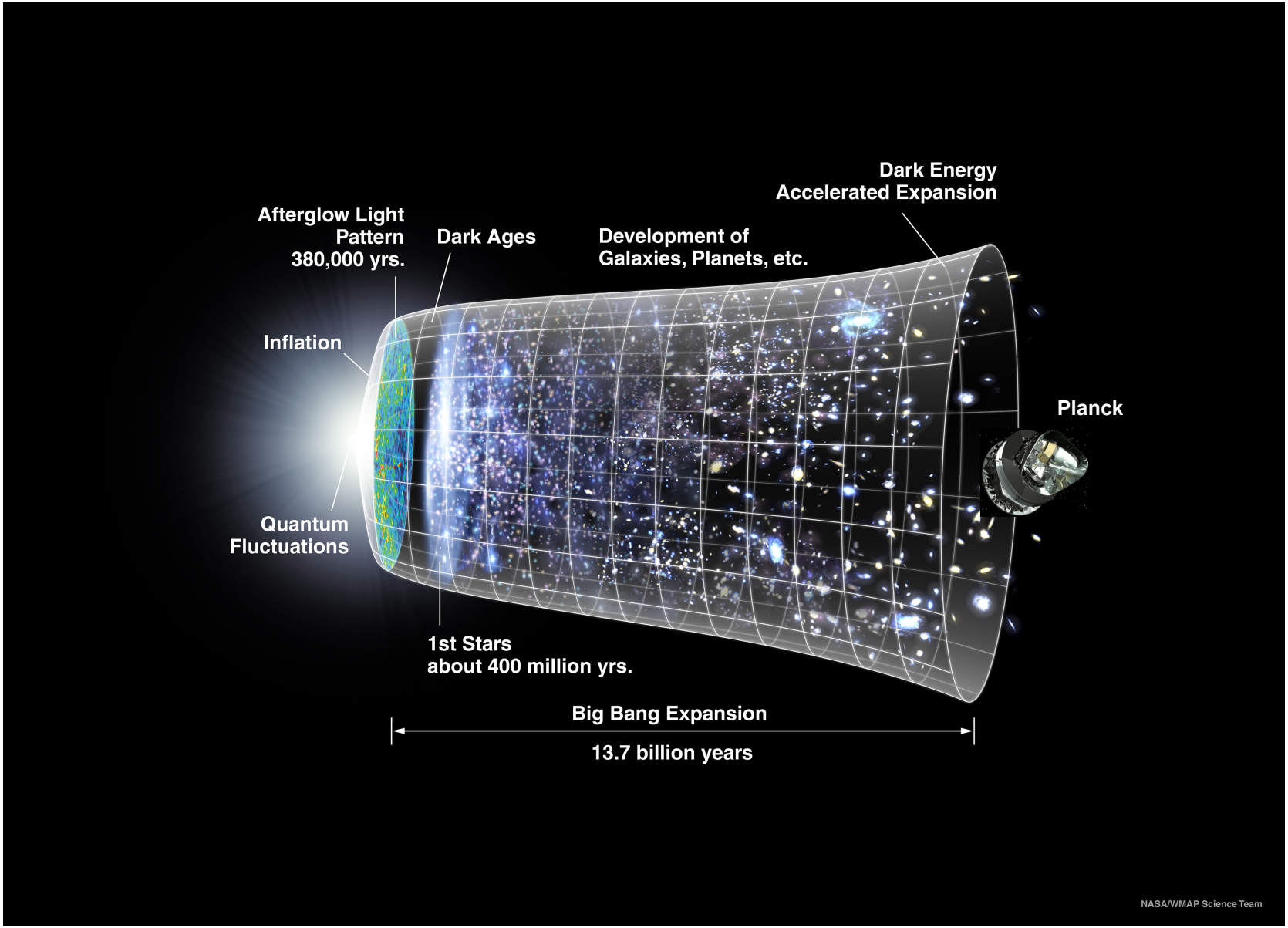
Cosmic Microwave Background Data Analysis

Julian Borrill

Computational Cosmology Center, LBNL

Space Sciences Laboratory, UCB

A History Of The Universe



CMB Science

- Past:
  - Existence => evidence of Big Bang over Steady State
  - Temperature anisotropies => fundamental parameters of cosmology, complementary constraints (SN1a)

- Present:
  - Temperature & E-mode polarization anisotropies => precision cosmology, complementary constraints (DE)

- Future:
  - B-mode polarization anisotropies => measurement of lensing potential, evidence of & constraints on Inflation

2 Nobel Prizes awarded, 1 supported, 1 anticipated!

The CMB Data Challenge

- Extracting fainter signals (polarization, high resolution) from the data requires:
  - larger data volumes to provide higher signal-to-noise.
  - more complex analyses to control fainter systematic effects.

Experiment	Start Date	Observations - N_t	Pixels - N_p	Signal/Noise
COBE	1989	10^9	10^4	10^5
eg. BOOMERanG*	2000	10^9	10^6	10^3
WMAP	2001	10^{10}	10^7	10^3
Planck*	2009	10^{12}	10^9	10^3
eg. PolarBear*	2012	10^{13}	10^7	10^6
eg. QUIET-II*	2015	10^{14}	10^7	10^7
CMBpol*	2020+	10^{15}	10^{10}	10^5

- 1000x increase in data volume over previous & forthcoming 15 years
  - need algorithms/implementations to scale through 20 M-foldings !

CMB Computing Challenges

Dominated by manipulations of time-ordered data

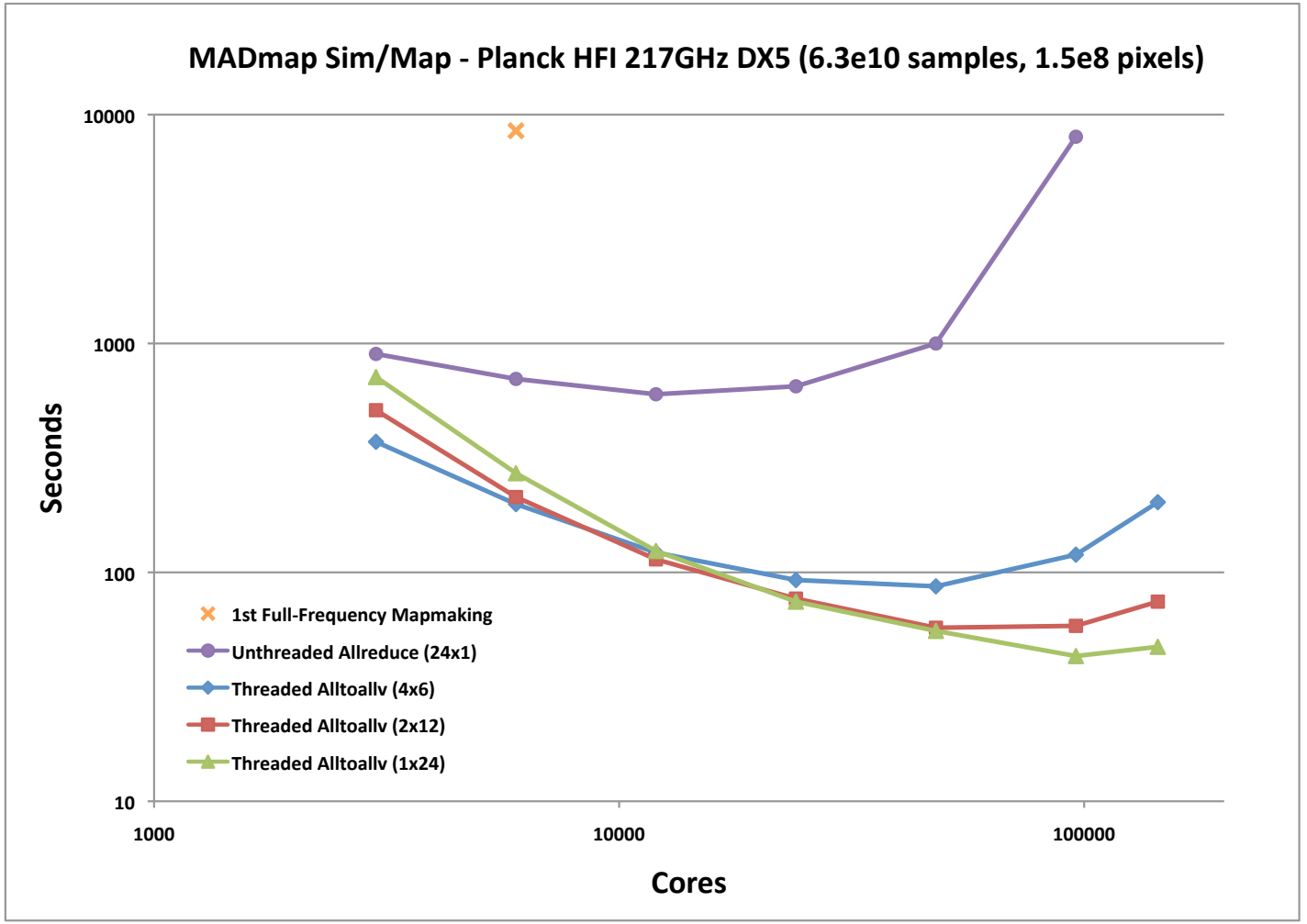
1. Mapping real data:
   - Read/distribute detector data & flags, mission pointing
   - Generate detector pointing by interpolation & rotation
   - Solve for map using PCG over FFT-based matrix-vector multiply
     - Reduce distributed submaps at each iteration
   - Write map
2. Simulating & mapping synthetic data in a Monte Carlo loop
   - Read/distribute detector flags & mission pointing
   - Generate detector pointing by interpolation & rotation
   - For each realization
     - Generate simulated data on the fly using PRNG, FFT, SHT
     - Solve for map as above
     - Write map

Scaling

	Real Map	MC SimMap
Calculation	Iterations x Observations	Realizations x (Simulation + Iterations) x Observations
Communication	Iterations x Pixels	Realizations x Iterations x Pixels
I/O	Observations	Realizations x Pixels
Calc/(Comm+IO) per Iteration	1	Observations/Pixel ~ S/N

- Observations & pixels are properties of the data set
- Iterations depend on the quality of the preconditioner (research in progress)
- Simulation depends on the mission properties (beam asymmetry, noise correlations)
- Realizations are a requirement of the science (10^2 - 10^4 for 10 - 1% uncertainties)

2006-12: 250x Speed-Up



16x Moore's Law (6,000 – 100,000 cores)
16x code optimization (breaking & re-breaking I/O & comm scaling bottlenecks)

Current CMB Work At NERSC

- Multiple suborbital experiments share a single repo (mp107 – 1.5M MPP-hrs)
  - typically O(10) experiments & O(100) data analysts
    - Currently EBEX, PolarBear, QUIET, SPTpol, Spider, etc
  - annual allocation shared by internal negotiation
  - also supports domain-generic algorithm & implementation work
- Planck satellite has two repos (planck, usplanck – 11M + 7.5M* MPP-hrs)
  - planck: standard allocation for O(100) Planck members worldwide
  - usplanck: dedicated resources for O(50) US Planck members
    - initially a stand-alone cluster (256 cores)
    - now a cabinet of carver (640 cores)
      - run as an allocation* with a reservation & dedicated queue
- Also
  - common/public + mission-specific/private software modules
  - dedicated project spaces (including Planck's purchased 100TB ELS)
  - pseudo-user accounts to own/manage the data

The Planck Example

- Planck at NERSC has been a spectacularly successful relationship – possibly a model for other large experiments or possibly domains.
- Under a DOE/NASA MoU
  - DOE provides a guaranteed minimum annual NERSC allocation (2M MPP-hrs) for the lifetime of the mission - though we can and do ask for much more.
  - NASA provides 2.5 FTEs to develop the tools to exploit this resource efficiently.
- Because of the security afforded by the MoU
  - Planck locates its dedicated hardware on the NERSC floor
    - Executed as “exceptional level of service” – even better than ownership!
  - NGF supports both general and dedicated resources – no data replication*
- Because of the long-term Planck/NERSC collaboration
  - Planck is an early adopter (and occasional driver) of new technologies
    - initial NGF buy-in, cluster-to-reservation transition, pseudo-users, etc
  - NERSC regularly provides extraordinary service
    - eg. user boost for collaboration meetings/mission-critical runs

2017 Projections

- We will have 100x the current data volume, but only 10x Moore's Law.
- Detecting fainter signals means more tightly controlling systematics
  - The relatively modest jump to NERSC-7 is a concern – it is essential for NERSC-8 to overcompensate!
  - The breadth of requirements NERSC serves don't map to a single system.

- Data movement at scale has the potential completely to disable us.
  - NGF is essential in a multi-platform environment, but its performance is already a concern.
  - MPI-like communication will still be essential, but its performance at moderate to high concurrency (even at 1 process per node) is a concern.

- Next-generation architectures will require major code re-working
  - Domain-generic elements (algorithms, implementations, shared data) must be consolidated.

Requirements Summary

	Used in 2012	Needed in 2017
Compute Hours	20M	500M
Typical #cores/run	30,000	Full system
Maximum #cores/run	30,000	Full system
Data I/O/run	2 TB in, 5 TB out	50 TB in, 5 TB out
Bandwidth	10 GB/s (hopper)	250 GB/s (ngf)
% IO time/run	35%	35%
Filesystem space	0.2 PB	5 PB
Archive space	0.5 PB	50 PB
Memory/node	1-2 GB	1-2 GB
Total memory	0.5 TB	5 TB

Domain Dimensions

- Computer, computational & physical science
  - multi-focus, overlapping collaborations, overlapping projects/funding
- Capability, capacity, on-demand & experimental computing.
  - different platforms, same data
- Experiment-specific & domain-generic software + data.
  - collaboration of collaborations
- Multi-agency (DOE + NSF, NASA, ESA, JAXA etc)
  - primary responsibility for computing varies with domain/experiment
  - formal responsibility may not match actual resources
- Multi-institution
  - multi-hatted individuals, given DOE lab constraints

The Conundrum

"As a cosmologist:

- want access to medium sized systems with short queues and no downtime
- often would benefit more from high disk bandwidth than super fast interconnect
- appreciate a chance to prepare my software for future systems at my own pace (e.g. Dirac GPU testbed)
- can tolerate longer run times with fewer cores
- would often greatly benefit from more memory per core

As a HPC software developer:

- want access to largest systems ever built
- want access to latest technology that might not even make it to production
- give less weight to disk performance
- can tolerate intermittent system access
- am less likely to exhaust memory due to large concurrency"

Current Challenges

- File-system performance is very unpredictable – very rarely matches its formal specifications.
- NGF is often too slow for production computing, but staging files to Hopper scratch is painfully time- and space-consuming.
- Archiving and retrieving data is slow & inflexible.
- Scheduling does not support the complexity we would like to use.
- Low-hanging algorithm/implementation fruit have (largely) gone.
- Next-generation architectures are disparate, heterogeneous & unpredictable.

One day in October …