EIC Software Infrastructure Review

Production Strategies

Joe Osborn

On behalf of the EPIC Collaboration

EPIC

		Discussion topic(s)	Decision topic(s)	comments	Point of Contact
May	4	*AIWG*			
	11	Transition Period	Present procedure. Decide on list and order of decision topics		
	18	*No meeting (Streaming Readout X Workshop)*			
	25	Code Repository 	Repository: - Location (GitHub, GitLab+Host) - Admins - Access		David Lawrence
Jun	1	Discussion Schedule	Schedule: - Decide most critical decisions to make before July 27th EICUG meeting - Schedule of topic discussions		
	8	Geometry 	Geometry: - Package (e.g. DD4HEP)		Markus Diefenthaler
	15	Data Model 	Data format - Generated events - Simulated data - Processed data (e.g. ROOT w/ specific tree format)		Whitney Armstrong
	22	Data Model			
	29	Reconstruction Framework 	Reconstruction Framework - Package		Wouter Deconinck
Jul	6	Reconstruction Framework			
	13	Data and Analysis preservation 	Data Preservation - What is preserved (simulated, DSTs, ...) - Location(s) - Access (S3, xrootd, rucio, ...)		Kolja Kauder
	20	Documentation 	Documentation: - Location of User documentation (wiki, repository,...) - Who will set up skeleton with list of topics (e.g. "Getting Started")		Dmitry Romanov
	27	*EICUG Meeting*			
Aug	3	Schedule realignment/ WG Business			
	10	Overview reports	DD4hep, JANA2, CI, ...	*reports*	
	17	Containerization Official builds	Containerization - platform (Singularity, Docker, multi, ...) - Supported OSes - Location of images (e.g. cvmfs) Official builds - Location (e.g. cvmfs, container image, ...)		
	24	Distributed Campaign Workflow	Distributed Campaign Workflow - Package (DIRAC, PanDA, STAR(?), ...)		
	?	Calibration DB Conditions DB	Calibration / Conditions DBs - Package - Server/Host - Access		

S&C Review

Requirements From EIC Statement of Software Principles

1. Aim for rapid turnaround of raw data to online and offline productions
    a. Compatible with streaming readout and near real-time physics ready productions
2. Enable distributed workflows on computing resources worldwide, leveraging HTC/HPC
3. Modular and robust against changes/differences in compute environment
4. Production workflows must be reproducible and re-interpretable
5. Ensure critical software is not dependent on the expertise of a single person
6. Should not “reinvent the wheel” if an existing technology satisfies all requirements laid out by ePIC Computing and Software community
7. User centered design and user friendliness paramount

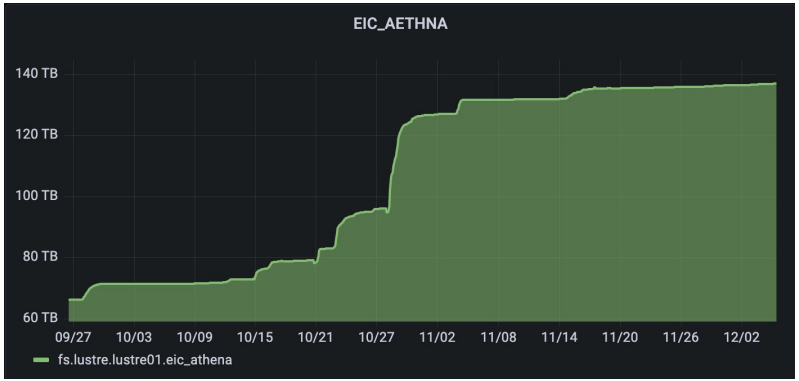
Additional Considerations

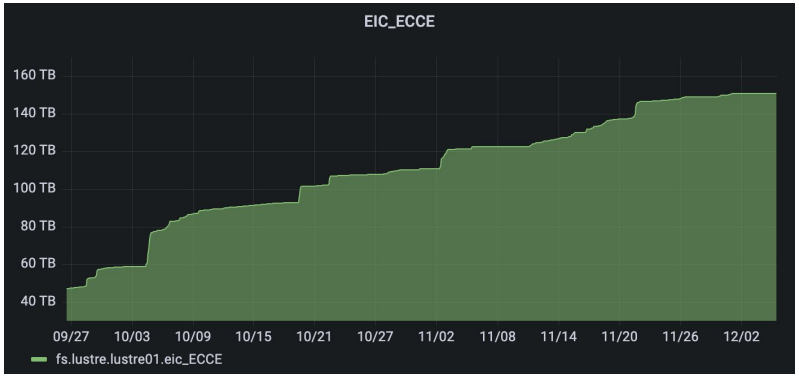
- Scalability
- Take advantage of opportunistic resources, if available
- Compatible with ML workflows
- Remaining questions:
  - Do existing technologies satisfy the requirements?
  - What would a “home grown” solution cost in terms of risk/resources if available technologies don’t satisfy the requirements?

To be discussed and decided on in forthcoming workflows meeting

Strategies During Proposal Period

- ATHENA and ECCE proto-collaborations produced ~300 TB of simulation data, utilizing O(10M) CPU hours for O(100M) of physics events for analysis
- Carry-over - liaisons from Physics/Detector Working Groups to communicate and request productions from SimQA working group
  - Benefited from greater communication to physics working groups and wide variety of Monte Carlo Event Generators utilized for EIC science mission





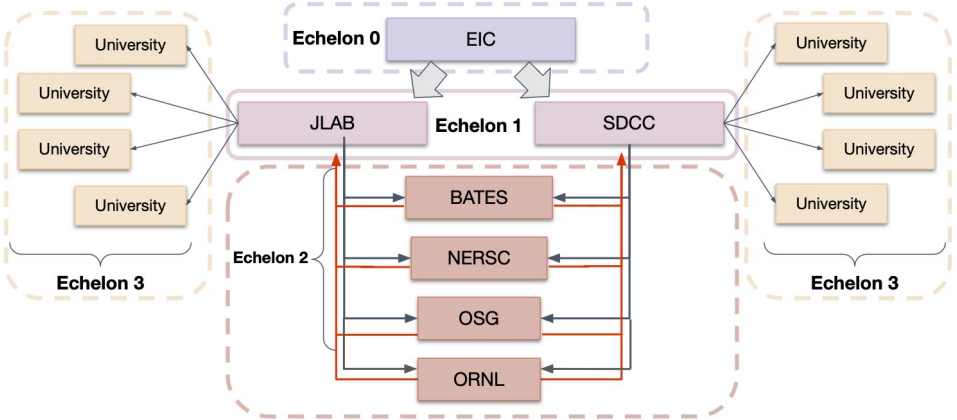
Strategies During Proposal Period

- Both ATHENA and ECCE proto-collaborations developed and deployed in house solutions for producing simulation data
- Series of scripts that could submit jobs to HTCondor (BNL/OSG…) or Slurm (JLab/ORNL…)
- Pros
  - Agile for quick development during condensed proposal timeline
  - Limited configurations necessary
- Cons
  - Not scalable
  - Not easily accessible/maintainable by the community
  - Limited number of developers know how system works

Previous solutions do not meet EIC Software Principles!

Strategies During Proposal Period

- Federated computing architecture deployed by both proto-collaborations
  - Very successful and highly desired moving forward
- WLCG style architecture envisioned, utilizing e.g.
  - Tier 1 sites - BNL and JLab
  - Tier 2 sites - Additional large compute sites, e.g. OSG, NERSC, others…
  - Tier 3 sites - Local universities/small compute sites

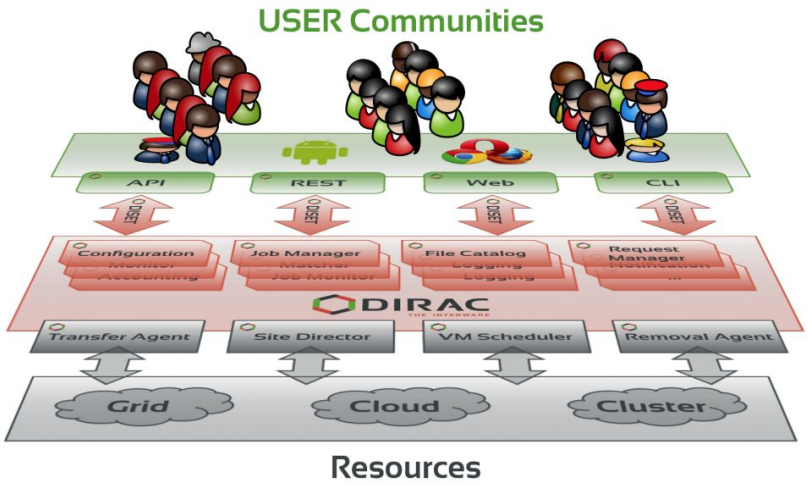


Software and Computing Roundtable

- Dedicated S&C roundtable discussion in May on production workflows
  - https://indico.jlab.org/event/505/
- Presentations and discussions on
  - DIRAC
  - PanDA
  - Parsl
- NHEP community has been developing, discussing, and producing solutions for production workflows in a variety of ways
- EIC software community engaging with broader HEP software through e.g. HEP Software Foundation

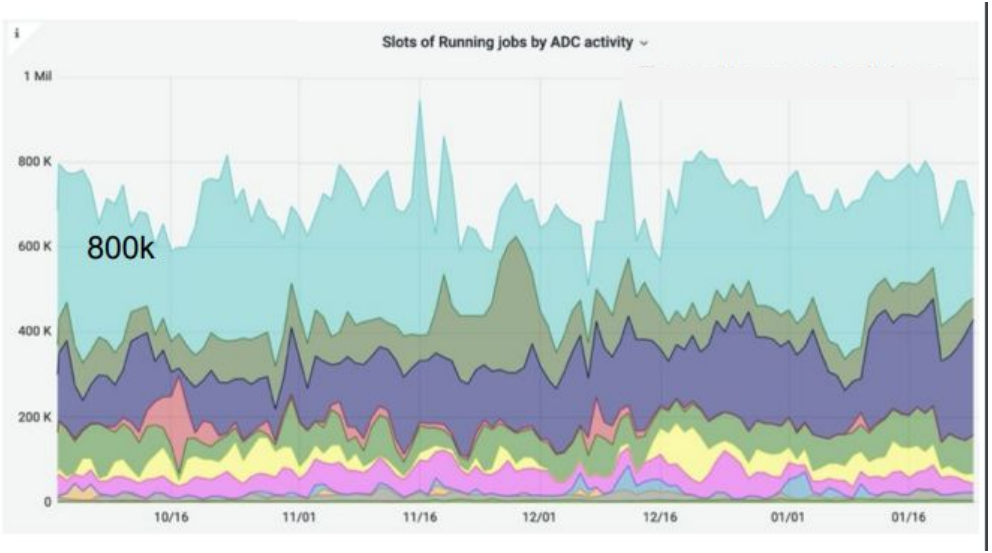
Existing Technology: DIRAC

- DIRAC developed by LHCb experiment at CERN, starting 2009
- Interfaces users to resources for job scheduling and handling
- User base grown since initial development and use at LHCb, e.g. at Belle2 and T2K



Existing Technology: PanDA

- PanDA developed and in use by ATLAS experiment at CERN since 2005
- Manages different types of workflows at range of compute centers
- User base expanded beyond ATLAS to NP (sPHENIX, COMPASS) and astrophysics (LSST)



Conditions Database

- Discussion on conditions database also scheduled after distributed workflow management
- Evaluation of EIC requirements and available technologies/expertise within the community ongoing
- EIC community engaged through e.g. S&C roundtable discussion
- To be discussed:
  - Available software packages? Do they meet EIC requirements?
  - Where are DBs hosted? Who maintains?
  - How is access handled? How will this be integrated into main software stack?

Conclusions

- Following software decision schedule, distributed workflow management system discussion to be held in forthcoming weeks
  - Evaluation of EIC requirements and available technologies forthcoming
- Technical solutions deployed by both proto-collaborations in proposal stage are **not** adequate long term
  - Nonetheless, deployment of federated architecture was successful and desired moving forward
- EIC S&C community has engaged with development teams of available technologies, e.g. DIRAC and PanDA
  - Already experience available in the community