Evolution of ATLAS analysis workflows and tools for the HL-LHC era

D.Cameron, A. Forti, A. Klimentov, A. Pacheco, D. South
on behalf of the ATLAS Collaboration

vCHEP
17-21 May 2021

Analysis Tools Evolution

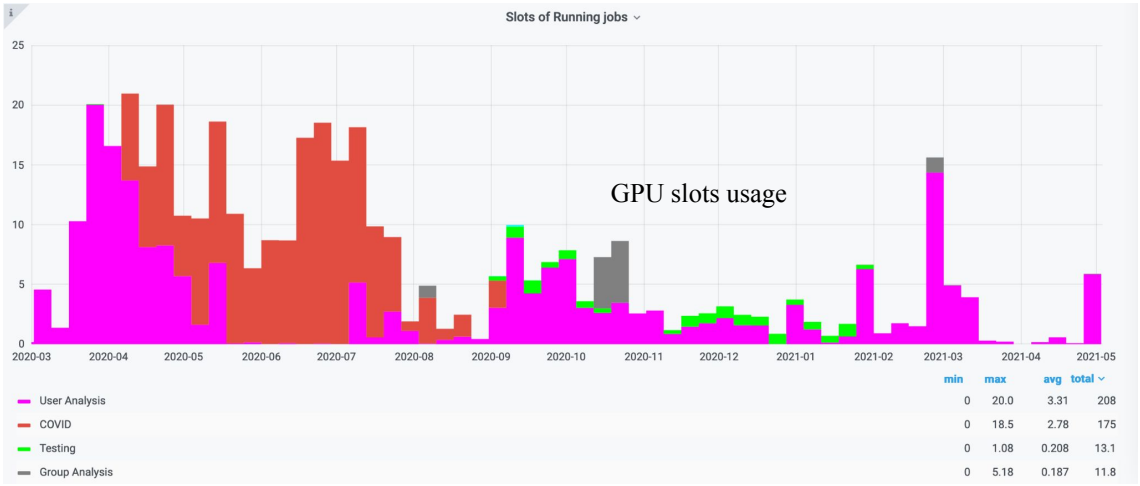
- Data Reduction
  - Reduce amount of data required by end users
    - HL-LHC format for analysis ~10kB/ev
      - ~2PB per year
- Tools evolution
  - Ecosystem developed outside of HEP
    - pandas, numpy, Dask, HDF5, scikit-learn, matplotlib
      - PyHEP building specific HEP tools based on these
  - ROOT ecosystem also evolving to allow for more efficient storage and I/O
  - Data transformation and delivery services
    - Columnar data analysis "array-at-a-time" instead of "event-at-a-time"
  - Machine Learning workflows
- End user analysis can then be highly declarative and carried out on small data formats using the above

Facilities for Analysis

- Possible implementations :
  - Tier-3s evolution - Co-location with T1s and T2s - Single or federated - Dedicated AF (national or ATLAS wide) - DAaaS (Data Analysis as a Service) frontend for Grid - HPC as AF - Commercial clouds for ATLAS analysis
  - Likely a mixture of solutions
- Many questions on access, support and funding models
  - Need to support +1500 ATLAS users
- Grid resources access quite democratic
  - Can this be replicated with interactive distributed AF and more complex ecosystem?
- What characterizes an AF is interactive access to analysis resources, support of services and diverse software.
- Core of R&D is to give opaque access to AF, grid and cloud type of resources

Containers and Authz

- Containers and authz are the underlying glue between different resources.
- Standalone containers give users the flexibility to run what code they want in different environments
  - Most of the AF services developed are container based
  - ATLAS runs standalone containers also on the grid, HPC and clouds
    - Standalone containers distribution still in development
    - GPU enabled workflows for users



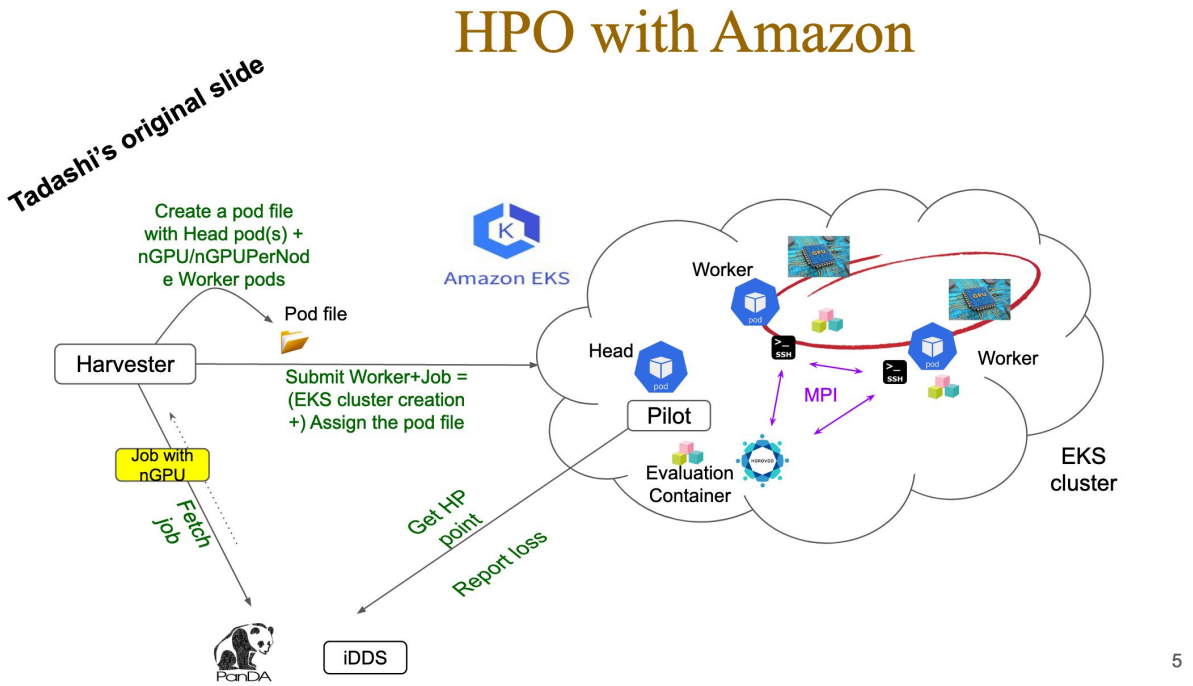
- Oauth2.0 authz is important for the integration of cloud type of technology like kubernetes and jupyter
  - ATLAS services have already token enabled authorization
  - Integration with WLCG tokens missing
  - Users using new workflows will be the first to use tokens

AF Services

- An important part of future analysis will be interactive services and data delivery/transformation services.
  - iDDs, ServiceX (columnar data), Jupyter Hubs, etc
- iDDs is already heavily used in production for streamlined data delivery from tape
  - Has been extended to support analysis iterative and chain workflows
    - HPO, active learning, chain workflows
- ServiceX developed to deliver columnar data to an AF
  - Could be integrated in WFMS
- Jupyter Hubs, REANA, binder
  - Seamless access

An Example

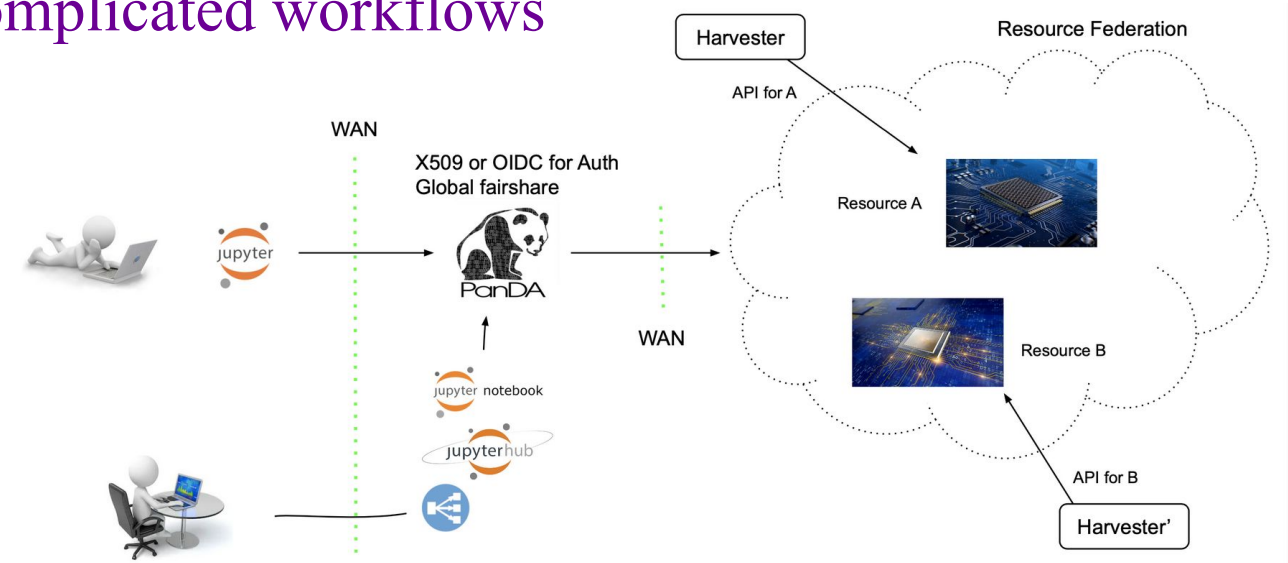
- HPO (Hyper Parameter Optimization) on AWS resources brings several aspects of the current R&D together



- Makes use of iDDS functionalities for iterative workflows
- Is tested on a kubernetes cluster
- ...on the Amazon cloud.
  - HPO workflows started on grid sites GPUs
- This workflow uses also horovod to scale deep learning training to multiple GPUs

Jupyter as UI

- Jupyter NB is an important tool
  - A Jupyter hub is one of the services most commonly mentioned as part of an AF
- Integration with panda clients with Jupyter NB to offer the same interface being prototyped
  - Using the PanDA API as the command line tools
  - Possible to interact also with iDDs to control more complicated workflows



- rucio also has a prototype integration with their clients

Conclusions

- We need to define the analysis facilities best suited for ATLAS and use Run 3 to define and conduct R&D effort to be ready for HL-LHC
  - We will have a mix of AF implementations for Run 4
  - R&D work is to smooth access to completely different resources.
    - Jupyter interface, container integration, kubernetes integration, spill over of machine learning analysis and similar access to the data are all examples of this.
- R&D projects in pre-production during Run 3, some of them have already started
  - AF prototyping
  - Analysis using commercial clouds
  - Jupyter notebook to submit analysis jobs