Comp-Cosmic liaison report: algorithmic challenges

Brian Yanny, Debbie Bard

August 10, 2020

Cosmic Frontier Experiments (past,present, and future):

SDSS I-III, DES: Past,Present: Data available for reproc/analysis
DESI: Present/Future: large spectroscopic instrument
VRO/LSST -- large imaging survey
CMB-S4 -- large microwave background survey
LIGO Gravitation Wave Detector and Follow-ups
Redshifted 21-cm 'cosmic dawn' EDGES and followup
Other radio projects (SKA)
Event Horizon Telescope (Black Hole imaging)
Massive spectroscopic surveys (VRO followup)
Space based WFIRST, Euclid, Gaia

HPC vs. HTC (from Wikipedia -- note Cosmic Frontier folks are still new to these terms):

HPC tasks are characterized as needing large amounts of computing power for short periods of time, whereas HTC tasks also require large amounts of computing, but for much longer times (months and years, rather than hours and days).[1] HPC environments are often measured in terms of FLOPS.

The HTC community, however, is not concerned about operations per second, but rather operations per month or per year. Therefore, the HTC field is more interested in how many jobs can be completed over a long period of time instead of how fast.

From:
The CMB-S4 Science Book:
 arXiv:1610.02743

Showing compute capability of supercomputers and data volume and compute needs of CMB experiments

CMB-S4 Data Volume: 10s PB.

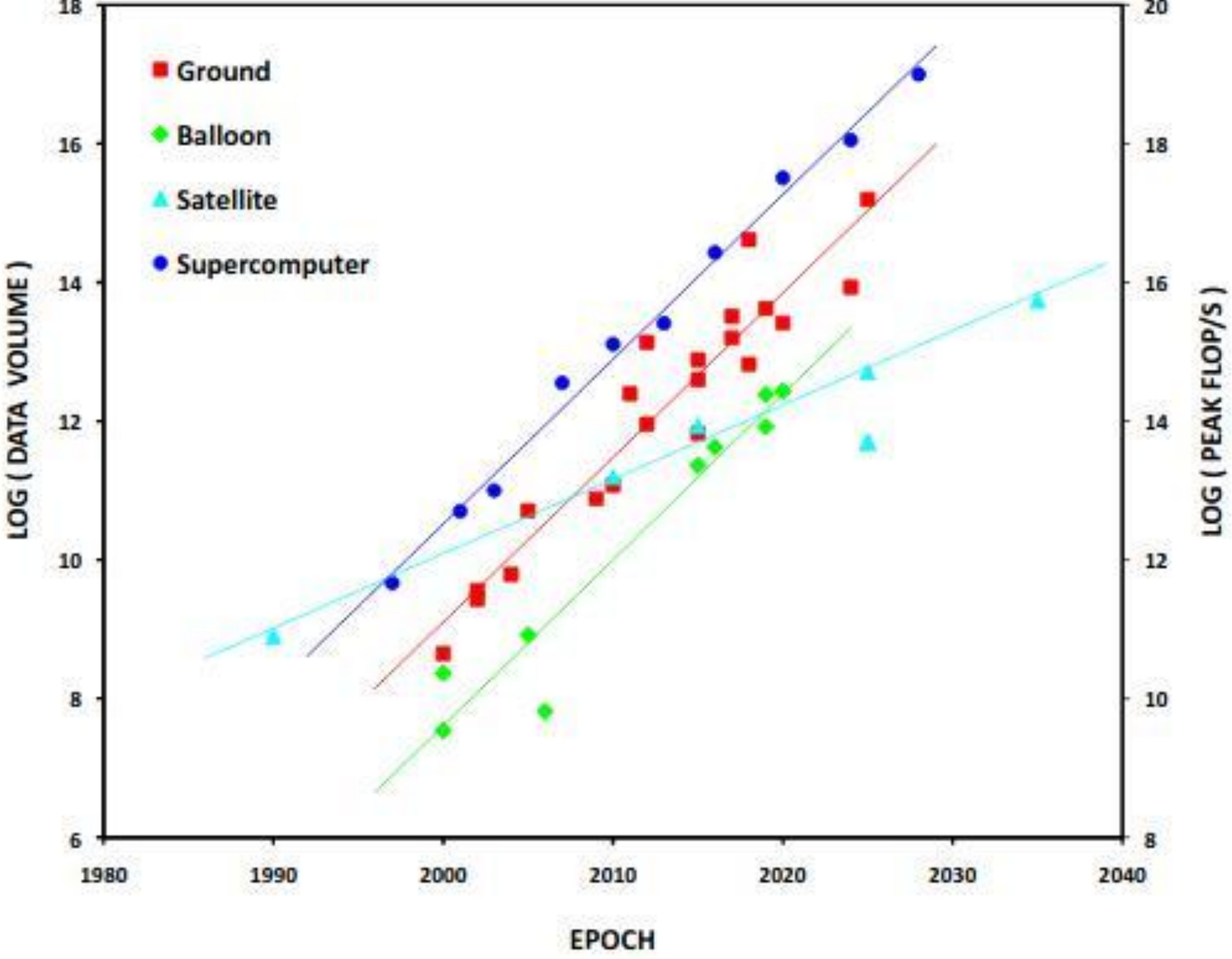

Figure 62. *Exponential growth of CMB time-ordered data volume and HPC capability: 1990 – 2030.*

The quest for ever-fainter signals in the CMB drives us to gather ever-larger time-ordered data (TOD) setsto obtain the necessary signal-to-noise to uncover them. As Figure 62 shows, the volumes of ground-based,balloon-borne and satellite CMB data sets have exhibited exponential growth over the last two decades,and are anticipated to do so again over the coming two. Moreover, for suborbital experiments the exponentexactly matches that of Moore’s Law for the growth of computing capability, where we use as a proxy here thepeak performance of the flagship high performance computing (HPC) system at the DOE’s National EnergyResearch Scientific Computing (NERSC) Center at any epoch (reflecting the widespread use of NERSC forCMB data analyses over the last 20 years).As noted above, in the absence of a full data covariance matrix we rely on Monte Carlo methods foruncertainty quantification and debiasing, and achieving the desired percent-level statistical uncertaintyrequires us to simulate and reduce 104realizations of the data. This implies that all TOD-processing steps (insimulation or analysis) **must employ algorithms that scale no worse than linearly** in the number of samples,and that these algorithms mustcollectivelybe implemented efficiently on the largest high performancecomputing (HPC) platforms available to us

From: https://arxiv.org/pdf/1610.02743.pdf

The **critical challenge for CMB-S4** will be to develop these capabilities for **a dataset 1000x the size ofPlanck’s and 100x the size of those from existing S2** ground-based experiments. This scale of computing willrequire substantial development effort from the CMB community, but is **still much smaller than some existingexperiments (e.g. ATLAS, CMS**) and, with appropriate tooling, should be possible on existing or forthcomingcomputing facilities.

Requirements and opportunities: more on simulation,production,analysis algorithms

- Ray-tracing **simulations** photon-by-photon simulations of sky+atmosphere+telescope mirrors and lenses + ccd detector characteristics

Highly parallelizable but also highly CPU/GPU dependent

CMB-S4: large arrays of bolometers detecting microwave photons from the Cosmic Microwave background -- massive FFT (Fast Fourier Transforms) for **production** of maps

Big MCMC chains ‘error ellipse plots’ and corner plots’ for **analysis** -- require hundreds to thousands of CPUs talking to each other ‘HPC’

Example: https://www.osti.gov/servlets/purl/1412701

Requirements and opportunities: more algorithms

VRO/LSST: 0.5 EB of images and 3 PB of catalogs by 2034

Vera Rubin Observatory/LSST: Requirements for both **Production**, **Archive** and **Analysis/End-User-Access** aspects:

Production: Requirement: Need to '**bring the data to the large processing workhorse centers** efficiently with tracking of metadata and not losing track of things or unnessairily duplicating production'.

Archive: Requirement: **Cost-efficient long term storage**, with possibility of re-processing from 'raw' at a future time as algorithms improve or there are new ideas for new algorithms (i.e. a better weak lensing shear algorithm)

Requirements and opportunities: more algorithms

Analysis/End-user-access: Requirement: Need to bring users ‘**code to the data’** (different paradigm than for production) to allow thousands of users access to same MultiPetabyte data set (catalogs and image cutouts backed by exabyte scale full image dataset) and sufficient access to compute resources where the data is to enable users to run their algorithm on significant subsets ‘bring home heavily processed or sub-selected’ results.

Requirements and opportunities: unique algorithms

AI and Machine Learning applied to large imaging data sets -- **rejecting artifacts** (non-stars, non-galaxies) from images, **identifying outliers-of-astronomical interest** in large imaging data sets (i.e. transients, variables, moving objects) in an automated fashion.