SuperCDMS SNOLAB Computing Status and Future needs

Tina Cartaro
SLAC

DANCE Workshop
Rice University, Houston, TX
October 28th, 2019

OUTLINE

- The SuperCDMS Collaboration
- SuperCDMS SNOLAB overviews
  - The Super Cryogenic Dark Matter Search experiment at SNOLAB
  - Offline computing
  - Analysis tools
- Looking forward to the data

The SuperCDMS Collaboration

- ~100 physicists in 25 Institutions, including 3 US National Labs, 2 Canadian Labs

California Inst. of Tech.
CNRS-LPN*
Durham University
FNAL
NISER
NIST*
Northwestern
PNNL
Queen's University
Santa Clara University
SLAC
South Dakota SM&T
SMU
SNOLAB
Stanford University
Texas A&M University
TRIUMF
U. British Columbia
U. California, Berkeley
U. Colorado Denver
U. Florida
U. Montréal
U. Minnesota
U. South Dakota
U. Toronto

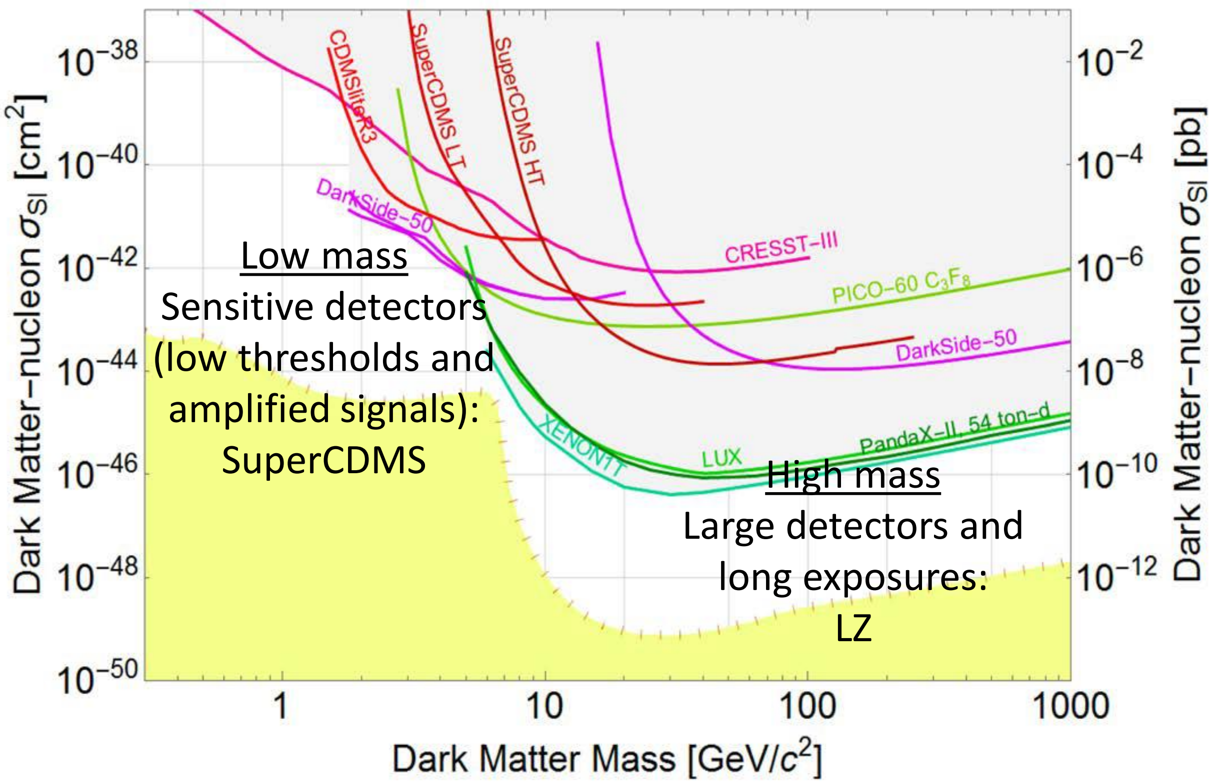
* Associate members

- Funded by:

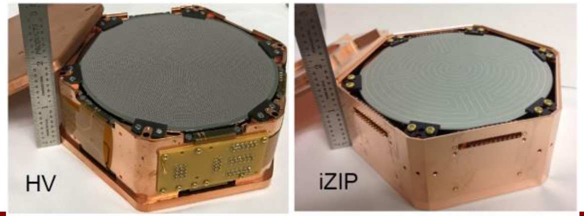
U.S. Department of Energy
NSF
Innovation.ca – Canada Foundation for Innovation / Fondation canadienne pour l'innovation
Arthur B. McDonald Canadian Astroparticle Physics Research Institute
NSERC CRSNG
DAE
Emmy Noether-Programm DFG

Concept for SuperCDMS SNOLAB

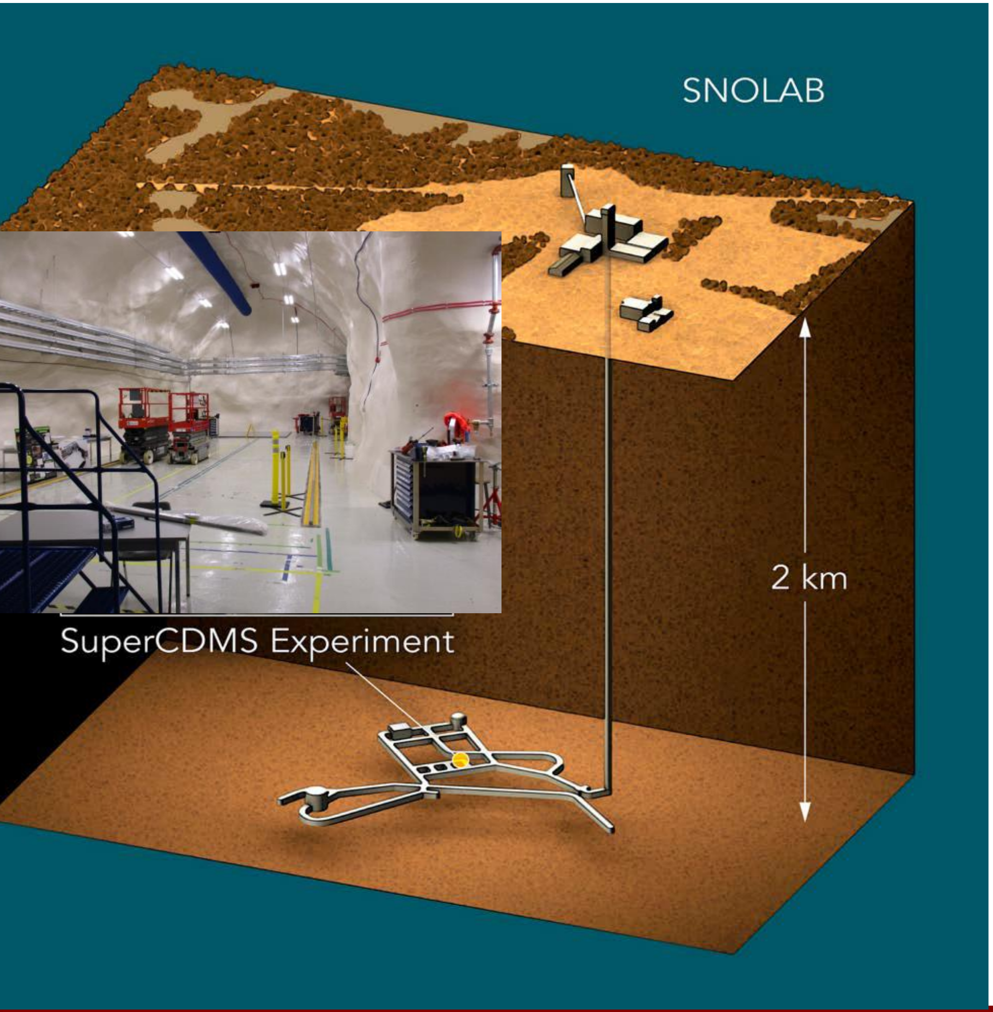
- SuperCDMS SNOLAB is a G2 experiment for direct detection of Dark Matter (WIMP)
- Two possible places where to look



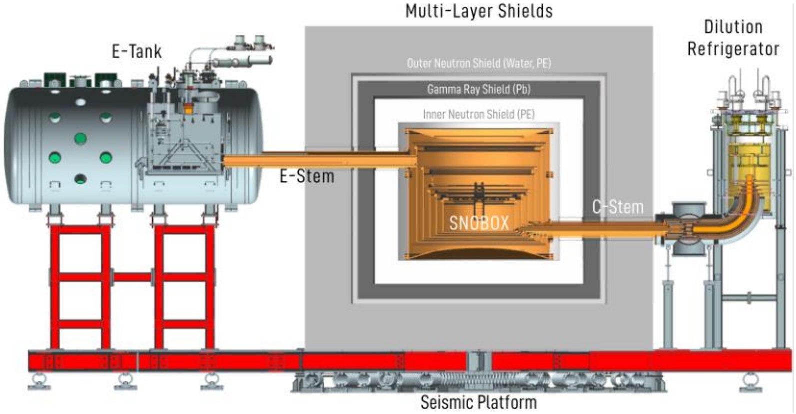
- Ge and Si solid state detectors at cryogenic temperatures (~15mK) shielded against cosmic and environmental radiation
- Two types of detectors, HV and iZIP, optimized respectively for sensitivity to small signals from low mass DM, and background rejection from electron recoil and surface events



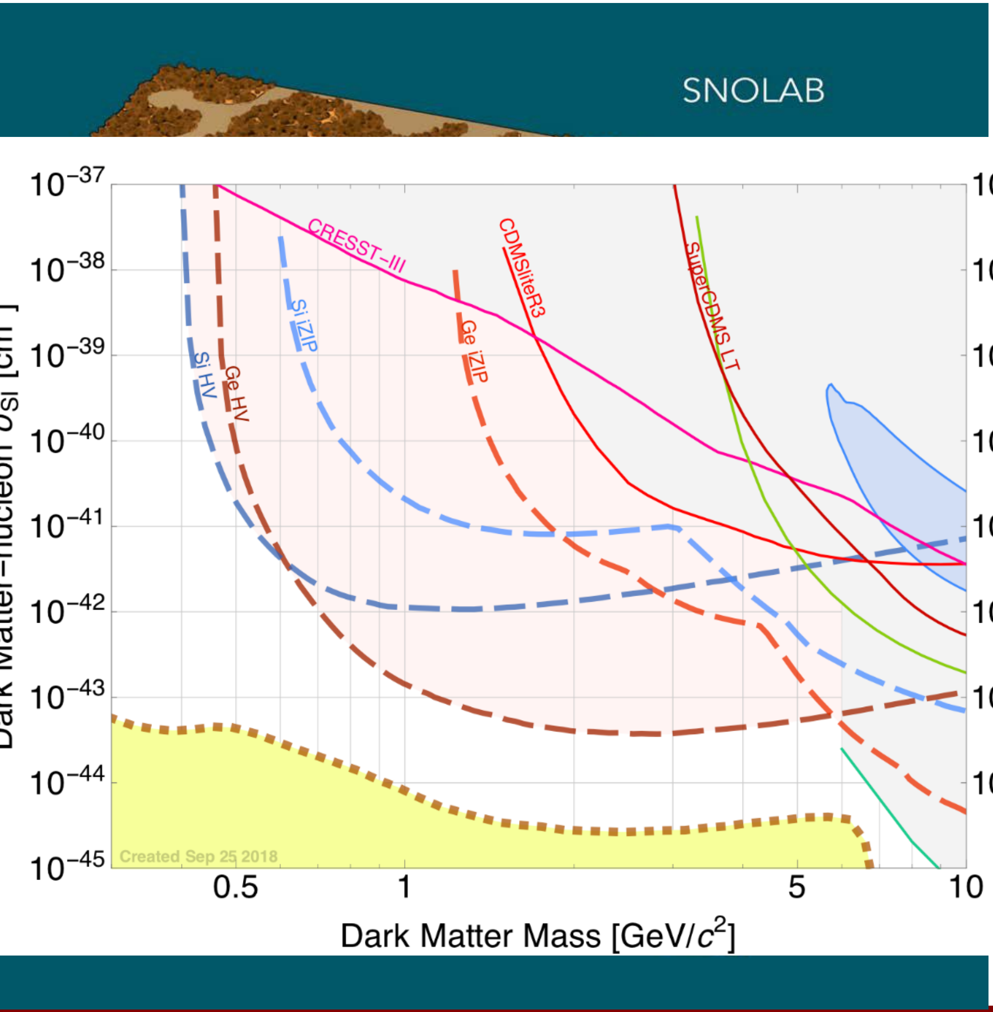
The SuperCDMS SNOLAB Experiment



- Experimental setup at SNOLAB
  - Construction started, data taking expected to start in 2021
- Initial payload: 4 towers, each containing 6 detectors
  - Full payload: 31 towers
- 5 years science mission



THE SUPERCDMS SNOLAB EXPERIMENT

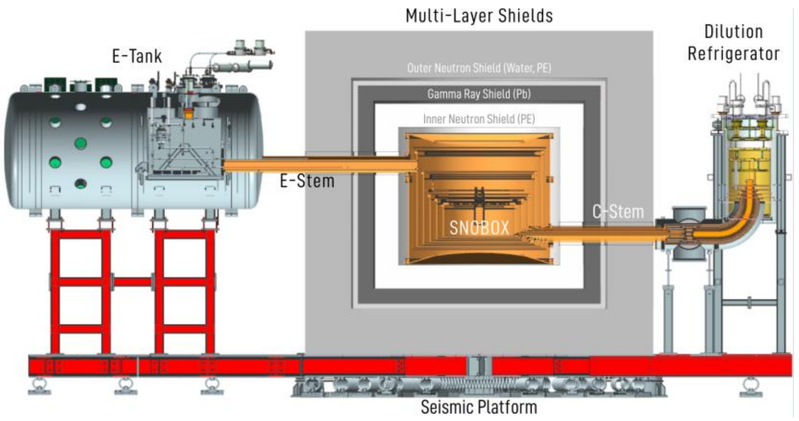


- Experimental setup at SNOLAB
  - Construction started, data taking expected to start in 2021

Initial payload: 4 towers, each containing 6 detectors

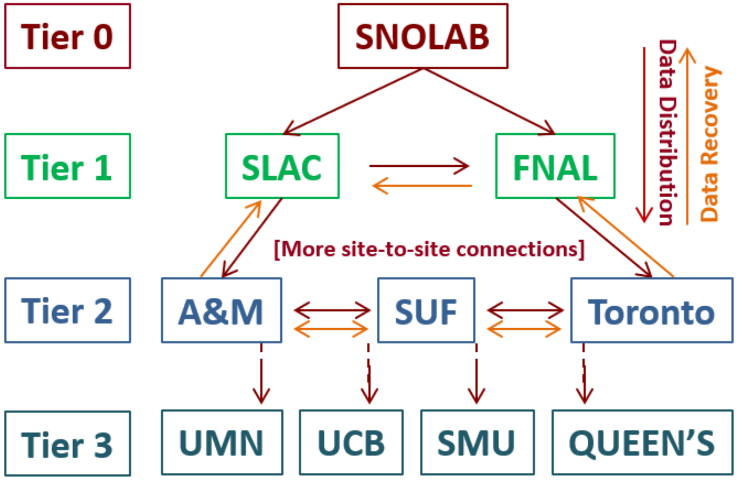
  - Full payload: 31 towers

5 years science mission



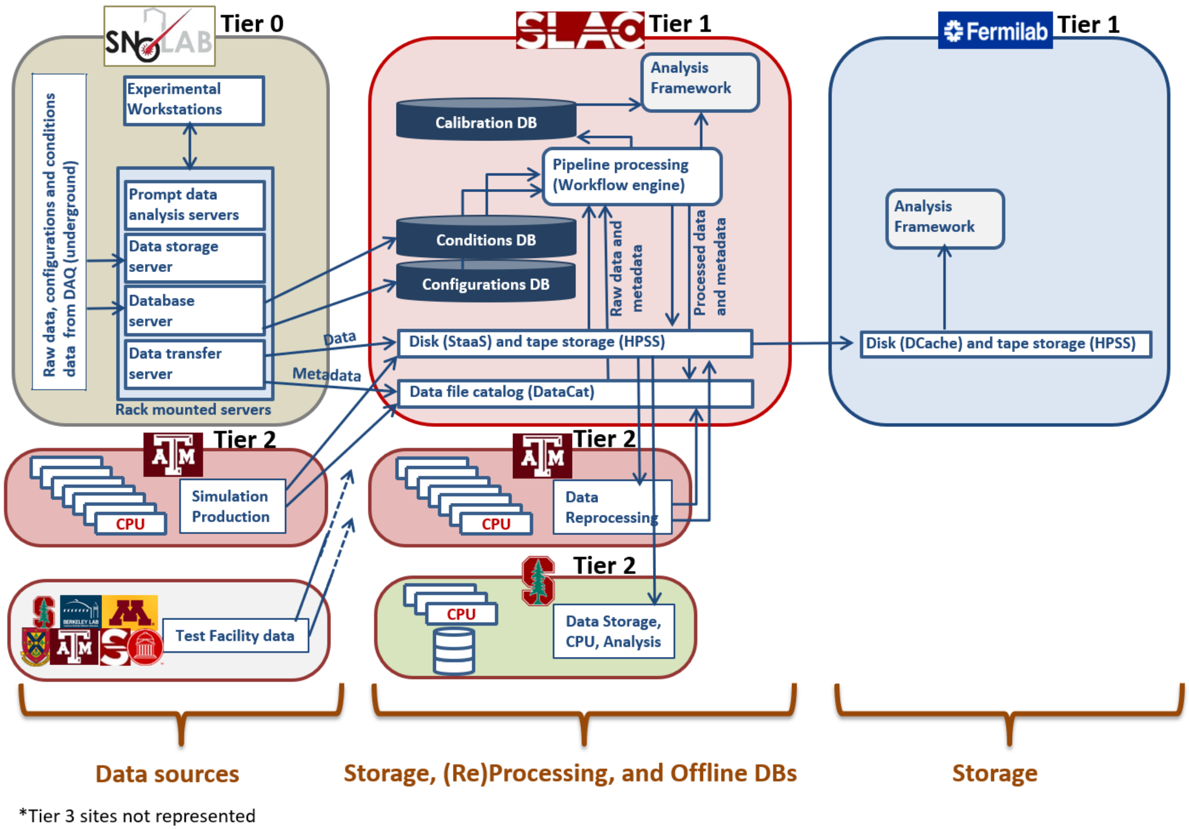
OFFLINE COMPUTING

- Principles for our offline computing
  - Application based for maximum automation
  - Distributed to maximize resource utilization
  - Modular for robustness
  - Scalable to the Petabyte
  - Reuse existing tools and applications whenever possible
- A tiered organization to leverage the resources of the collaborating institutions
  - Tier 0, SNOLAB, will not hold the data on the long term
  - Tier 1, SLAC and FNAL, host a complete copy of all data (real and simulated) and are primary distribution source of data
  - SLAC also supports data processing, data analysis, user accounts, and disk space
  - Tier 2: host a (partial) copy of the data, support either data processing, simulation, or data analysis depending on their capacity.
  - Tier 3: opportunistic resources, support for local groups



- Except for the Tier 1's, this layout of the institutions' roles is just an example
- CERN experiments use a slightly different definition for the Tier levels

DATA PATH



*Tier 3 sites not represented

- Raw data is exported from SNOLAB to SLAC and FNAL for long term storage
- Prompt processing will take place at the SLAC batch farm through existing allocation
- Large scale Monte Carlo production and data reprocessing will take place at data centers that the collaboration has access to and take advantage of the grid where possible

The raw data is in MIDAS format,
the processed data is in ROOT format

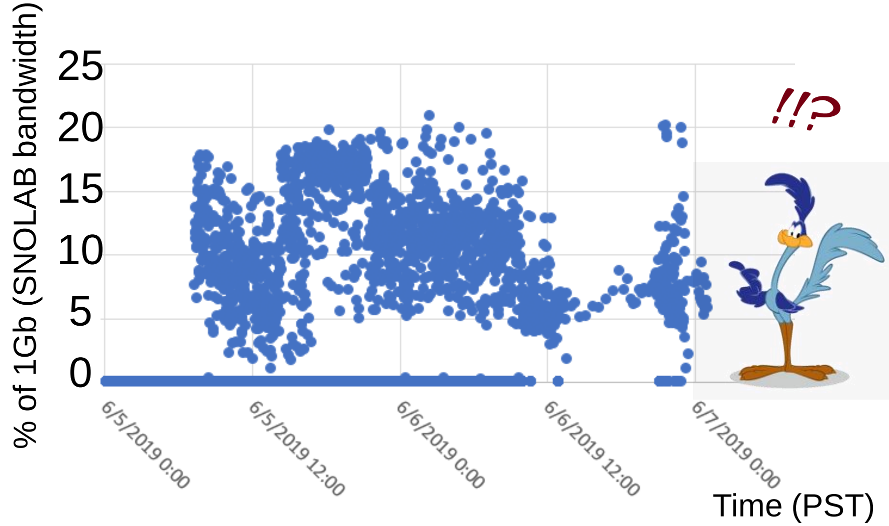
DATA VOLUME AND BANDWIDTH

- Raw data volume estimates:
  - 64-103 TB/year:
    - Low/High noise scenario, compressed
    - gzip/bzip2 gives us a factor 2 reduction
  - Dominated by noise traces for hybrid optimal filter
- Network link from underground to the Surface:
  - 10 Gb/s
- Network link from SNOLAB to Canadian Research Network:
  - 1 Gb/s (= 3.75 PB/year)
  - Shared by all SNOLAB experiments
  - Upgrade to 10Gb/s foreseen in ~2020

Total MC+Data volume estimate:
104-250 TB/year (low/high noise, compressed)

BANDWIDTH TESTS

- From SNOLAB to SLAC single rsync
  - Our data volume is ~4TB/week
  - Single rsync stream uses ~15% of bandwidth with peaks at 20%
    - Use bbcp (multi stream) if needed
  - SNOLAB Disk buffer
    - ~3 week disk buffer in case of network outages or slow network
    - The DAQ will have disk space underground in case of outages to the network link to the surface
- Data transfer between SLAC and Texas A&M regularly ongoing for simulation production
  - Globus Connect python interface



27 GB of data from TAMU to SLAC (single stream):

- rsync: ~32 MB/sec
- bbcp: ~38 MB/sec
- globus online/connect: ~70 MB/sec

81 GB in 3 streams of 27 GB each

- GO: ~150 MB/sec

5 streams of 27 GB each:

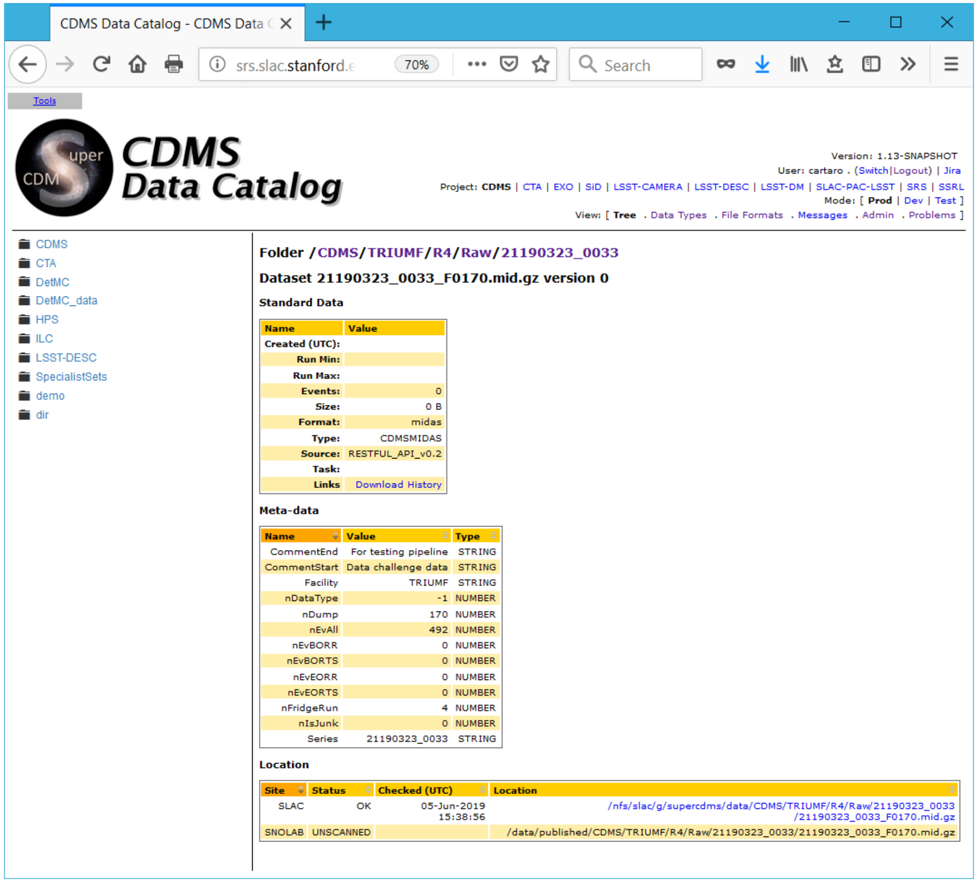
- rsync: ~106MB/sec (individually topped off at ~22 MB/sec)
- bbcp: ~169 MB/sec (individually topped off at ~33 MB/sec)

CPU Needs

- Prompt processing for SNOLAB data at SLAC:
  - ~100 cores needed (our current allocation is almost a factor 2 higher, plus opportunistic use of idle CPUs)
- Single MC event:
  - 4 CPU s/eV energy deposit in an iZIP, about 10 CPU s/eV in an HV or CDMSlite detector Vs. 17s for a data event
- Data re-processing and large-scale MC production:
  - Need a lot of CPU to finish in a timely fashion
  - Run through existing allocations and opportunistically wherever we can
  - At SLAC, during a reprocessing effort it is possible to shift CPU resources from the larger experiments for a limited amount of time
  - Significant resources at Texas A&M routinely used for SuperCDMS Soudan MC production

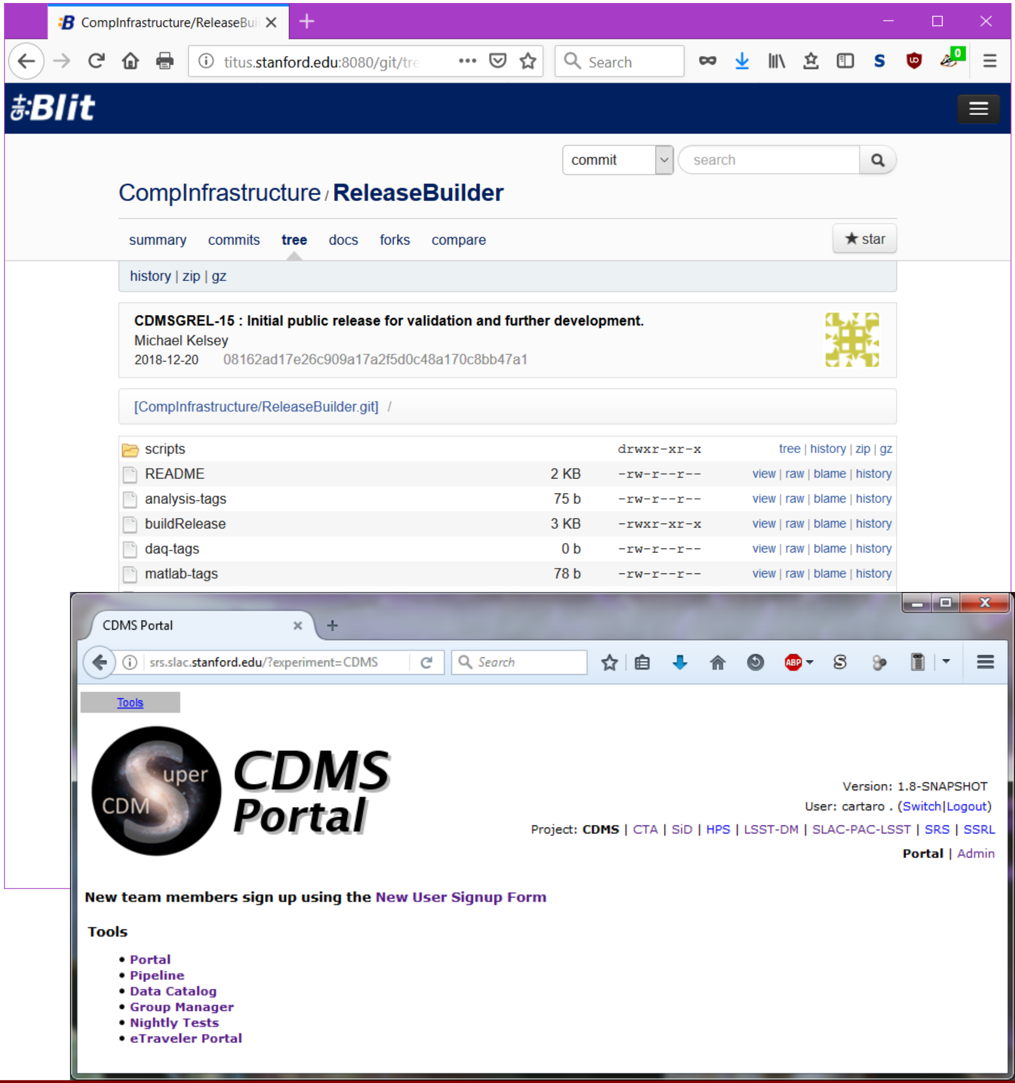
PROCESSING AND BOOKKEEPING AT SLAC

- Batch jobs scheduled through a workflow engine, the (Fermi-GLAST) Pipeline, for data processing
  - Automated bookkeeping, roll-back of failed jobs
- File bookkeeping through Data Catalog
  - Multiple locations
  - Download tool
  - Metadata search
  - Python API
- Question
  - Long time support?



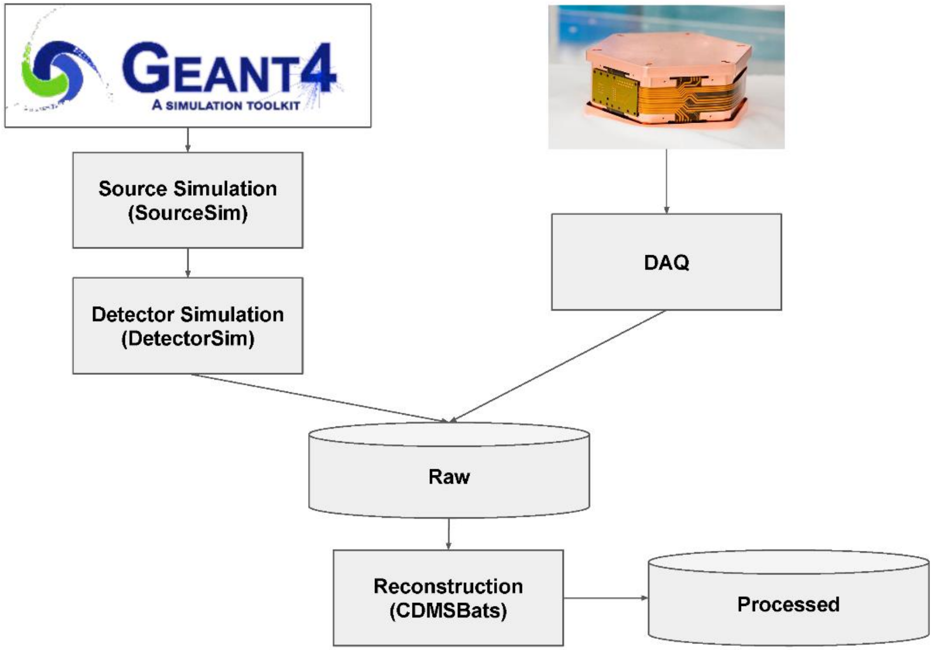
SOFTWARE RELEASE

- The SuperCDMS software is based on few large software packages (CDMSBATS, CDMSTOOLS, SuperSim, DMC) with loose compilation or runtime dependencies
  - Main packages are in C++ and/or Python3
- A private GitBlit repository is used for version control
  - Installed at the Stanford Underground Facility cluster
  - Considering migration to GitLab
- Software Release Builder implemented for reproducibility and for package compatibility
- Continuous Integration software: Jenkins
- Jenkins, Pipeline, and Data Catalog are all interfaced together (CDMS Portal)
- Code distribution: CVMFS at FNAL



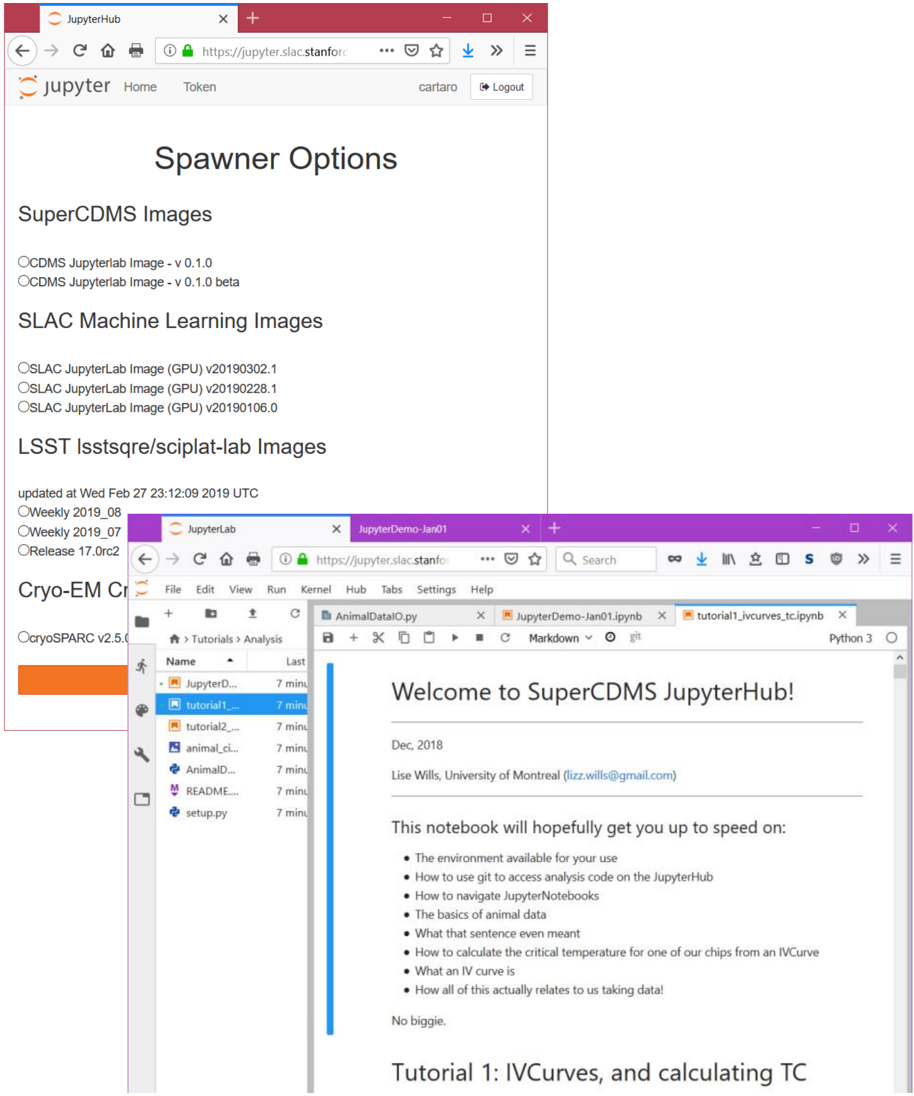
SIMULATION

- Use the Geant4 toolkit, and custom Condensed Matter Physics tools (G4CMP) to simulate the detector
- Goal: All the steps needed to fully simulate detector/readout response to create events in Raw data format
- Process events using the standard Reconstruction tools



JupyterHub @ SLAC for Analysis

- The Software Release environment is going to be embedded in a SuperCDMS singularity image
- The Software Release itself is independent from the image
- The image, preconfigured with the existing SuperCDMS environment, is available for data analysis through the SLAC Jupyter Hub
- Python3 is the primary analysis language and we're looking at pandas DataFrame objects for the default data representation
- Tutorials and documentation available



A Look Ahead

- Always keep an eye on infrastructure and tools
  - Long term support of applications developed/maintained outside of the collaboration (Pipeline and DataCat, GlobusConnect) are not guaranteed
  - GEANT 4 expertise is quickly disappearing
  - Expert computing and software manpower is generally insufficient
    - Also difficult to face national labs infrastructure changes when they happen
  - JupyterHub at SLAC is only in beta and we're definitely testing its limits
- On the bright side, even if the data taking will start in 2021, several test facilities are already operating, and we have most of the infrastructure that we need in order to manage their data (in most cases it is perfectly good DM data)
  - IMPACT@TUNL (neutron beam, HVeV detectors), NEXUS at FNAL (HVeV, HV), CUTE at SNOLAB (Detector characterization, HV, iZIP)
- Looking forward to exciting new science and hopefully to new possibilities for computing to support that science!