DIRAC @ PNNL

MATHEW THOMAS
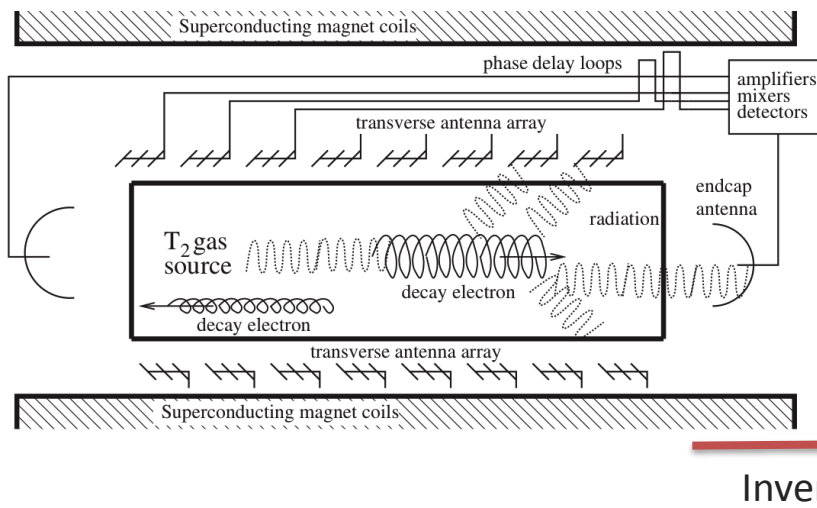ON BEHALF OF PNNL PROJECT8, AND CCSDI

Pacific Northwest National Laboratory

Cyclotron Radiation Emission Spectroscopy
(measure frequency of radiation from magnetically trapped electrons)

Invert $f(E)$

Tritium β-decay electron endpoint

Bodine, Parno and Robertson, *Phys. Rev. C* **91** (2015).

$Q \approx 18.6$ keV
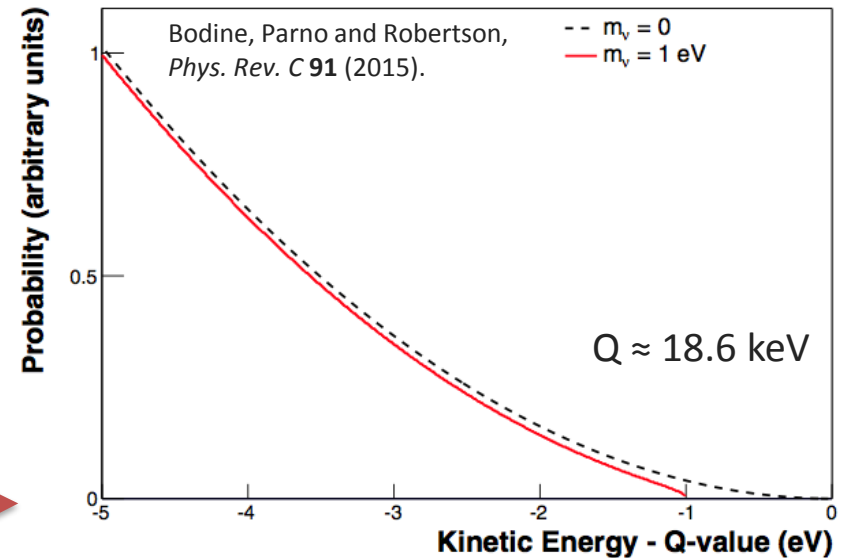
▶ Tritium Beta Decay: $^3\text{H} \rightarrow {}^3\text{He}^+ + e^- + \nu_e + Q$.

▶ High-precision spectroscopy on the e^-.

▶ Neutrino mass manifests as a deviation at the energy endpoint.

▶ Fit the spectral shape with $m_{\nu e}^2$ as a free parameter:

Project8 Data Taking and Computing

▶ Expected data rates to increase at each experimental phase:

- Phase 1 & 2 (now): ~0.5PB
- Phase 3 (FY2021): 10-20PB
- Phase 4 (FY2025-2030): ExaBytes

▶ Processed data samples will be distributed within the Continental U.S.

▶ Project8 leverages metadata information in conjunction with the transformation system to automate the production efforts.

▶ PNNL leads the design, deployment, and operations of the computing effort

▶ Leveraging virtualization expertize to deploy a robust and scalable solution using containers, Kubernetes, and HELM. Used for core Project8 computing system at PNNL and Yale.

Project 8 and DIRAC

▶ Project 8 currently uses DIRAC for all aspects of data and work management:

- RAW data transfers from experiment DAQ at University of Washington (UW) to storage element at PNNL and Yale
- Automated analysis of calibration and slow control data, with results returned to operations web server at UW
- Analysis of fast DAQ time series data
- User access to data and CPUs at PNNL
- Transformations to merge and plot data files based on metadata

# Current Project8 DIRAC Computing Setup
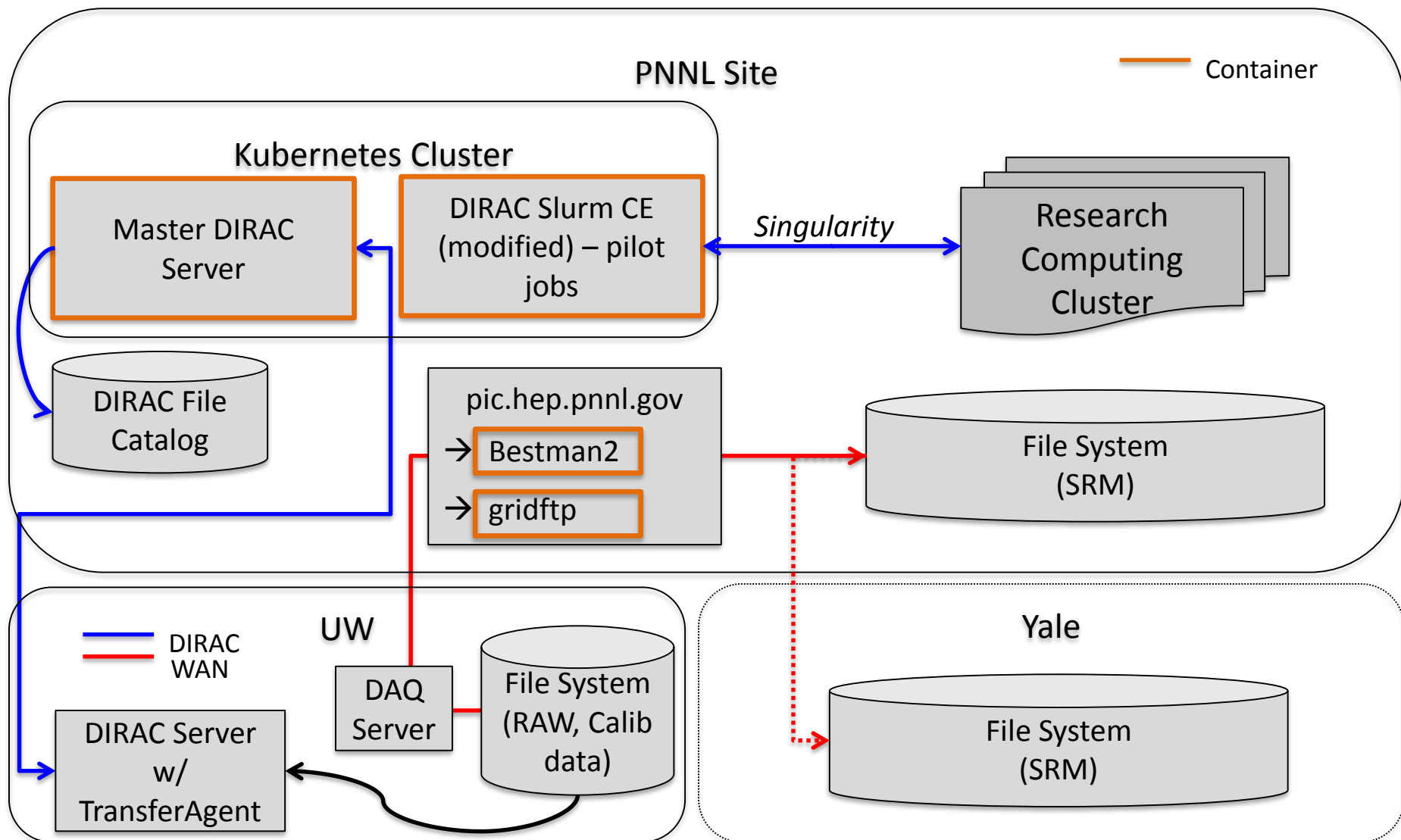
Automating Project8 data processing using DIRAC TS

- Raw data is produced at the University of Washington and registered to the DIRAC File Catalog with well defined metadata
- A compute transformation is defined:
  - Retrieve files with metadata with DataType='Data' and DataLevel='RAW'
  - Run reconstruction jobs
  - Upload output processed files to storage and assign ancestry and metadata, such as DataLevel='Processed' and SoftwareTag='XYZ'
- A merging transformation is defined:
  - Retrieve files with metadata DataType='Data' , DataLevel='Processed', and SoftwareTag='XYZ'
  - Merge files and upload output with ancestry and metadata, such as DataLevel='Merged'
- An analysis transformation is defined:
  - Retrieve files with metadata with DataType='Data' , DataLevel='Merged', and SoftwareTag='XYZ'
  - Upload outputs with ancestry and metadata, such as DataLevel='Analysis' or DataLevel='Monitoring'
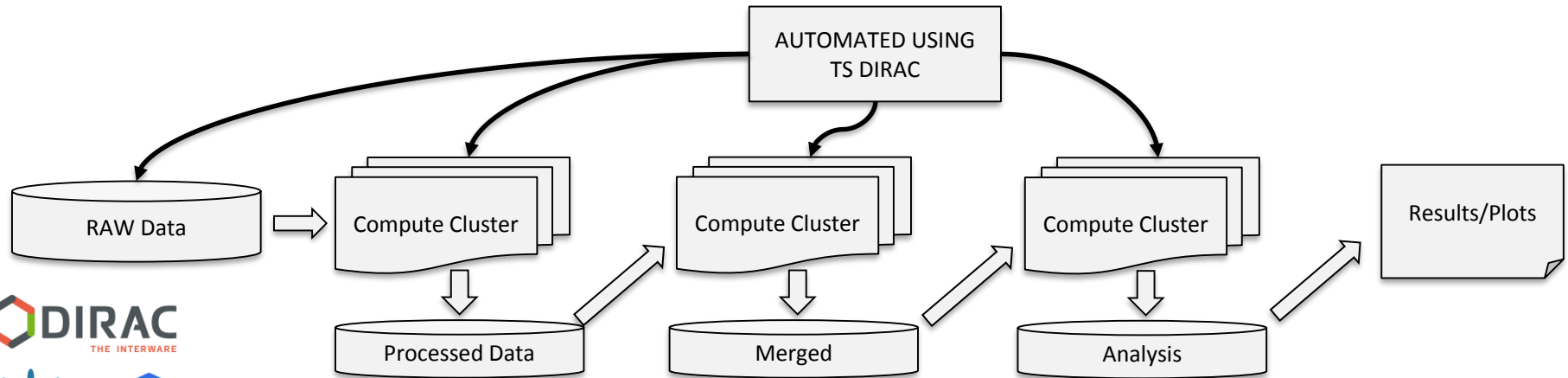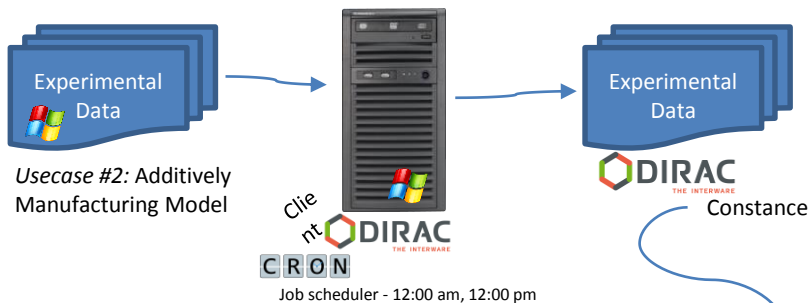
Complex Chemical System Dynamics Initiative

- Goal: Develop the scientific foundation necessary to understand and predict the temporal evolution of complex chemical systems in real-world environments.
- DIRAC provides a common framework across projects to study chemical system dynamics.
- Extensions are being developed to automate workflows and provide reproducibility.
- DIRAC Infrastructure provides the ability to process and access data:
  - A file catalog for experimental, model, and analytic information
  - Design and setup automated workflow for measurements, models, and comparisons.
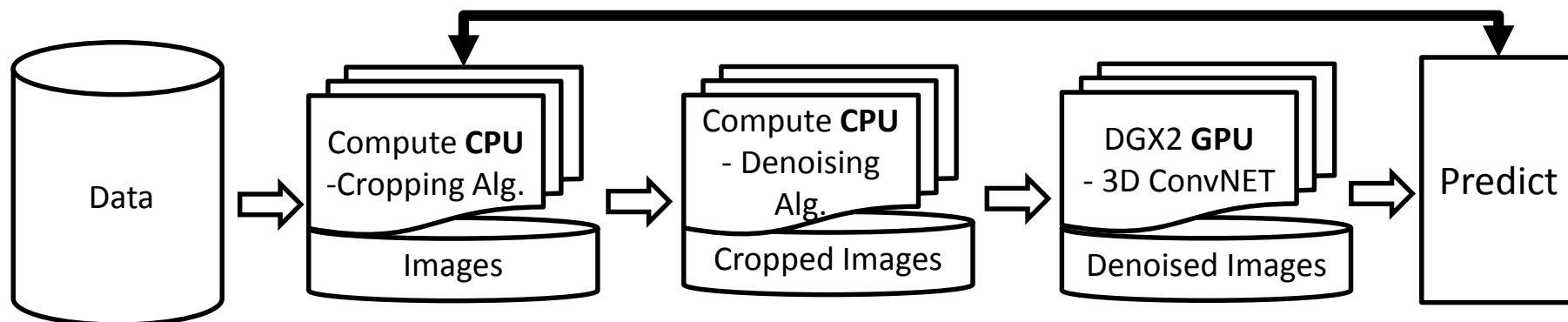
Workflows, Data Management, and Machine Learning (ML) provenance

Experimental Data

Usecase #2: Additively Manufacturing Model

Clie nt **DIRAC** THE INTERWARE

CRON

Job scheduler - 12:00 am, 12:00 pm

Experimental Data

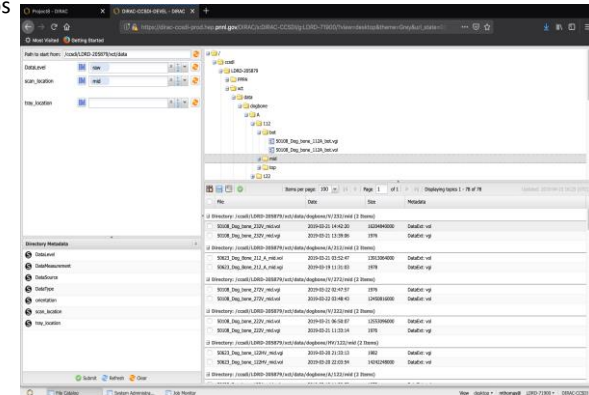**DIRAC** THE INTERWARE

Constance

DIRAC Webapp:
➢ Browse data
➢ Filter to display files with specific scan location
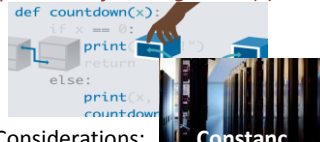➢ Monitor, submit jobs

DIRAC Client:
➢ Use DIRAC to get the paths to files/directories that need to be processed based on metadata
➢ Launch, monitor jobs

DIRAC Client on Windows has a cron job that does the following:
➢ Looks for new 3d .vol files
➢ Cleans up the folder and file names
➢ Uploads data to DIRAC storage on Constance
➢ Adds necessary metadata

Data Transfer

DIRAC

Steps to transform Algorithm(s) to workflows...

```
def countdown(x):

    print
    return
    else:
    print(x
    countdow
```

Constanc

Design Considerations:
➢ Inputs: Accept paths to input data.
➢ Outputs:
  ▪ Temporary files: write any temp files to the current working directory.
 ▪ Results: output results to usecase based folder architecture.

Job Metrics

DIRAC

DIRAC

3d vol → 2d slices

2d slices → cropped 2d slices

Register output files to **DIRAC** THE INTERWARE

Define the metadata fields and ancestry based on usecase architecture.

Write the script to launch jobs using DIRAC

Usecase #2 Automated Workflows

3d vol → 2d slices → cropped 2d slices (method 1) → denoising → porosity calculation → cropped 2d slices (method 2) → denoising → porosity calculation

ML Study   ML Study   ML Study

mlflow http://mlflow-ccsdi.pnl.gov/#/
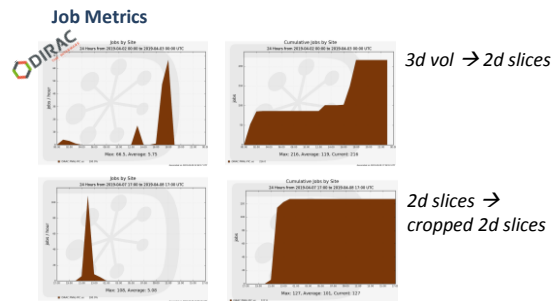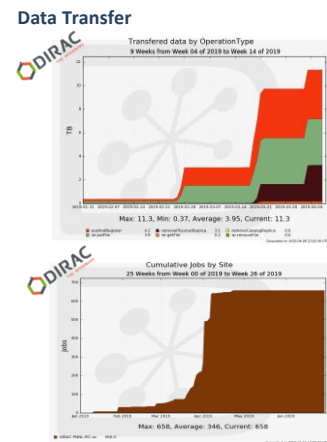
ml*flow* Tracking	ml*flow* Projects	ml*flow* Models
Record and query experiments: code, configs, results, ...etc	Packaging format for reproducible runs on any platform	General model format that supports diverse deployment tools

Mlflow is an open source platform to manage the ML lifecycle, including experimentation, reproducibility and deployment

Summary of DIRAC related work at PNNL

- We are using and developing DIRAC components for 2 active projects
- We are investigating ML techniques to improve overall efficiency
- We heavily use the Transformation System to automate workflows
- General interest in using DIRAC for other projects
  - Evaluations are ongoing or planned
- Trying to upgrade the version of DIRAC used