Rubin Observatory USDF-at-SLAC Update and HTCondor-at-NCSA 2023

Throughput Computing 2023, July 12, 2023

Greg Daues, NCSA

Vera C. Rubin Observatory Overview

- Rubin Observatory construction continues
  - Installation at Cerro Pachon, Chile
  - `Telescope Structure' complete
  - On track to receive the observatory's 8.4-meter mirror, 3200-megapixel LSST Camera
- Rubin Observatory will conduct the Legacy Survey of Space and Time (LSST)
  - First Photon: mid 2024; First Light: late 2024
  - Operations begins in 2025
  - 10 year survey - 20 TB of data each night, 10 PB year

Data and Processing Overview

- Data is transferred from Chile to:
  - USDF-SLAC

  and then replicated [Rucio / FTS] to
  - FRDF-CC-IN2P3
  - UKDF-RAL
- Multiple Types / Scenarios of Processing
- Production / Data Release Processing (DRP)
  - Multi-Site (USDF/FRDF/UKDF) -> Panda
- Support for Developers / Scientists at USDF - SLAC
  - Algorithm testing / devel at USDF -> HTCondor (Slurm)

LSST Batch Production Service (BPS)

- BPS = Project software to support general Workflows
  - ctrl_bps https://github.com/lsst/ctrl_bps
  - https://arxiv.org/abs/2206.14941
  - Different workflow systems supported via Plugins
    - ctrl_bps_panda
    - ctrl_bps_parsl
    - ctrl_bps _htcondor
      - Utilizes DAGMan workflows

USDF at SLAC Resources

- SLAC Shared Scientific Data Facility (S3DF)
  - LCLS, Rubin, and others
  - Slurm Cluster, multiple partitions
    - roma 40 nodes, 128 cores, AMD EPYC 7702
    - milano 134 nodes, 128 cores, AMD EPYC 771
- From *ctrl_bps_htcondor* view, no persistent HTCondor pool
- Glide-in solution for users =>
  - Rubin/lsst legacy packages
    - ctrl_execute, ctrl_platform_*
    - https://github.com/lsst/ctrl_execute
    - Supporting CM, scheds on devel login nodes

Glide-in Packages and allocateNodes

- ctrl_execute provides the *allocateNodes.py* utility

```
% allocateNodes.py -n 20 -c 32 -m 4-00:00:00  -q roma,milano -g 900    s3df
% allocateNodes.py --help
     positional arguments:              platform         node allocation platform
 options:
  -n NODECOUNT       number of glideins to submit;
                     these are chunks of a node
  -c CPUS            cores / cpus per glidein
  -m MAXIMUMWALLCLOCK  maximum wall clock time; e.g., 10:00:00
  -q QUEUE           queue / partition name
  -g GLIDEINSHUTDOWN     glide-in inactivity shutdown time in seconds     ….
```

- ctrl_platform_s3df provides templates
  - glide-in condor config, Slurm submit file, bash script to srun

ctrl_execute Usage and Status

- < 2013 TACC Lonestar, SDSC Gordon
- 2013 NCSA Blue Waters (PBS) => 10,000 cores exercise
- 2019 NCSA Slurm Cluster
- 2023 Mitaka NAOJ (PBS Cluster)
- 2023 S3DF - USDF at SLAC Slurm Cluster
  - Current status at USDF at SLAC: users running workflows across 1000's of cores; so far so good
  - Problem/Limitations: ctrl_execute does not maintain glide-ins
    - Running workflows end-to-end problematic
    - users have to replenish if glide-ins expire (by hand, cron)
    - Possible development to have BPS workflow manage

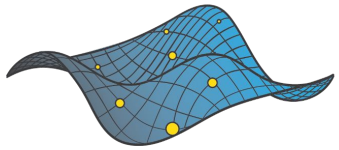
Future Plans at USDF-SLAC

- Decision Point for USDF work:
  - Continue to ‘tune’ ctrl_execute / allocateNodes, or
  - Should we try running ‘GlideinWMS’ ?
    - opensciencegrid/gwms-factory container
    - opensciencegrid/vo-frontend container
    - Test with docker; Running in k8s likely final target
    - Security infrastructure
  - Differences for users
    - Ownership of files (if factory uses service account)
    - One size of glide-ins from factory?
- Use of HTCondor at USDF motivates Pool at Summit (k8s)

NCSA Activities 2023 with HTCondor

- Support Rubin Obs USDF, Summit
- Dark Energy Survey DM late stages
  - Illinois Campus Cluster DES pool, CAPS partition
  - FNAL (jobsub-lite)
- Support South Pole Telescope (SPT) :
  - OSG Frontier Squid setup at Illinois Campus Cluster
- Prototyping in FABRIC testbed for CMB-S4

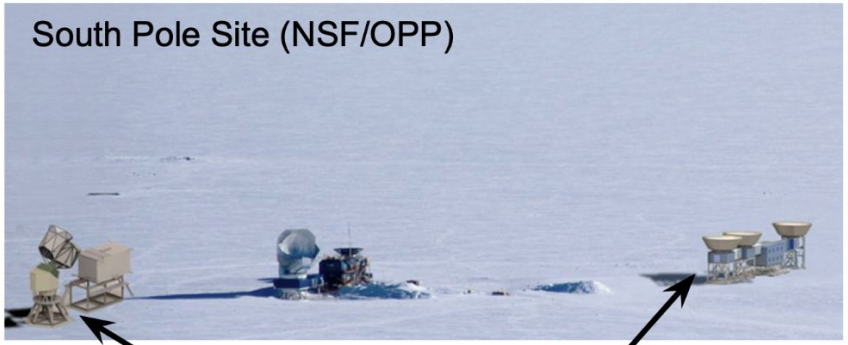
DARK ENERGY SURVEY



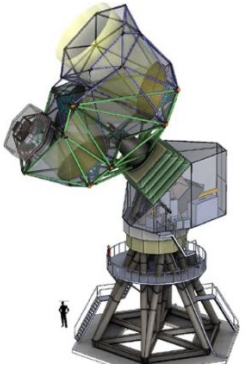
FABRIC

CMB-S

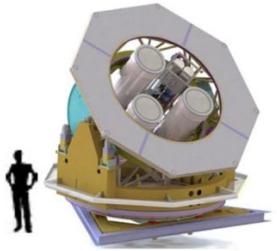
CMB-S4 Sites & Instrumentation

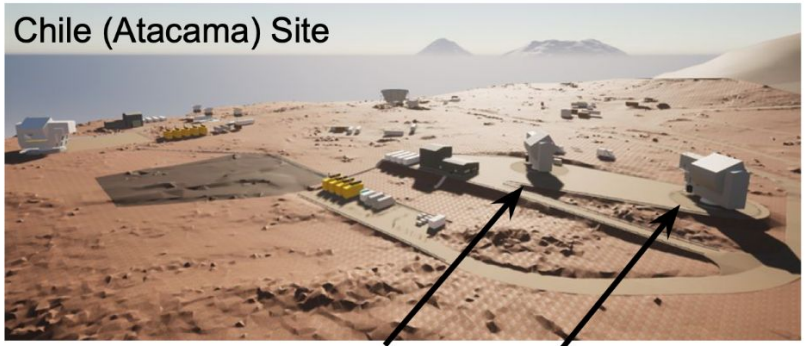


1 Large Aperture
(5 m) Telescope

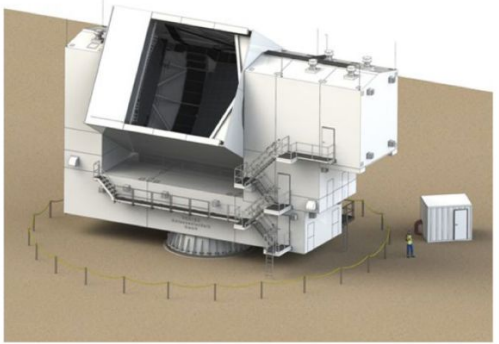


3 Small Aperture Telescopes
(9 0.5-m aperture optics tubes)





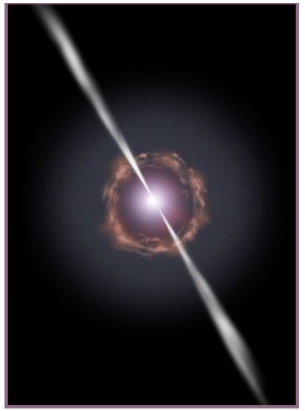
2 Large Aperture (6 m)
Telescopes



CMB-S4: Transient Phenomena

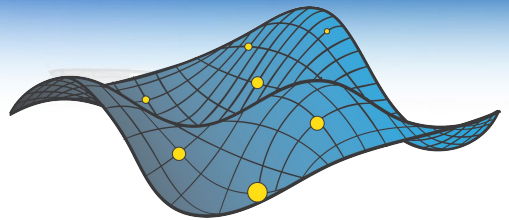
Astro2021: CMB_S4 will produce data sets of unprecedented sensitivity, cadence and spectral coverage…. **Seek broad engagement with astronomers** … **maximize opportunities for transient science**…..

- Stellar Flares
- Tidal Disruption Events
- Active Galactic Nuclei
- X ray Binaries
- Classical Novae
- Planetary Nebulae
- SN Type 1 progenitors
- Magnetars
- Solar system Objects

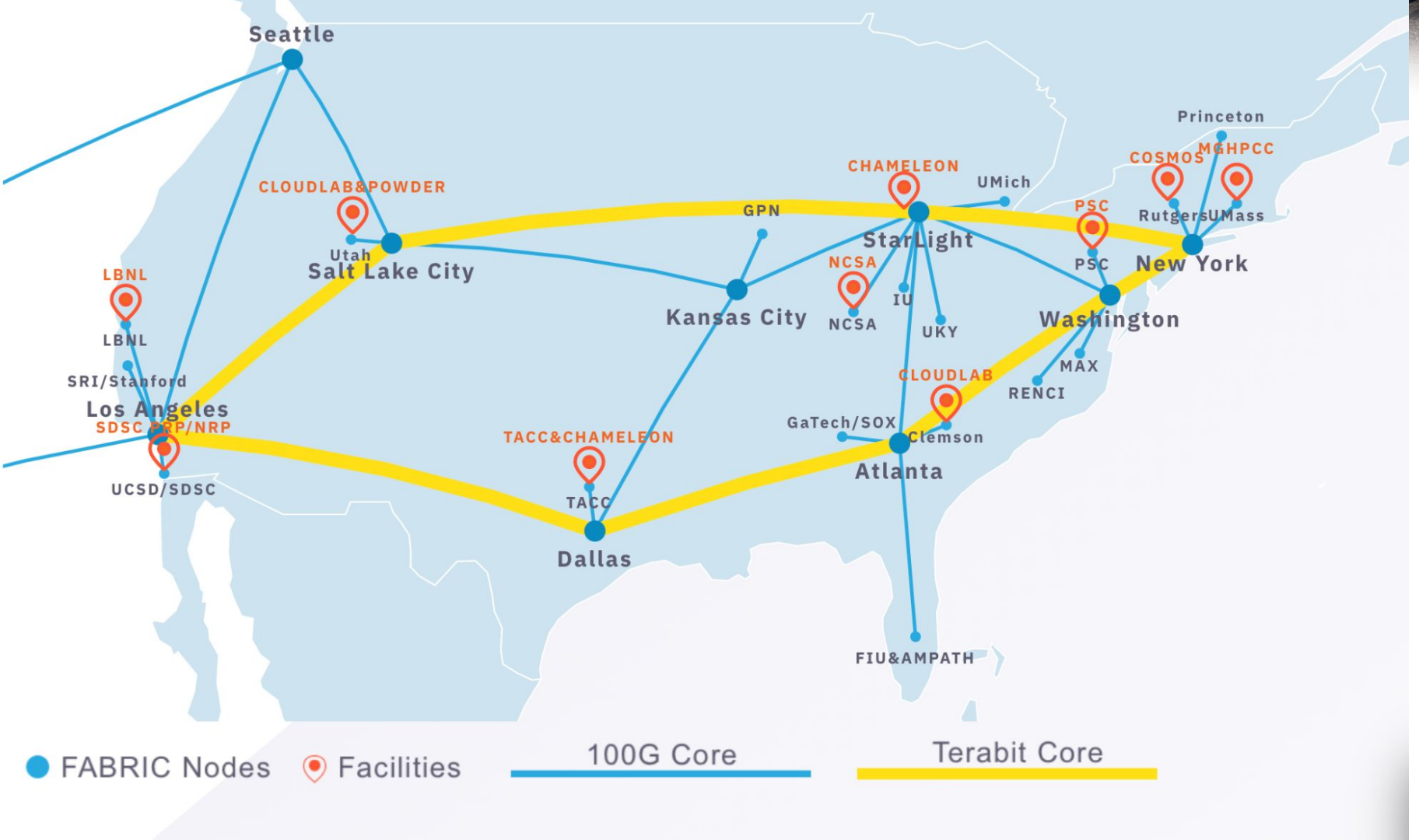


The infrastructure we are prototyping in FABRIC its to detect and announce announce initial detections to the community. In era of CMB-S4 we are in “discovery space”

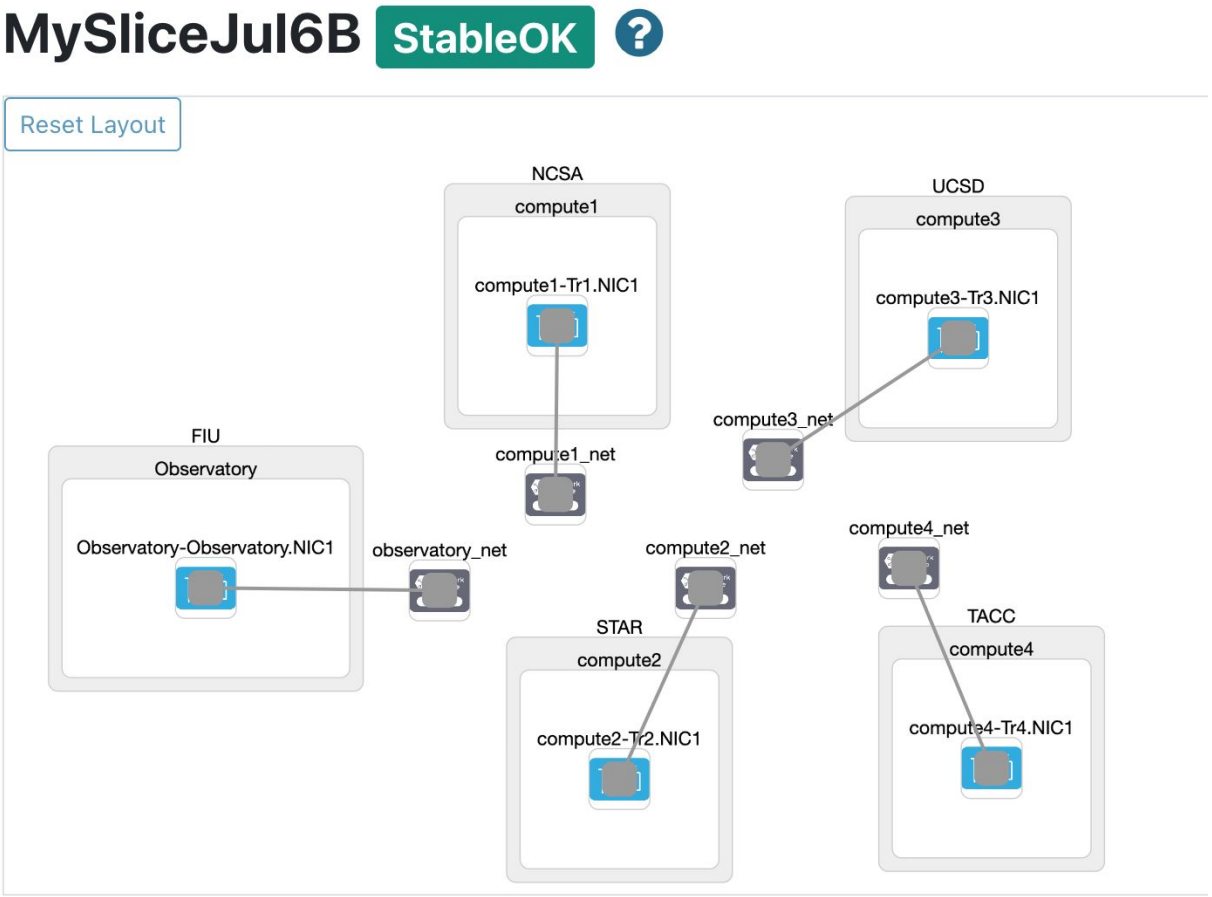
FABRIC Testbed



- Adaptive Programmable Research Infrastructure for CS and Science Applications
  - https://fabric-testbed.net
- 29 FABRIC sites with compute & storage interconnected by high speed, dedicated optical links
- Users create Experiments with VMs, NICs, networks, routes, etc. via Python API: https://pypi.org/project/fabrictestbed-extensions/


Seattle
CLOUDLAB&POWDER
Utah
Salt Lake City
LBNL
LBNL
SRI/Stanford
Los Angeles
SDSC PRP/NRP
UCSD/SDSC
GPN
Kansas City
TACC&CHAMELEON
TACC
Dallas
CHAMELEON
UMich
StarLight
NCSA
NCSA
IU
UKY
GaTech/SOX
Clemson
Atlanta
CLOUDLAB
FIU&AMPATH
PSC
PSC
New York
Princeton
COSMOS
MGHPCC
Rutgers
UMass
Washington
MAX
RENCI
FABRIC Nodes
Facilities
100G Core
Terabit Core

Multisite Slice on FABRIC



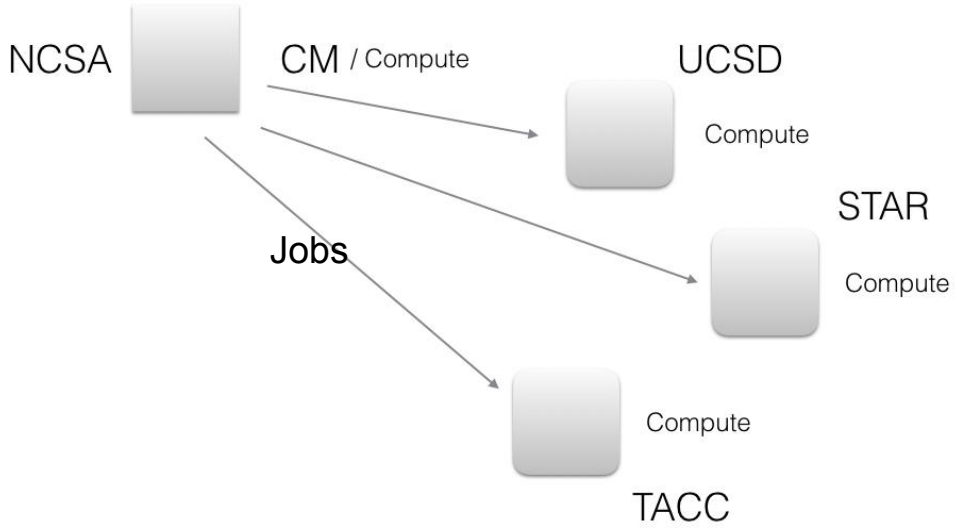
HTCondor Computing Pool on FABRIC

- Create an L3 IPv6 network at each site, add site VM, add routes between all

- All nodes configured for an HTCondor role (Manager, Execute)

COLLECTOR_HOST=<NCSA IPv6 IP#>
NETWORK_INTERFACE=<Site IPv6 IP#>

- Compute Nodes communicate with a CM using L3 networks/routes
- Jobs submitted across pool



Data Flow

- A main workflow on CM discovers available data at FIU / Observatory
- All jobs on compute nodes copy data from persistent storage at FIU / Observatory using L3 networks/routes (data plane)