Distributed Computing for the Project8 experiment

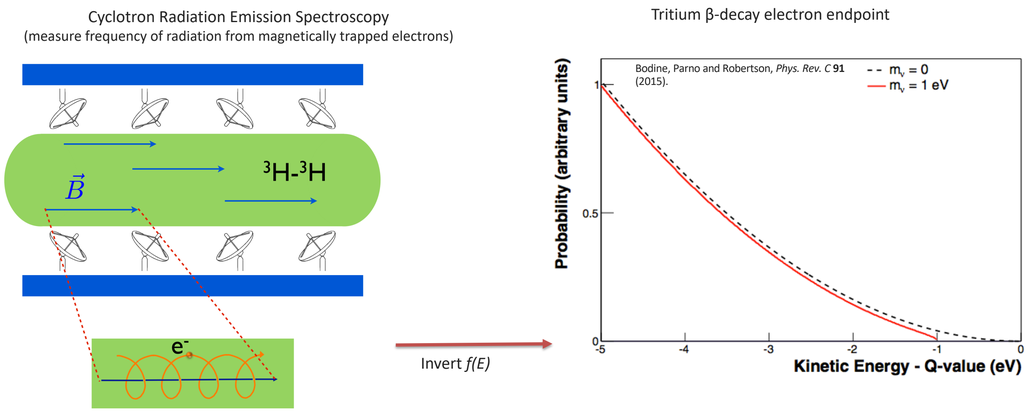
Malachi Schram, Mathew Thomas, Kevin Fox, David Cowley, and Benjamin LaRoque for the Project 8 Collaboration

PROJECT 8

Pacific Northwest NATIONAL LABORATORY

Introduction

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)



Computing requirements

The current data rates are modest, however, they are expected to increase significantly in the later experimental phase:

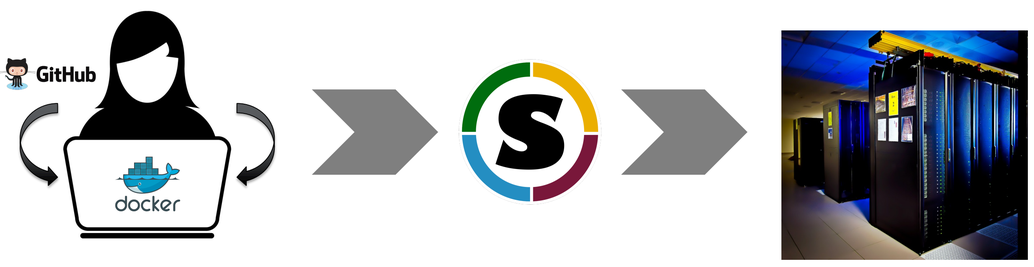
- Phase 1 & 2 (now): ~0.5PB
- Phase 3 (FY2021): 10-20PB
- Phase 4 (FY2025-2030): ExaBytes

Processed data is currently distributed to a select number of sites.

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.



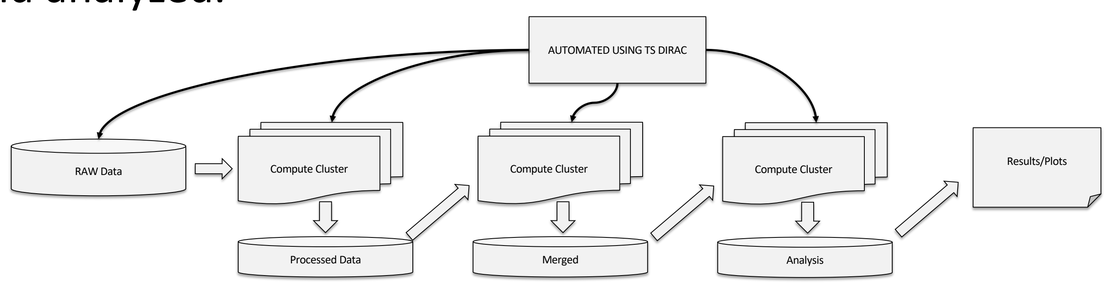
Distributed computing and Automating Workflows

Project8 has adopted DIRAC (Distributed Infrastructure with Remote Agent Control)

DIRAC THE INTERWARE

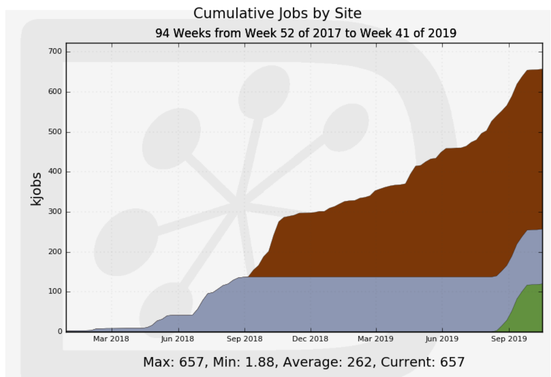
INTERWARE as it's distributed computing workflow because it provides and complete solution which include automation features. These automated features are critical for smaller scale experiments such as Project8.

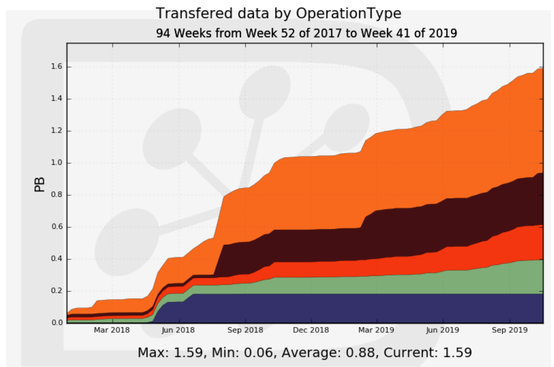
The raw data is produced at the University of Washington and transferred to PNNL using a dedicated DIRAC agent. The files are registered to the DIRAC File Catalog with well defined metadata in order to trigger the raw data processing workflow. Once a data production run is finished all raw processed 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 be executed and nearly 2PB of data movement. We expect an increase in computational usage in order to preform simulation studies for Phase 3 and beyond.





Acknowledgment

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