Introduction to the AF forum

Analysis Facilities Forum Kick-off

25th March 2022

Practical information




Conveners:

Alessandra Forti (ATLAS)

Lukas Heinrich (ATLAS)

Diego Ciangottini (CMS)

Nicole Skidmore (LHCb)

Mailing lists:

HSFAFFORUM

Mattermost:

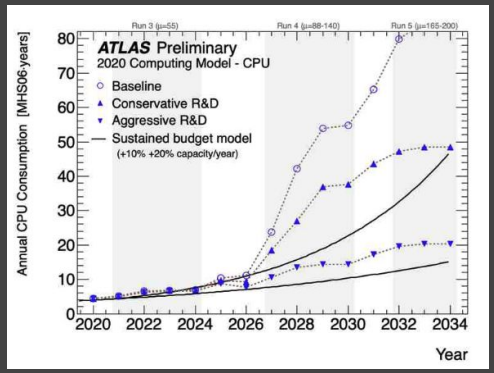
hsf-af-forum

Why is this important now?

LHC future

HL-LHC (2027) will see orders of magnitude more data taken by the LHC experiments delivering an unprecedented scientific data volume at multi-exabyte scale

TL-SOFT-PROC-2021-010



- Current LHC computing model will not provide the required data processing capabilities even with foreseen hardware evolution
- Current “local” end-user data analysis methods and corresponding tools will not scale with event count
- Sharing and optimising the efficient use of specialized infrastructure (increasingly used in end-user data analysis) will become more and more important

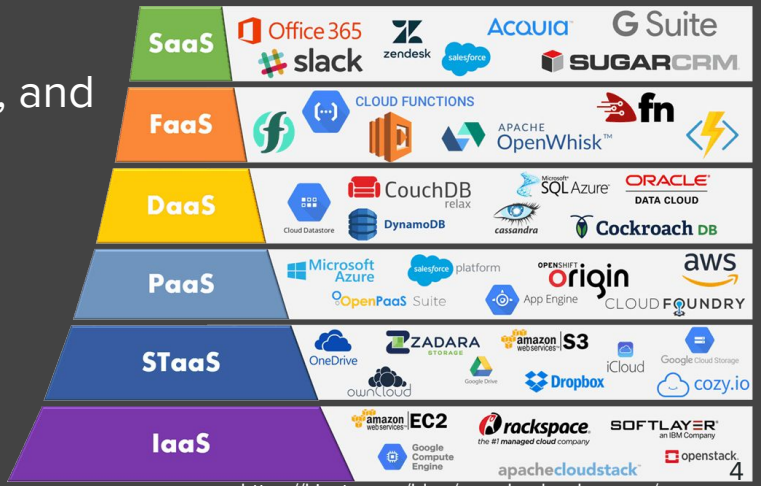
Future experiments with big data

- Dune, LSST, SKA, EIC - share experience, innovation and resources

Why is this important now?

Technologies evolution

- New technology developments in DOMA areas
  - Development of innovative data delivery services
  - Advanced caching techniques
- Switching to use bearer token eg. SciTokens authorisation
- Integrating different types of resources/"aaS" available on demand: public and private clouds, and Kubernetes (k8s) with increased analysis containerization
- Evaluation of successful AF characteristics



https://kinsta.com/blog/google-cloud-vs-aws/

*DOMA: Data Organization, Management and Access

Why is this important now?

New analysis techniques

- Development of new analysis workflows for more efficient analysis
  - Columnar analysis instead of traditional event loop approach
  - Workflow management tools
  - Increased use of alternative hardware
- Increased interoperability between ROOT and Python data-science tools
- Integration and adoption of industry tools in HEP analysis frameworks
  - Including machine/deep learning algorithms and sophisticated analytics engines like Apache Spark, DASK
- New user interfaces for interactive analysis (e.g. Jupyter notebooks)
- Analysis preservation - REANA platform

What will this forum achieve?

- This forum will bring together invested parties for dedicated, **technical** discussions **on a bi-weekly basis**

- The goal is to build/foster a **technical community** and **serve as a “bridge” between various involved parties**: experiments, software stakeholders, data centres, WLCG, IRIS-HEP and HSF

- Ultimately this forum would lead to the evaluation of proposed solutions and a white paper after one year outlining the community vision for future AFs, especially designed for HL-LHC scale analysis

Meetings proposed and next topics

- AF infrastructure requirements - from both analysts, developers and data centres view points - what is the Venn diagram?
- Industry approaches for scalable analytics (Dask, Apache Spark, Ray frameworks) and its adoption in HEP (in general) / AFs
- HEP analysis frameworks (state of the art, perspective, integration and adoption by AFs)

Related events

Analysis Ecosystems workshop II May 2022

Glossary

- DOMA: Data Organization, Management and Access
- WLCG: Worldwide LHC Computing Grid
- HTTP: Hypertext Transfer Protocol. HTTP is the protocol used to transfer data over the web. A typical flow over HTTP involves a client machine making a request to a server, which then sends a response message.
- HTTPS: Hypertext Transfer Protocol Secure - the secure version of HTTP used for secure communication over a network
- SciTokens: The SciTokens project builds a federated ecosystem for authorization on distributed scientific computing infrastructures.
- IAM: Identity and Access Management
- OIDC: OpenID Connect. An authentication protocol which verifies user identity when trying to access a protected HTTPs end point.
- aaS: “as a Service”. Eg. PaaS = Platforms as a Service, SaaS = Software as a Service
- Kubernetes: (k8s) is an open source container orchestration platform that automates many of the manual processes involved in deploying, managing, and scaling containerized applications.
- Apache Spark: Apache Spark is an open-source unified analytics engine for large-scale data processing.
- Dask: flexible library for parallel computing in Python. Similar to Apache Spark but integrates with existing Python tools.
- Ray: Ray is a high-performance distributed execution framework targeted at large-scale machine learning and reinforcement learning applications
- REANA: reusable and reproducible research data analysis platform
- HSF: HEP Software Foundation
- ServiceX: ServiceX is a data extraction and delivery delivery service
- XCache: cache-based data approaches to increase efficiency of CPU use (via reduced latency) and network (reduce WAN traffic)
- Data lake: storage service geographically distributed across large data centers connected by fast network with low latency. Alternative to running jobs at site where files are located.

What is an Analysis Facility?

"infrastructure and services that provide integrated data, software and computational resources to execute one or more elements of an analysis workflow. These resources are shared among members of a virtual organization and supported by that organization."

People

Dedicated support staff

Maintenance and development

Services

Access to experimental data

Storage space for per-group or per-user data

Access to significant computing resources

Hardware

CPUs and disks

Growing need for GPUs

Software

ROOT

Python-based ecosystem

Interactivity