Data Management Update

Julian Borrill

CMB-S4 Collaboration Meeting
April 3-6, 2023

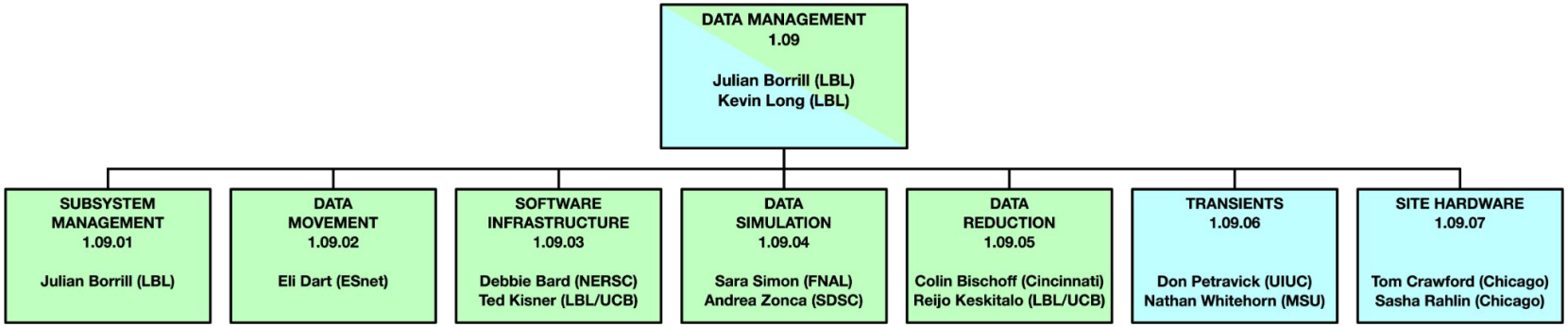
Data Management Charge & WBS

By the start of commissioning, deploy and test all the infrastructure necessary to

- Receive raw data from DAQ at each Site
- Deliver reproducible science-quality data products to the Collaboration & Community:
  - Daily: map each observation, assess data quality, identify transients, issue alerts.
  - Annually: reprocess to science-grade maps & metadata, distribute to the collaboration.
  - Periodically: release well-documented science data products & software to the community.

During construction, use these tools to support the experiment design.

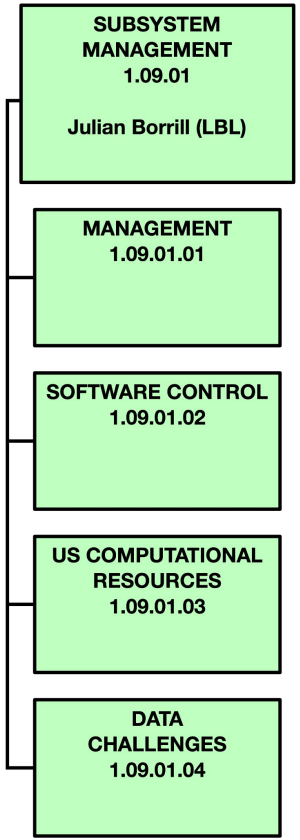
Each L3 presenting here progress in the last year and how/where to get involved



1.09.01
Subsystem Management

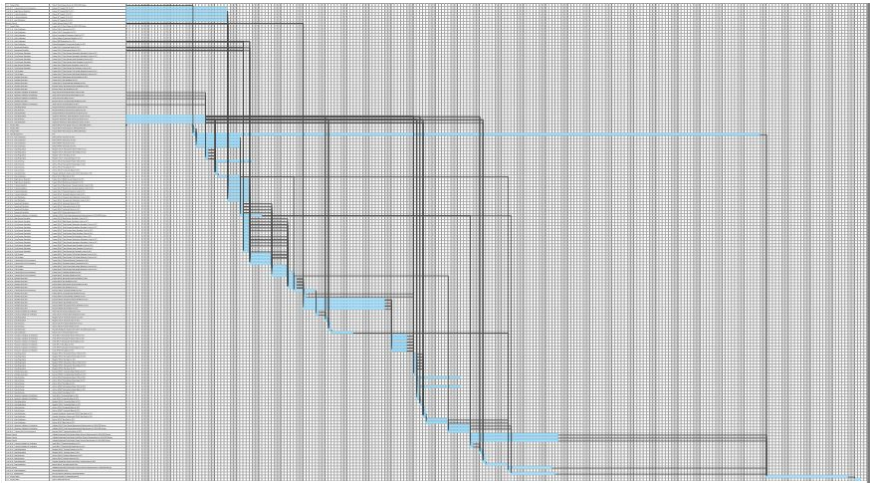
1.09.01 Overview

- Who
  - Julian Borrill (Scientist)
  - Kevin Long (Cost Account Manager)
- What
  - Manage subsystem schedule, cost, risk
  - Organize training events & hackathons
  - Define software standards/controls
  - Acquire US computational resources
  - Coordinate data challenges



1.09.01 Recent Highlights

- Leading roles in Analysis of Alternatives
- Developing a completely revised schedule
  - Reflecting impact of AoA
  - Capturing all dependencies
  - Aligned with new review gates
  - Alternating phases of
    - Design, development and deployment
    - Data challenge to validate point design
- Coordinating Data Challenge 0
  - See tomorrow's session
- Managing NERSC allocation issues
  - Post-Moore's Law scarcity + significantly increased demand from DOE projects



Waterfall schedule for Data Challenge 0, from point design freeze to OPA/CDR review

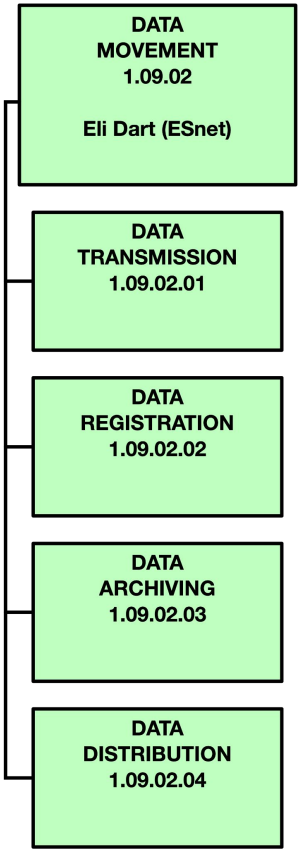
1.09.01 Opportunities For Engagement

- Mailing lists
  - datamanagement@cmb-s4.org
- Slack channel
  - data-management
- Telecons
  - DM-wide: every other Thursday at noon Pacific
  - Details on calendar
- Feedback
  - What software training would you like to see?
  - What foregrounds & systematics should go into Data Challenge 1?
    - see tomorrow’s session

1.09.02
Data Movement

1.09.02 Overview

- Who
  - Eli Dart (ESnet)
  - Need more engagement!
- What
  - Receiving the raw data from DAQ at the sites and transmitting it to the US
  - Registering/archiving raw and reduced data
  - Distributing data to DM, the collaboration, and the community



1.09.02 Highlights

- Coordinating with Simons Observatory & REUNA on Chile site connectivity, leading to multiplexed path to ALMA and beyond.
- Discussions with HEP experts on data registration/replication/archiving about repurposing their software (eg. RUCIO + AMI from ATLAS and others)
    - Potential French partnership on AMI
- Prototype CMB-S4 data portal stood up to distribute DC0 - see Wednesday
    - Add your Globus account identity to your membership record!

Google Earth image showing planned fiber path from Alma to Cerro Toco

1.09.02 Opportunities For Engagement

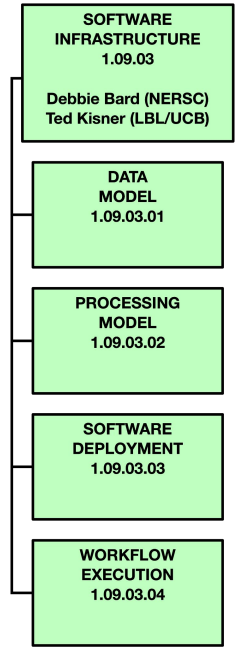
- Huge need for help here, especially in investigating tools like RUCIO for automating Data Registration/Distribution/Archiving
- Contact Eli and/or Julian if you are interested in helping.
  - dart@es.net
  - jdborrill@lbl.gov
  - datamanagement@cmb-s4.org

1.09.03
Software Infrastructure

CMB-S4

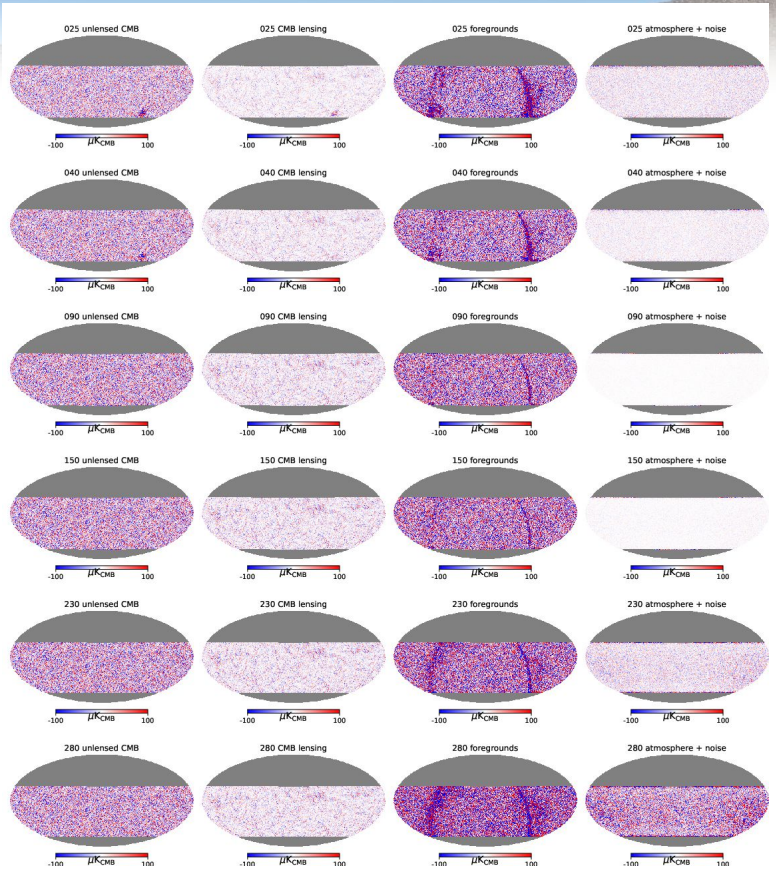
1.09.03 Overview

- Who (points of contact):
  - Debbie Bard (NERSC)
    - NERSC and cross-facility tools
  - Ted Kisner (LBNL)
    - Data formats and simulation / analysis interfaces
  - Reijo Keskitalo (LBNL)
    - Data challenge workflows and execution
  - Nestor Demeure (NERSC)
    - GPU porting
- What:
  - Overall data model: formats and interfaces to data in memory and on disk. Metadata indexing.
  - Overall processing model: APIs for simulation and reduction operations. Construction of workflows and code optimization.
  - Software deployment: Policies for dependencies, releases etc. Installation tools for full stack across computing resources
  - Workflow execution: Job setup across computing resources. Job and data product provenance tracking.



1.09.03 Highlights

- Executed DC0
- Prototyping several GPU porting options for TOAST
- TOAST benchmark for NERSC-10 system RFP
- Implemented FLAC compression of detector data within HDF5, used to archive DC0 data
- Enabled support for observing common simulated atmosphere with more efficient per-wafer domain decomposition in simulation workflows
- Improved internal instrument coordinate systems for better compatibility between CMB-S4 and Simons Observatory



DC0 simulated data across components and frequencies.

1.09.03 Current Opportunities For Engagement

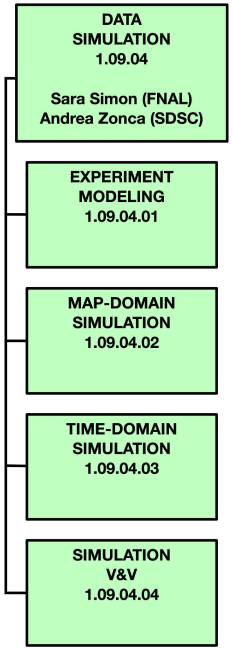
- Extend TOAST / SPT3G interfaces to include new data schema being developed by DAQ
- Performance and archival study of SPT3G and HDF5 data formats
- Expand workflow construction tools for upcoming DC-1
- Beginning stages of metadata and provenance tracking design, including tests running on NERSC spin platform and parsing metadata from existing DC-0 data
- Python packaging (conda / wheel) of remaining core software tools
- Communication on general email list (datamanagement@cmb-s4.org) and new slack channel (#dm-software-infrastructure)

1.09.04
Data Simulation

CMB-S4

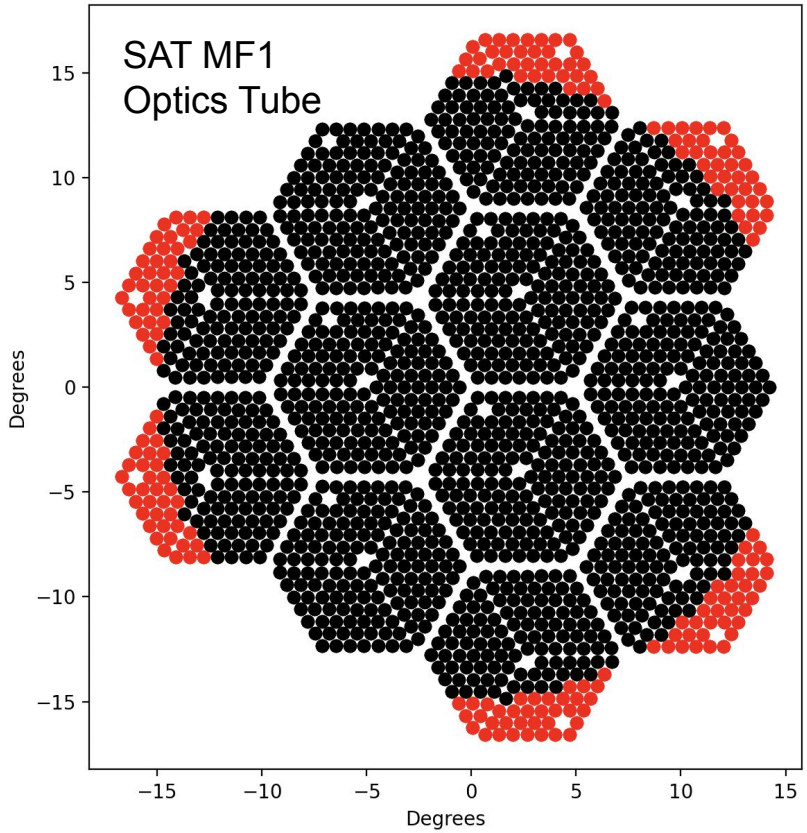
1.09.04 Overview

- Who
  - Sara Simon (FNAL)
  - Andrea Zonca (UCSD)
  - Reijo Keskitalo (LBNL)
- What
  - Experiment Model
    - Instrument Model: full and complete description of instrument point designs
    - Observation Model
      - Sky Model: CMB, lensing, galactic emission, extragalactic emission
      - Observation Efficiency: Includes detector yield, season dates, calibration time, downtime
      - Scan Strategy: Includes target fields, scan speed/accelerations/length, scan style, boresight rotation, Moon/Sun avoidance
  - Map-domain simulations
  - Time-domain simulations
  - V&V: Validation and verification of experiment model, map-based sims, and time-ordered data



1.09.04 Highlights

- Simulations for the Analysis of Alternatives→ 14 instrument + scan configurations
- CHLAT: Completed experiment modeling + map domain & TOD sims for DC0
- Implementing experiment model for SPLAT and SPSAT for DC0 (post-AoA configuration)
  - Dead pixels for mechanical pin/slot
  - Select only “optically good” SAT detectors
  - Updated noise + instrument parameters
- Developed sky model delivery in collaboration with Pan-Experiment Galactic science group:
  - New PySM 3 high resolution models for Dust and Synchrotron
  - Set of sky models at 3 levels of complexity for simulations shared with LiteBIRD / Simons Observatory



1.09.04 Opportunities For Engagement

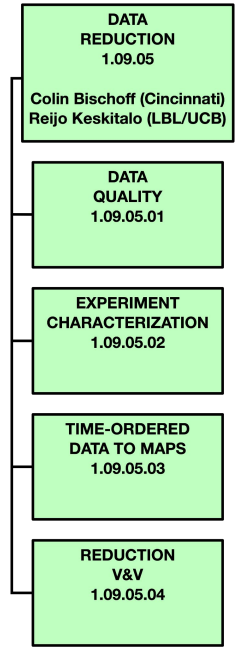
- Upgrades to the instrument + observation models (point person: Sara)
- Observation strategy: cross-cutting group for observation efficiency and scan strategy (surveystrategy@cmb-s4.org, #surveystrategy, point person: Sara)
- Map domain or Time domain systematic studies in coordination with Systematics Working Group and Data Reduction (Systematics group lead: John Ruhl, systematics@cmb-s4.org, time-domain point person: Sara, map-domain point person: Andrea)
- In-depth validation of new Dust and Synchrotron models to guide further improvements (point person: Andrea)
- Validation and Verification Efforts (point people: Andrea + Sara)

1.09.05
Data Reduction

CMB-S4

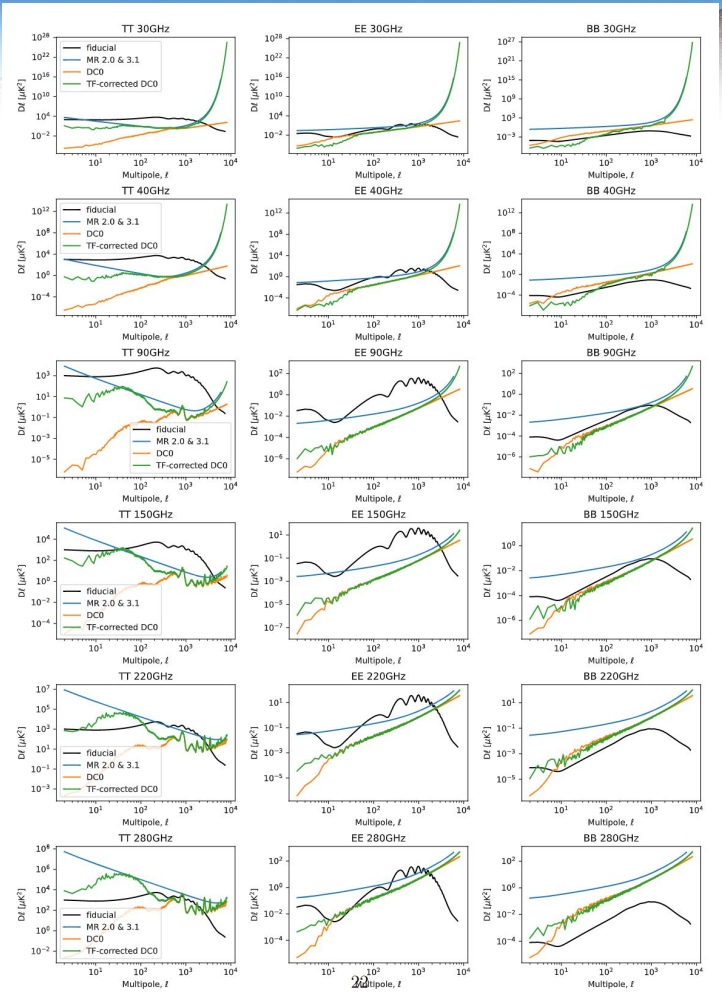
1.09.05 Overview

- Who
  - Colin Bischoff (Cincinnati)
  - Reijo Keskitalo (LBNL)
  - Hamza El Bouhargani (LBNL)
  - Jeremy Webb (Cincinnati)
- What
  - Data Quality
    - Cut statistics
    - Jackknife tests
    - Human interface for data inspection
  - Experiment Characterization
    - Gain and pointing calibration from CMB observations
    - Analysis tools for specialized calibration observations
  - TOD to Maps
    - Comparison of mapmaking algorithms
    - Filtering of time-ordered data
    - Systematics mitigation
  - Validation and Verification
    - Test that maps produced for Data Challenges meet expectations



1.09.05 Highlights

- Configured a filter stack that was performant enough to reduce ~140,000 CHLAT observation maps (5,711 observations, 6 frequencies and 4 signal flavors)
- Working on a database and user interface for data quality statistics collected during the simulation
- Working on the specifics of computing the SPSAT observation matrix



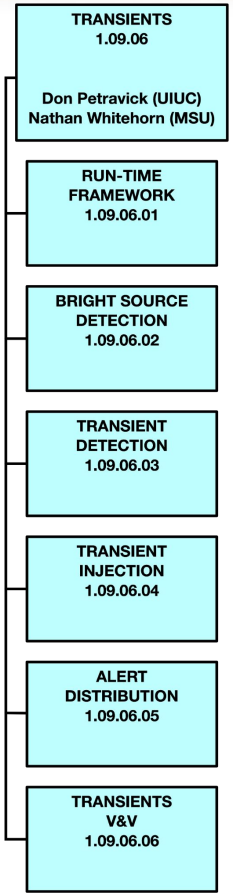
1.09.05 Opportunities For Engagement

- Data Reduction telecons every other Thursday at 8:15 am pacific time (next DR telecon April 20).
- #dm-data-reduction Slack channel.
- Point person for TOD-to-Maps: Reijo Keskitalo
  - Developing, optimizing, and running pipelines for Data Challenges
  - Intend to carry out comparative study of mapmaking algorithms before DC1
- Point person for Data Quality: Colin Bischoff
  - Using TOD and maps from DC0 to explore cut statistics, jackknives, and develop web interface for data inspection.
- Point person for Experiment Characterization: Colin Bischoff
  - Think about simulation/analysis of calibration data and systematics for DC1
- Point person for Validation & Verification: Colin and Reijo
  - V&V of CHLAT, SPLAT, and SAT maps from Data Challenges

1.09.06
Transients

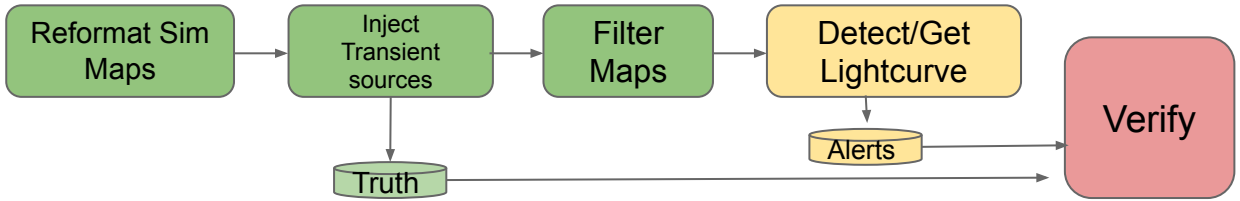
1.09.06 Overview

- Who:
  - Greg Daues (NCSA)
  - Felipe Menanteau (UIUC/NCSA)
  - Don Petravick (NCSA)
  - Nathan Whitehorn (MSU)
- What:
  - Prompt Detection of Transients
    - Issue Alerts to the collaboration and community.
    - Using data from the SP and Chilean LAT's
    - Where Gamma Ray Afterglows is the driving use case for DC0.
  - Support early development by inserting synthetic transients into sims.
  - Build Bright source catalog.
  - V&V: Validation and verification of data challenges

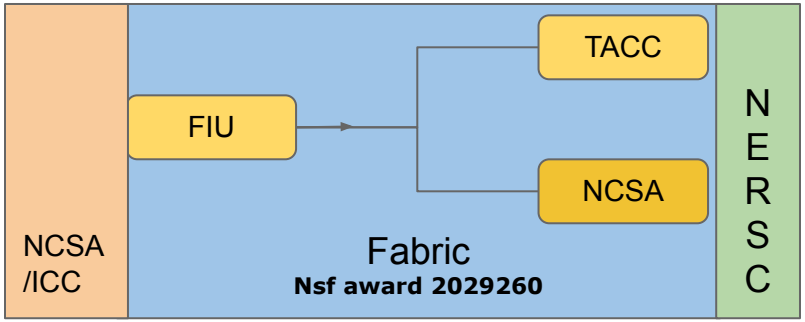


1.09.06 Highlights

- Processing
  - First version of filtering
  - First version of injecting transients into maps
  - Beginning to run detection code.



- Prompt computing on FABRIC
  - Connectivity within Fabric is a “private network”.
  - Florida International University represents the Observatory.
  - NCSA and TACC (Texas Advanced Computing Center) represent alternate processing centers.



1.09.06 Opportunities For Engagement

- Focus -- *On-project* Implementation of…
  - Prompt initial detections of transients.
  - Announcements to community.
  - “Prompt” computing to support this activity.
  - CMB-S4 SLACK: dm-transients channel.
  - Zoom: Friday @1100 Pacific time every other week.

- Focus -- All of relevant transient science .
  - Sources and Transients *Analysis Working Group.*
  - CMB-S4 SLACK: sources_and_transients channel
  - Zoom: Friday@9:00 Pacific time

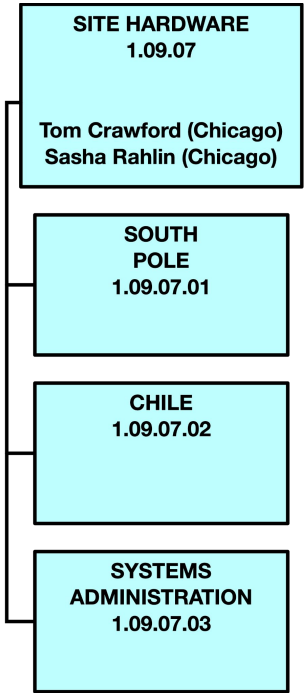
- Check CMB-S4 calendar for on/off weeks changes to ZOOM meetings.

1.09.07
Site Hardware

CMB-S4

1.09.07 Overview

- Who
  - Sasha Rahlin (UChicago)
  - Tom Crawford (UChicago)
- What
  - Design, procurement, installation, testing of computing systems at both sites (Chile and Pole).
  - Scope:
    - Current South Pole scope: responsible for all quick-turnaround computing (including data quality monitoring and transient detection / alert system) at Pole.
    - Current Chile scope: System at Chile (as currently envisioned) mainly a storage buffer.



1.09.07 Highlights

- Fun current projects:
  - Transient AWG thinks it would be super-interesting to detect things and issue alerts on sub-observation timescales (down to ~minutes). What kind of on-site (or in-the-pipe) computing would this entail?
  - The Analysis of Alternatives process appears to have resulted in a “new normal” of smaller footprint at the South Pole. How do we re-design the Pole computing system to reduce footprint (particularly power consumption) while not compromising the science?
    - Example project: Research low-power computing alternatives.
  - ICD with South Pole Site (1.11) and Chile Site (1.10) to fit within power and space constraints
  - ICD with DAQ (1.08) for data handoff at sites.

The SPT-3G on-site computing system currently at the South Pole

1.09.07 Opportunities For Engagement

- Help us design these systems!
- Contact us personally or via the Data Management email address or slack channel
  - Sasha: arahlin@uchicago.edu
  - Tom: tcrawfor@kicp.uchicago.edu
  - datamanagement@cmb-s4.org
  - CMB-S4 Slack: #data-management