Pacific Northwest
NATIONAL LABORATORY

Automated distributed computing for Project 8

October 28, 2019

Malachi Schram

on the behalf of the Project 8 collaboration
PNNL Data Science Architectures and Ai Team Lead

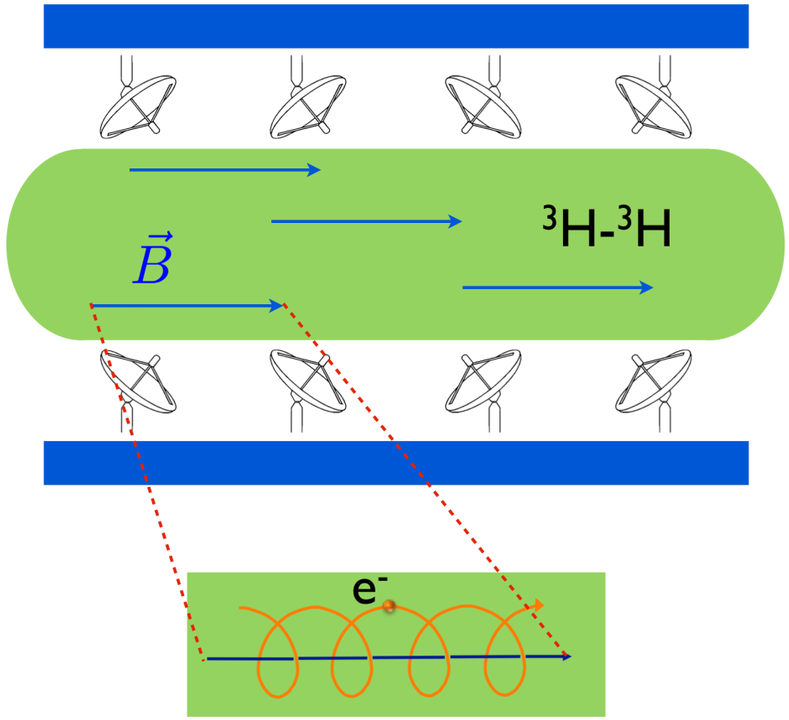
U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Project 8 high-level overview

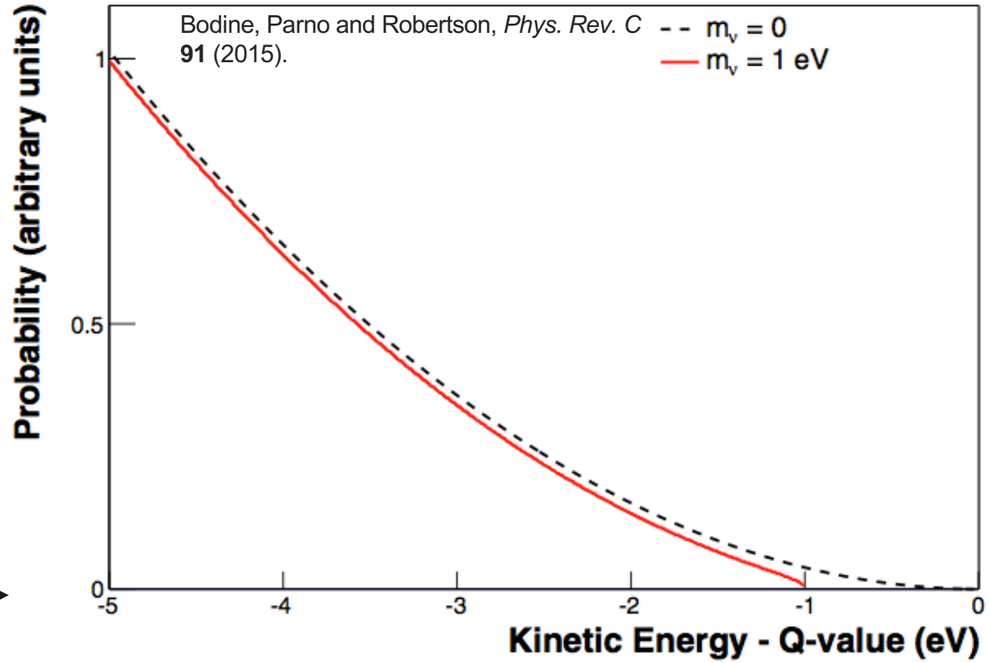
- The goal of the Project 8 experiment is to measure the absolute neutrino mass using tritium beta decays. The approach taken by the Project 8 collaboration is to make this measurement using a new method of electron spectroscopy, Cyclotron Radiation Emission Spectroscopy (CRES)

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



Invert *f(E)*

Tritium β-decay electron endpoint



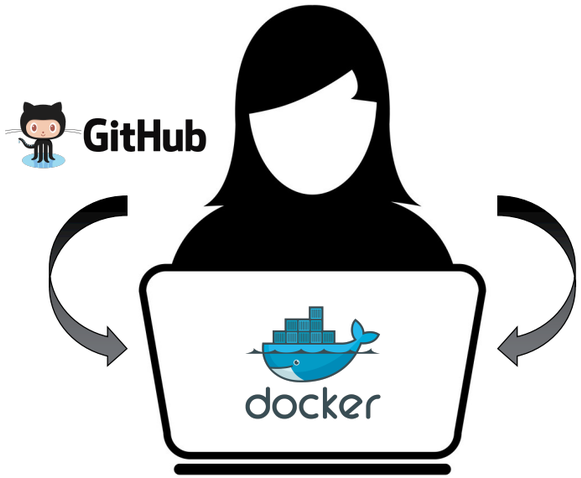
Pacific Northwest
NATIONAL LABORATORY

Computing Requirements

- The current data rates are modest, however, they are expected to increase significantly in the later experimental phase:
  - Phase I & II (now): ~0.5PB
  - Phase III (FY2021): 10-20PB
  - Phase IV (FY2025-2030): Potentially ExaBytes without triggering and data reduction
- Processed data is currently distributed to a select number of sites.
- Research for reducing the anticipated data volume is ongoing
- Close coordination between online and offline computing will be critical in the later phases of the experiment.

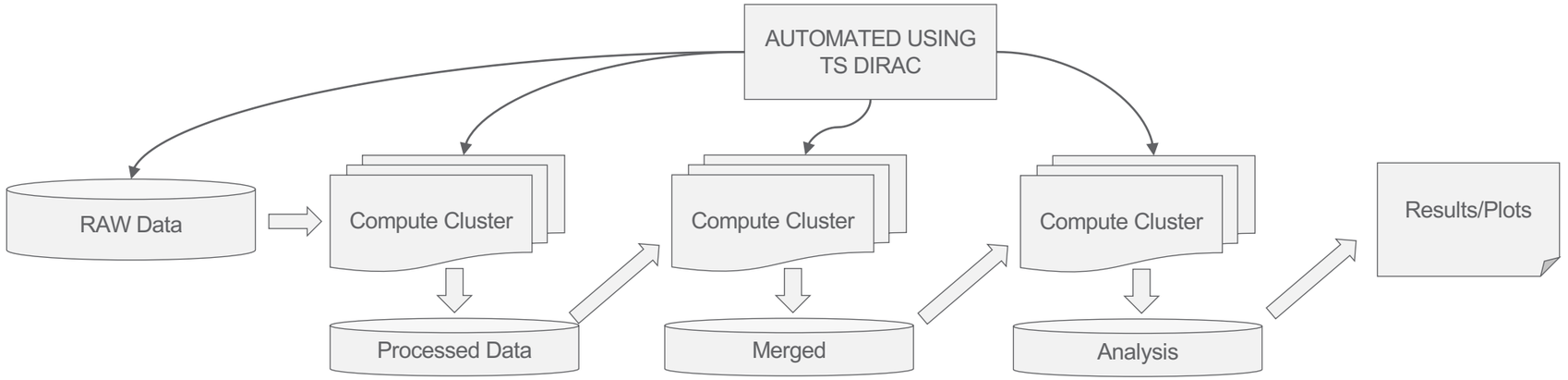
Designing a flexible Infrastructure

- The core computing services are hosted and managed at the Pacific Northwest National Laboratory (PNNL) using ***Kubernetes***.
- Individual containers are used to instantiate ***DIRAC*** services, agent, and databases and other core grid services.
- ***Docker*** containers are also used to build new instances of the development and production environment.
  - This provides maximal flexibility and satisfies the collaborations specific OS and libraries requirements.
  - Production containers are then converted to a ***Singularity*** image. Computation jobs are performed on the PNNL HPC cluster using a dedicated DIRAC agent mapped to the desired singularity image.



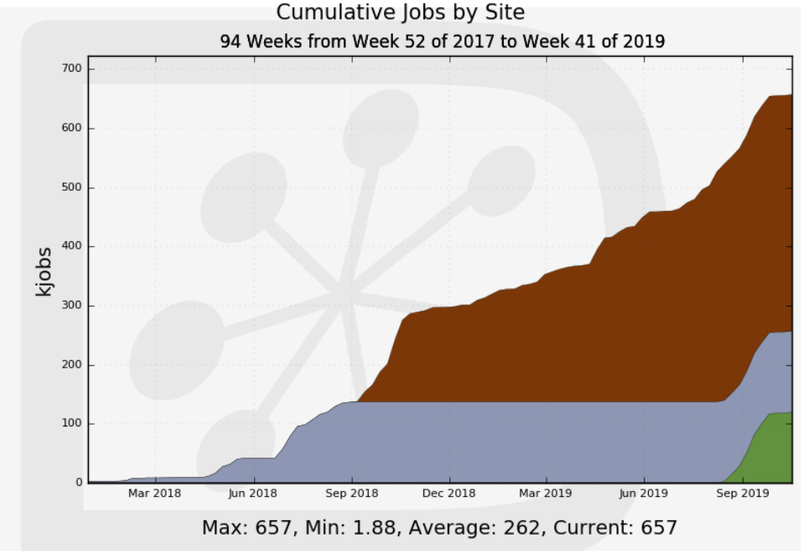
Distributed computing and automating workflows

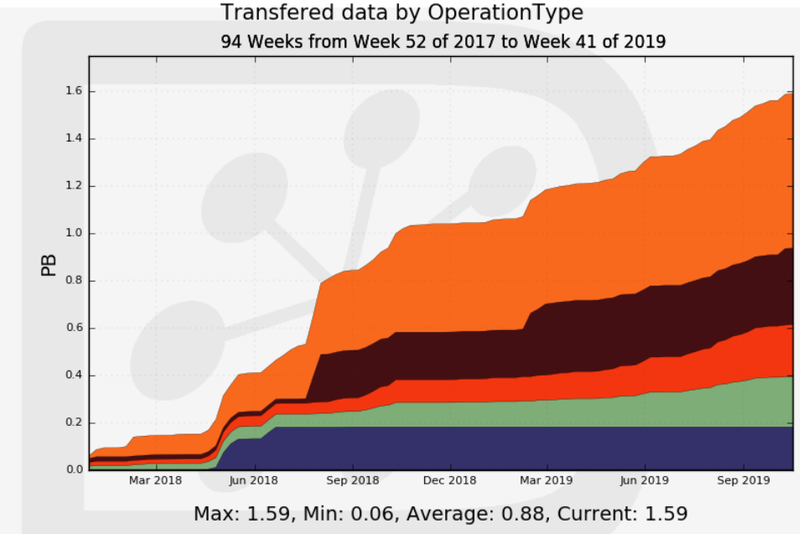
- Project 8 has adopted DIRAC (Distributed Infrastructure with Remote Agent Control) INTERWARE as it's distributed computing workflow because it provides a complete solution and includes automation features.
- These automated features are critical for smaller scale experiments such as Project 8.
- The raw data is produced at the University of Washington and transferred to PNNL using a dedicated DIRAC agent.
- The files are registered to the DIRAC File Catalog with well defined metadata in order to trigger the raw data processing workflow. Once a data production run is finished, all raw processed files are automatically merged and analyzed.



Computing usage to date

- The majority of the computing used by the production system has been to process and analyze the data.
  - Over 650,000 jobs have already been executed
  - Nearly 2PB of data have been moved.
- We expect an increase in computational usage in order to preform simulation studies for phase III and beyond.
- We are currently re-evaluating the offline computing in preparation for phase III





Pacific Northwest
NATIONAL LABORATORY

Thank you

This work was supported by the US DOE Office of Nuclear Physics, the US NSF, the PRISMA+ Cluster of Excellence at the University of Mainz, and internal investments at all collaborating institutions. The author recognizes support from the Laboratory Directed Research and Development program at PNNL. A portion of the research was performed using Research Computing at PNNL.