Data Management and Workflow Challenges in LArSoft

Michael Kirby, Fermilab/Scientific Computing Division

June 25, 2019

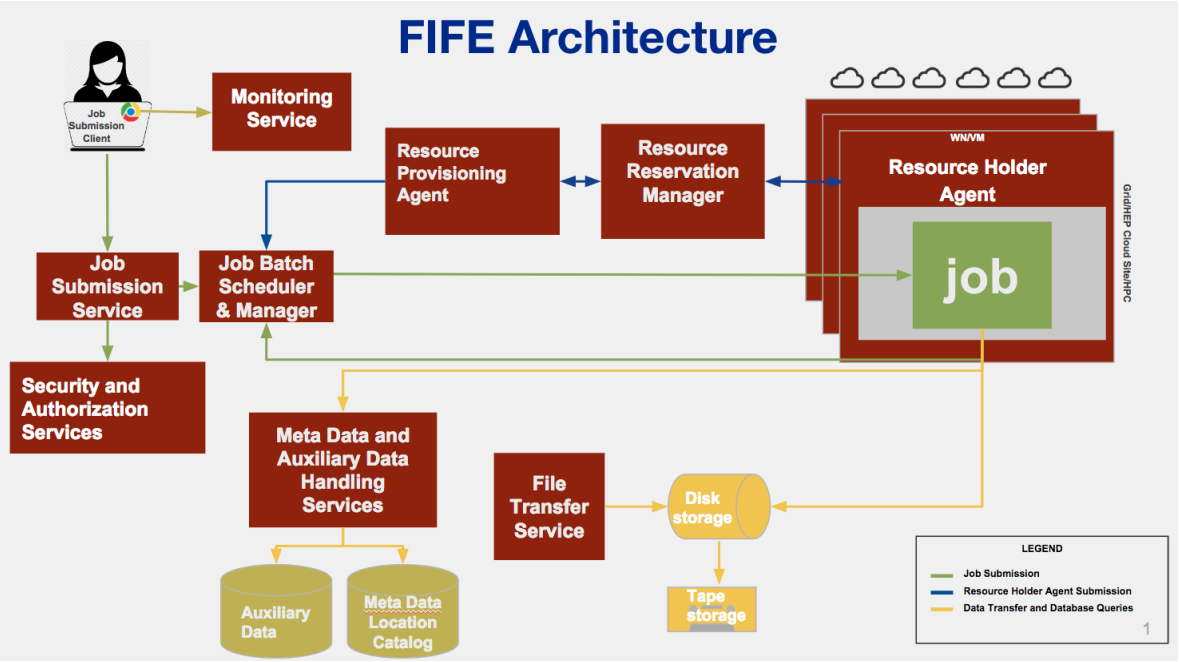
Data Volumes for LArTPCs of the future…

- "event" volumes for DUNE are an order of magnitude beyond collider events
  - already quickly reducing the data volume from raw to just hits 8 GB/trigger -> 100 MB/trigger
  - workflow question of persistent metadata of transient data structures
  - do we know if the current LArSoft framework is sufficient for analysis needs
- what is an event?
- handling of sub-events
- supernova readout
- proton decay event processing

Source	Annual Data Volume
Beam interactions	27 TB
Cosmics and atmospheric neutrinos	10 PB
Radiological backgrounds	< 1 PB
Cold Electronics calibration	200 TB
Radioactive source calibration	100 TB
Laser calibration	200 TB
Random triggers	60 TB
Trigger primitives	13 PB

DUNE TDR (June 2019 draft)

Data Management and Workflow Solutions needed in LArSoft



- Thinking about data management within the context of distributed computing and HPC
- Each provides separate challenges
- HPC may require either edge services or delivering the data to local SE
- HTC computing may require delivering jobs to the distributed dataset
- within the context of LArSoft though, these problems are independent of that

Fermilab

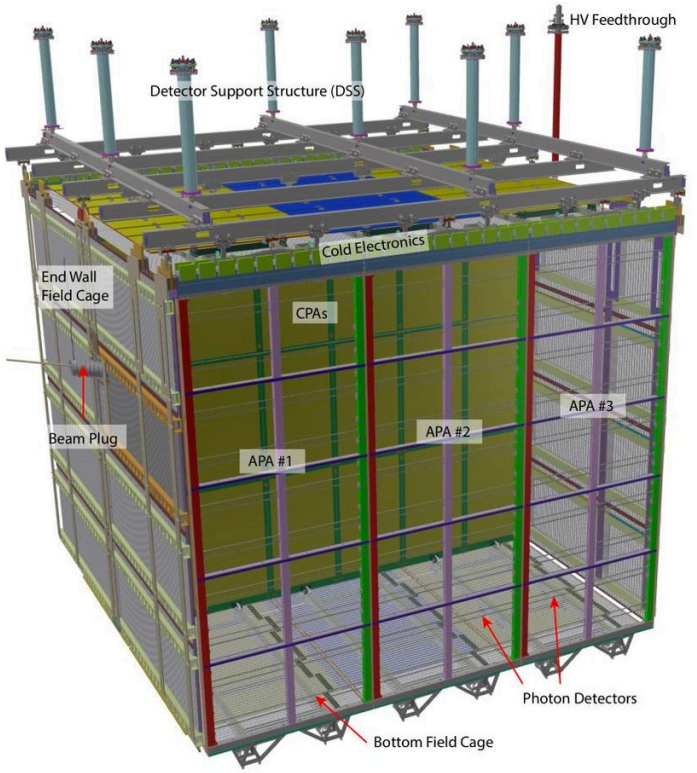
Data Management and Workflows in the era of HPCs

- HPC centers have incredible capabilities
- OSG accessible Storage Elements have not traditionally been one of those capabilities
- “edge” services get access to local storage (luster, etc)
  - stage large datasets into the HPC facility
  - request a reservation and process everything
  - stage data back to local SE
- incorporating event delivery services into the framework
  - ATLAS currently uses an event service to backfill idle cores on Texas Advanced Computing Center
  - how do we bookkeep those events and interface them with LArSoft services

LArTPC specific data management issues

- what is an “event”?
- DAQ software commonly uses the idea of a trigger record
- for a detector with 150 APAs, that may change
- active development on-going with processing each APA in ProtoDUNE separately (6 APAs)
  - do you copy 1 file to six locations?
  - share 1 file to 6 cores on the same node? (benefit from shared memory?)
  - can we distribute APAs across nodes using an “event service”?

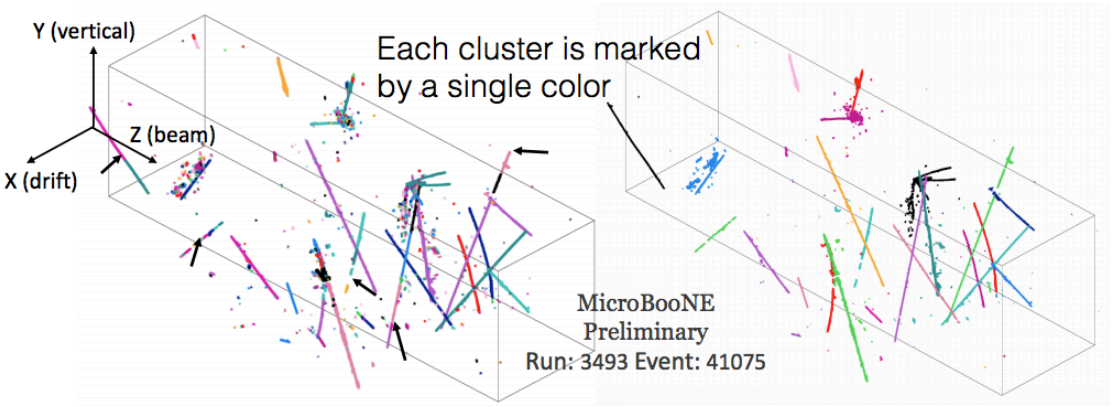
ProtoDUNE Single Phase



Fermilab

Regions of interest and path-level-parallelism

- single trigger record may contain many regions of interest (ROIs)
- how to map/bookkeep multiple ROIs from a single trigger record
- different ROIs processed through multiple paths in a single process
- DUNE not currently taking advantage of multiple paths within the LArSoft framework
  - what are the advantages to doing this?
  - are there features that are needed to make this more useful

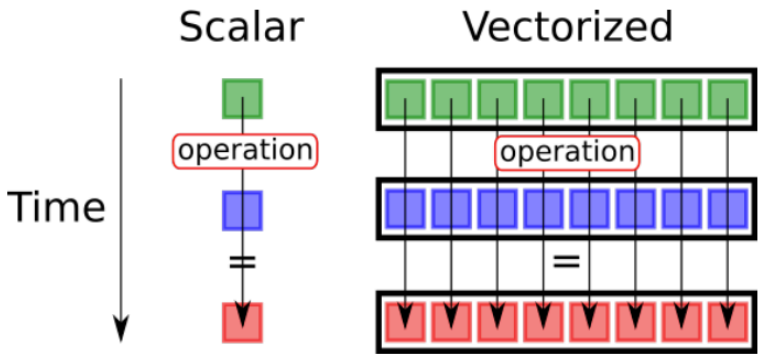


Courtesy Hanyu Wei, *Neutrino 2018*

Fermilab

Pipelined module paths

- parallelism will help address the problem of large memory requirements for LArSoft jobs
- almost immediately have to address the problem of CPU efficiency
  - an HPC cluster will not be overjoyed with users who occupy cluster and leave cores idle
- reading the full event into memory while backfilling idle cores just recreates the original problem of memory usage
- ability to stream subevents, data structures becomes an important part of the workflow

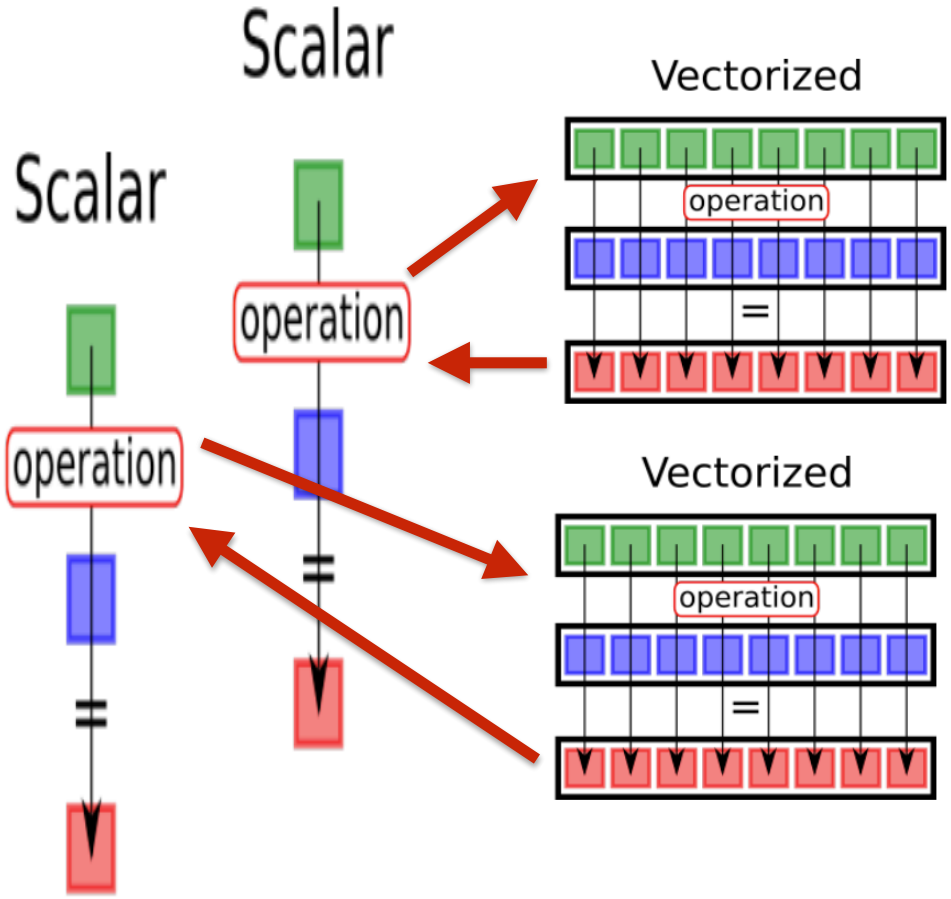


Courtesy of G. Cerati

Fermilab

Pipelined module paths

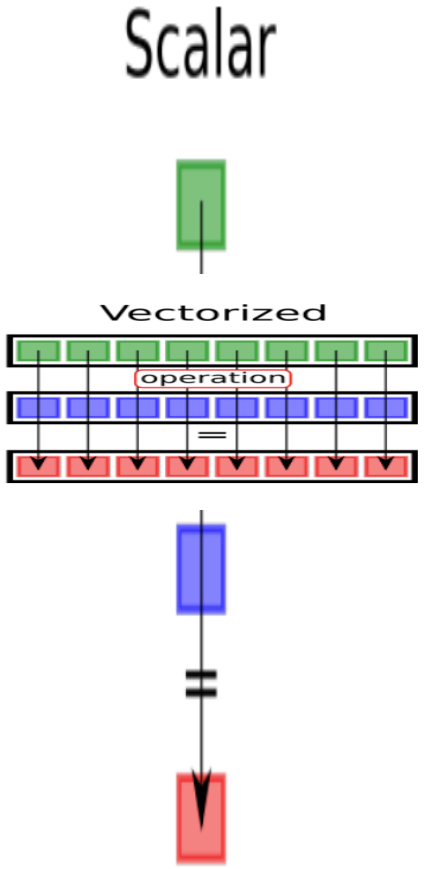
- parallelism will help address the problem of large memory requirements for LArSoft jobs
- almost immediately have to address the problem of CPU efficiency
  - an HPC cluster will not be overjoyed with users who occupy cluster and leave cores idle
- reading the full event into memory while backfilling idle cores just recreates the original problem of memory usage
- ability to stream subevents, data structures becomes important part of the workflow



Fermilab

Pipelined module paths

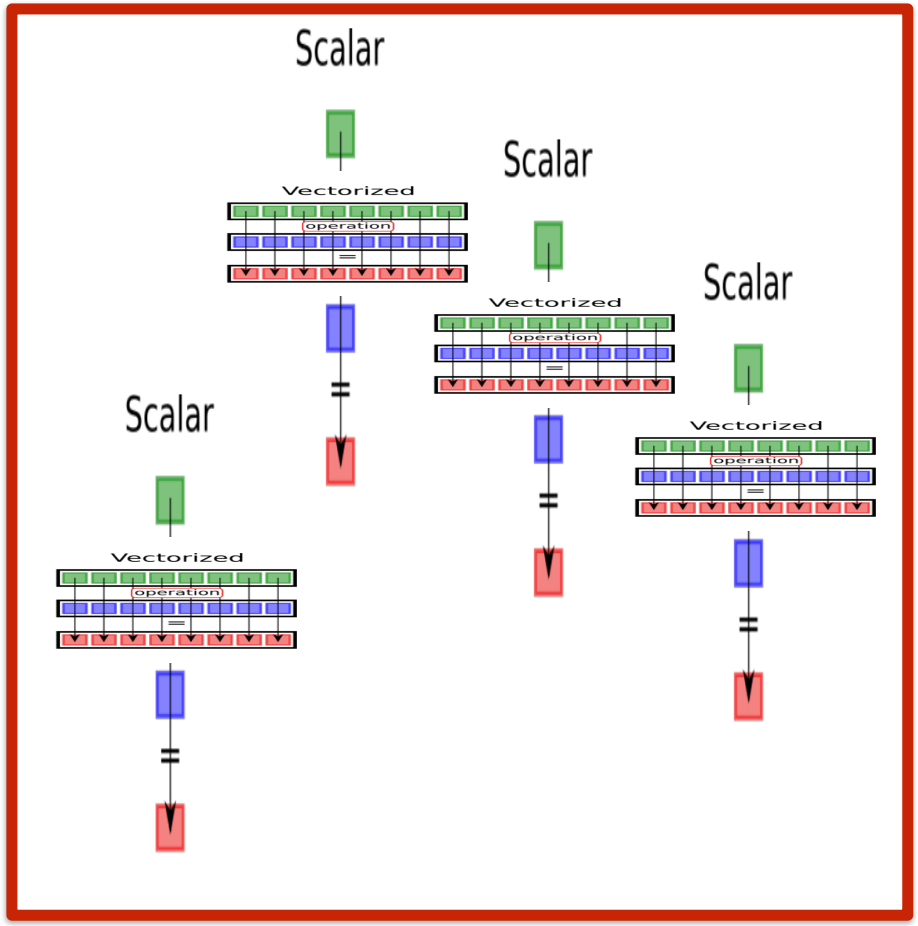
- parallelism will help address the problem of large memory requirements for LArSoft jobs
- almost immediately have to address the problem of CPU efficiency
  - an HPC cluster will not be overjoyed with users who occupy cluster and leave cores idle
- reading the full event into memory while backfilling idle cores just recreates the original problem of memory usage
- ability to stream subevents, data structures becomes important part of the workflow



Fermilab

Pipelined module paths

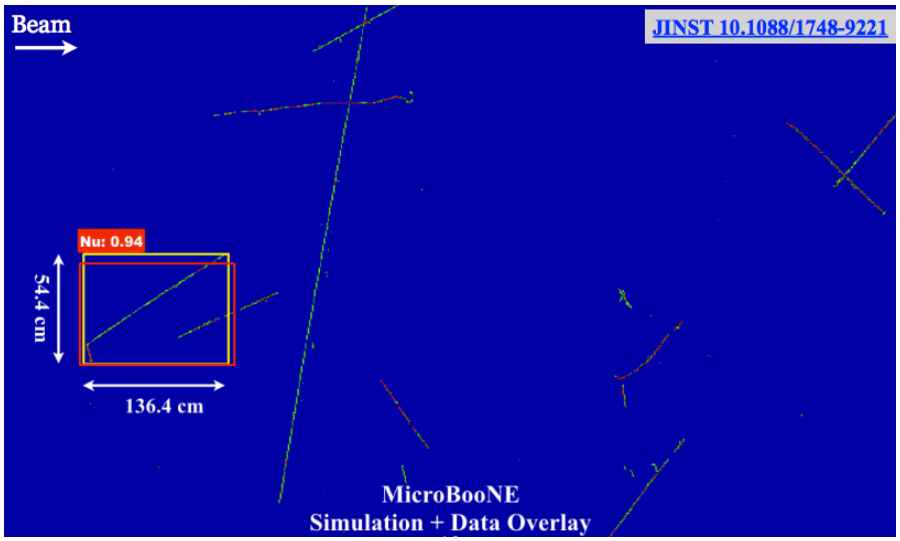
- parallelism will help address the problem of large memory requirements for LArSoft jobs
- almost immediately have to address the problem of CPU efficiency
  - an HPC cluster will not be overjoyed with users who occupy cluster and leave cores idle
- reading the full event into memory while backfilling idle cores just recreates the original problem of memory usage
- ability to stream subevents, data structures becomes important part of the workflow



Fermilab

Framework integration of overlay samples

- cosmic data overlay for DUNE is not as critical of a problem for the LBL oscillation analysis
- significant workflow for all SBN detectors and may be a concern for non-oscillation analyses (proton decay, supernova, solar neutrinos, etc)
- having the framework appropriately handle the sampling of a secondary dataset for merging/ overlay along with tracking metadata for the dataset is important
- as data taking periods become longer, bookkeeping of this information becomes important
- additionally, the framework and SAM don't currently work well together to ensure overlay data requests happen with priority based upon staging status



Current questions of workflow and data processing

- Please note: these are not necessarily issues of the framework!
- Efficient handling of sparse dataset and intermediate files that this processing can necessitate within POMS
- if a job is configured for multiple streams and there is a failure in one stream, how to recover without recreating duplicate files
- quarantining failed files is not currently possible within POMS - it would be extremely helpful to be able to remove a file from processing after N retries
- LArSoft framework works wonderfully for processing artroot files - there is a lack of a “framework” for processing non-artroot files (plain ntuples, etc) and this gap could be a problem
  - CAFAna is actively in use for DUNE and NOvA, but not a fully supported analysis framework

Summary

- LArTPC data volumes are not going to be the driver for data lakes, object stores, etc, but event volumes will be a driver for framework and data handling features
- LArSoft's ability to handle "large" events and transient data products will play a significant role in addressing this challenge
- DOMA middleware needs to prepare for handling these datasets on HPC through edge services and ensure that event size is not an issue
- processing trigger record across different architectures (i.e. numerous cores for same record) will require data delivery of sub-events
- framework will need to handle the transition from trigger record into ROIs, subevents, etc
- configuration of event staging to memory should make sure that path-level parallelism and pipelining of tasks doesn't contravene the memory benefits of threading
- current workflow tools have some limitations that would make a significant improvement to production efficiency

Big Thanks: Brett Viren, Tom Junk, Herb Greenlee, Ken Herner, Erica Snider, Giuseppe Cerati