Data Management Plan

SPT-3G produces approximately 1.2 TB of uncompressed raw detector data daily, or about 80 TB of compressed data (400 TB uncompressed) annually during normal operations. These data are then calibrated and processed as described in the Data Analysis section of the main proposal to create maps of the sky and power spectra for later science analysis and public release.

Raw Data Archival

Raw data are stored using a custom compressed file format at three sites: University of Chicago, Argonne National Laboratory, and NERSC (Lawrence Berkeley National Laboratory). The first of these copies, at the University of Chicago's Midwest Tier 2 Computing Center (MWT2), is held on spinning disks on a highly-parallel dCache-based file system, which is attached to an Internet2-connected GridFTP gateway for distribution of data to remote computing sites. The second copy is stored on a storage system maintained by the Argonne Laboratory Computing Resource Center and linked to the Crossover computing cluster. This copy is transferred from MWT2, for use as a spinning backup, as well as for access by SPT users on the cluster. The NERSC copy is held on tape as backup in case of a catastrophic failure of the spinning copies at MWT2 or Argonne. The three sites offer both extra levels of redundancy and mitigation of geographic risk by splitting the data between California and the Midwest. Intermediate data products (calibration results, etc.) that can be regenerated from the raw data in the event of such a catastrophic failure are stored at both MWT2 and Argonne. The copies at the national laboratories are provided externally to this proposal; hardware and personnel for the primary MWT2 copy for the expected duration of the project, including replacement disks for anticipated failures, are funded through this proposal.

The custom file format used by SPT-3G, based on that used successfully by the IceCube Neutrino Observatory for the last two decades, provides two features to ensure long-term readability of the SPT-3G data. First, the data structures in the custom file format are universally versioned internally, providing perfect forward compatibility of the on-disk format even in the face of changes to the on-disk structure in later phases of the project. Second, every data frame stored on disk, including reproducible calibration intermediates, is protected against data corruption by an individual 32-bit CRC checksum unconditionally verified by our data processing software at every read. In the event of corruption, individual files can be restored quickly from the offline tape backups at the national laboratories.

During transport of data from the Pole, both by satellite and hard disk transport, these checksums are verified on both ends of the connections and the raw data at Pole retained until at least two verified copies exist at our Northern archival sites. This prevents any risk of silent corruption during transport. Polar data taken during the winter prior to Northern archival are maintained on a set of double-parity RAID arrays with block-level SHA256 checksums in addition to the CRC32s. The disk arrays can withstand a large number of simultaneous failures and automatically replace failed disks in the event of either complete or persistent block-level failure with disks from a hot spare pool.

Software for Raw Data

The core data processing and IO software for SPT-3G, including the implementation of our custom file format, are provided as open-source tools under a 2-clause BSD license on the CMB-S4 GitHub workspace. This software is being baselined by the CMB-S4 experiment, which will help ensure long-term readability of SPT-3G data after the completion of the project. The software has also had significant uptake by the POLARBEAR, Simons Array, and Simons Observatory collaborations, which both provides a benefit to the community and shares the maintenance burden to some degree.

Public availability of this software also makes certain steps of our analysis externally reproducible. In particular, public availability of our core software framework will allow higher-level tools implemented in that software framework to be released externally, allowing sharing of methodology and reproducibility of those aspects of our high-level data releases by third parties.

Release of Final Data Products

The final data products from this work will be maps of the sky in our bands centered at 95, 150, and 220 GHz in temperature and polarization, CMB power spectra constructed from those maps, maps of the lensing potential, maps of the Compton-y parameter and other component maps, catalogs of galaxy clusters emissive sources, thumbnail maps and lightcurves of transient sources in the field (see main proposal, section 4). Because the size of these data products does not scale with detector count, maps and power spectra will not be meaningfully larger than those produced by SPTpol and SPT-SZ, and we plan to continue using the same techniques: publication on the NASA Legacy Archive for Microwave Background Data (LAMBDA) and as part of supplemental materials for publications on these topics. These data releases will be highlighted on the collaboration website and made available simultaneously with publications. Thumbnail maps and lightcurves for transient sources (new as of this proposal) will be made available via an interactive interface hosted on a public server maintained by the NCSA. Policies for reuse, including NSF support acknowledgments and citations, will be posted with the data. Maps will be released in FITS format in both flat-sky and HEALPIX projections.

Reproducibility of Results

Starting with SPTpol, we began a policy of tagging software versions and archiving data products associated with our publications. This is designed to allow at least internal reproducibility of all publications from the raw instrument data. We have continued to implement this policy with SPT-3G.