ATLAS Analytics and Machine Learning Platform

Ilija Vukotic
University of Chicago

CHEP 2018, Sofia, Bulgaria

Getting the most from distributed resources

- ADC (ATLAS Distributed Computing) needs to account, monitor and optimize usage of all computing **services** & **resources** available for ATLAS physics.
- We built an analytics platform that collects information supporting ADC analytics activities

What we want	What we need
To understand the system	A way to easily get global picture
To understand interplay of different systems and services	Collect all the data at one place. Be able to cross-reference.
Debug systems	Ability to drill down to the most detailed information
Run simulation, test models	Programmatic access to all the data
Alerts, sensing services	Continuously / periodically running services operating on raw / derived information fast enough for a real time feedback

ATLAS analytics infrastructure

CERN

- Data Sources, including:
    - file transfer data, dataset usage (Rucio)
    - job information (PanDA)
    - xAOD access information
- Primary purpose - real time monitoring and accounting

CERN IT provides the infrastructure for monitoring & analytics by the ATLAS distributed computing team (ADC)

- **DBs**

  ORACLE, MySql, Hadoop, Elasticsearch, Ingress DB

- **Transport**

  AMQ, Flume, Kafka

- **Processing**

  Pig, Spark, SWAN, Dockerized applications on OpenStack Magnum Kubernetes cluster

- **Visualizations**

  Custom made dashboards

ATLAS analytics infrastructure

ATLAS Midwest Tier2 Center @ University of Chicago

- Additional data sources:
  - Network data from WLCG/OSG
  - PerfSONAR
  - CPU benchmarks
  - IO (per file, ROOT collection)
  - Application logs (Frontier, Squids)
- Processing
  - Inline
  - Offline

University of Chicago and ATLAS provided infrastructure

- DB

  Elasticsearch cluster

- Transport

  RMQ, Flume, custom made collectors

- Processing

  Data collection, enrichment applications on SLATE kubernetes cluster

  Jupyter/Google apps for alarms/alerts.

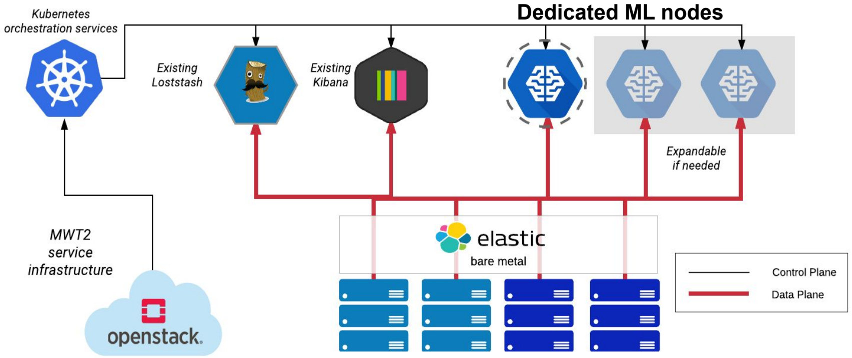
Current platform

Already a lot data in the system (most datasets span last 2-3 years).

Most of data in real or near real time.

Elasticsearch copes with more than 40k queries/s and 10k docs/s ingress.

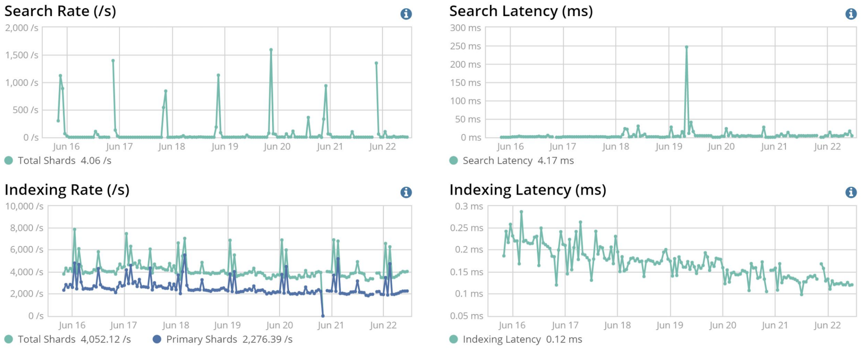
Running several production services.



ATLAS Elasticsearch cluster

Nodes: 15	
Disk Available	65.25% 29.6 TB / 45.4 TB
JVM Heap	43.77% 183.1 GB / 418.3 GB

Indices: 3,673	
Documents	32,503,045,818
Disk Usage	15.7 TB
Primary Shards	18,057
Replica Shards	3,752

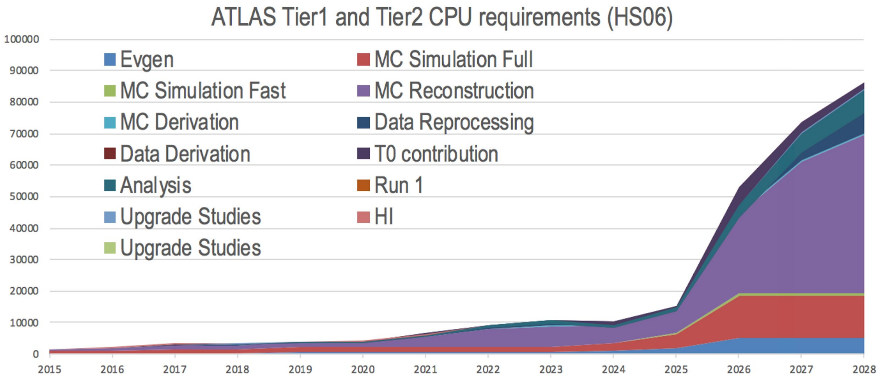


Some analytics studies

- CPU benchmarking (relative ranking, what CPUs are best for the job)
- Data usage (what datasets are popular, what data collections are not needed, how to optimize derivations)
- IO studies (performance of different formats, ROOT options, storages)
- Site monitoring and optimization
- File Transfer System (FTS) optimizer tuning, endpoint/link settings optimization
- Job wall/CPU time efficiency, job brokering studies.
- Network anomaly detection
- Local Cache simulations

Resources crunch - can ML help?

- Making computing operations more efficient:
  - Optimize data placement, job scheduling, data transfers
  - Anomaly detection (processing, data transport,...)
- Improving physics results while reducing computing cost:
  - Calorimeter simulations
  - Clustering
  - De-noising
  - Jet tagging
  - Trigger
  - Tracking
  - DQ monitoring



Methods:
- NNs
  - DNN
  - CNN, LSTM
  - VAE
  - GAN
- BDTs
- Bayesian analysis
- Data Smashing
- GA

Most of this requires special resources.

Already more than 40 users.

ML platform idea

At number of places there are “small” GPU equipped clusters.

Most of them are not user friendly.

Users want:

- “No hoops” access
- All the tools setup and functioning
- Not only support but tutoring

Have a distributed kubernetes cluster unifying resources scattered around.

- Pros
  - Frees sysadmins from software support (apart from k8s and drivers updates)
  - Gives users freedom and scalability
  - Central management of services (authentication, monitoring, storage, caching, etc)
  - Cheap - uses existing resources
- Cons
  - Applications have to be containerized

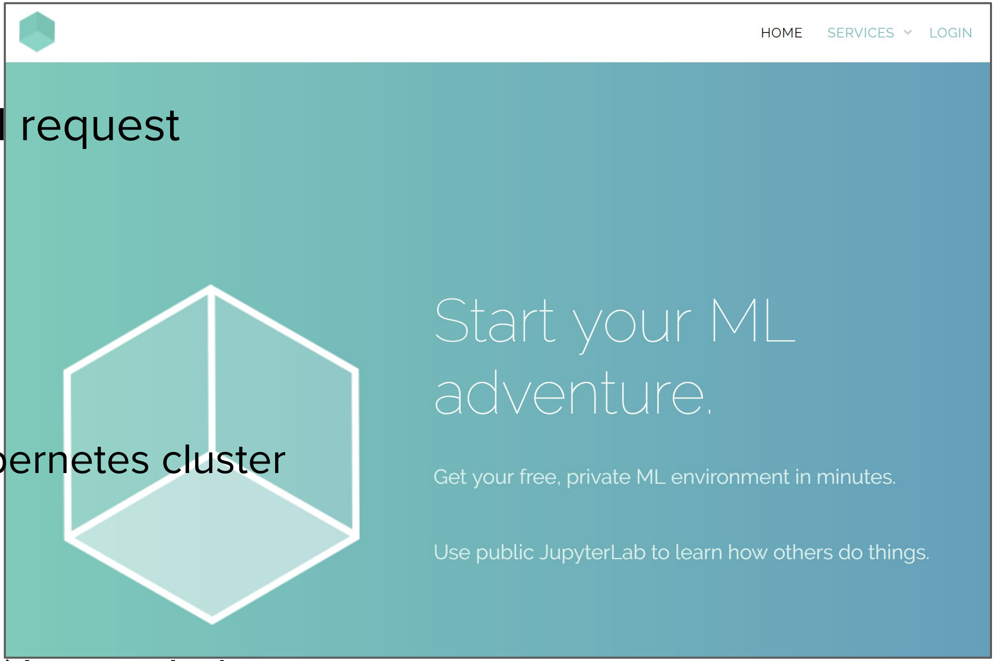
ML portal

A single point of entry for all ML/analytics needs.

Using Globus login.

Allows users to spawn applications and request resources:

- #cores/#GPUs
- Storage
- Duration



Applications are created on a federated Kubernetes cluster (currently SLATE)

Provides monitoring/logging.

Provides several frequently used apps (JupyterLab, ROOT)

Summary

ATLAS has a production level analytics infrastructure at both CERN and UChicago.

All services containerized and deployed in k8s clusters.

A lot of data from 40+ data sources collected in real or near real time.

Large number of studies done on both computing and physics side.

Machine Learning Platform is being developed to provide easy access to both data and hardware resources.

Questions or Comments?