making processing
- Scales as timeline length times number of detectors times number of simulations
- Propagating systematics through MC is very costly and not always straightforward
- Can incorporate foreground treatment and systematic marginalisation
- Viable for current/future missions if we can maintain efficiency

CMB data analysis pipeline



DA and HPC: some considerations

Several data processing modules, different requirements:

- For TOD processing several options, potentially HTC problem, new algorithm can help (e.g. NN based).
- PS and Parameter estimation, easiest to threat and with lowest cpu cost, here also new methods can improve the efficiency.
- Map-making heaviest problem, large data load, heavy communication
- Component separation, same as map-making. In the future surveys might be considered a single data-analysis step
- Simulation pipeline. Varied purposes. If used for propagating systematics and in “forward analysis extremely” costly. Emulators might help here. Potentially huge I/O.
- Map-making and simulation require HPC effort.

CMB and HPC: map-making

Map-making step

- Inputs: timelines
- Objectives:
    - collapse timelines in maps
    - reduce low frequency noise
    - marginalise over systematics
    - potentially solve for components
- Method:
    - algebraically corresponds to solving a huge linear system
    - jointly filter low frequency noise and solve for systematic templates
    - jointly solve for components: CMB and Foregrounds
    - architecture tuned for specific experiments but methods fairly general

CMB and HPC: simulations

Simulation pipeline

- Objectives:
  - generate timelines of signal (CMB, Dipole, Gal-foregrounds, ExG-foregrounds,…) and noise
  - include systematics
  - eventually simulate raw TOD processing
  - potentially generate map
- Method:
  - basic generations hugely parallelizable on independent tasks: signal and white noise
  - correlated noise and systematics require large communication
  - largest computational effort for 4π convolution
  - fairly general architecture, but each experiment has specific systematics

Conclusions

- Huge amount of data in the next years
- Analytical methods only feasible for low resolution data
- MC methods viable for current/future missions if we can maintain efficiency
- Approach requires: many cycles, high bandwidth/low latency communication, fast parallel I/O, significant storage.. i.e. high performance computing..
- Not grid/cloud/share at home/etc.
- Interesting cases for CNHPC:
  - Map-making
  - Simulation pipelines