Content delivery and caching

The DOMA ACCESS Working Group

Xavier Espinal on behalf of the WG

WLCG-DOMA F2F, FNAL 9 March 2020

Working Group Activity (1/3)

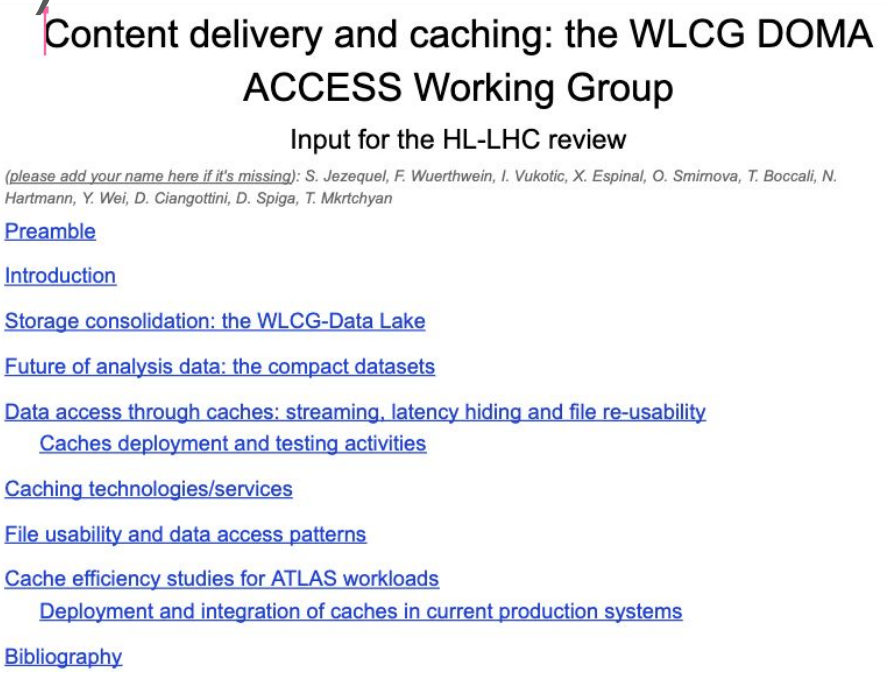
- Analysis evolution
  - discussions on the future of compact analysis datasets and analysis models evolution with HSF-Analysis WG
  - Joint workshop **HSF-Analysis and WLCG-DOMA ACCESS** working groups at CERN/Vydio
    - 17 March: https://indico.cern.ch/event/890991/
    - Some of the topics to be addressed:
      - Analysis: distributed CPU facility vs. local farm
        - Needs for analysis tasks (especially when largely different from reco/sim jobs)
        - Integration with services such as SWAN, jupyter, gitlab-ci, you-name-it
      - Existing, planned and new **infrastructures**
        - Usage patterns:
          - e.g. how often do you read xx events, how quick you need to get the result back, how much cpu time per event, which fraction of events is sufficient 90%, 99.9%?
        - Size of the input data that should be pre-staged/hot
        - How much would that data be shared with other analysis
        - Planning of resource usage vs free-bar

Working Group Activity (2/3)

- Caching infrastructures
  - following-up the status of the different deployments for ATLAS and CMS in US and EU
  - Several months of operation. Starting to get enough experience to evaluate from the performance point of view and from the administration point of view.
    - SoCal, LMU, BHAM, IT/INFN, i.e.:
      - BHAM is running production jobs with xcache (this is interesting because it was running simulation which used the same input file 5 to 10 times) .
      - LMU has tested production and analysis.
      - Dedicated talk form GridPP in two week time: https://indico.cern.ch/event/895815/
  - Introducing basic costs estimation (benefits?) with the collaboration of the cost model WG
    - First costing model to balance storage and networking costs presented at CHEP: https://indico.cern.ch/event/773049/contributions/3474404/attachments/1937844/3212558/DataAccessPatternsV2.pdf

Working Group Activity (3/3)

- Input to the HL-LHC review [doc]
  - Document is in good shape. Many contributions, both via comments and writing
  - Full edit mode, please do not hesitate to contribute (I tag versions)
- Confident to have a decent version by the deadline (end of the month)
- The input document for the HL-LHC review will be the basis for a white paper aimed to conclude the activity of the WG with recommendations for future areas of research/development



Content delivery and caching: the WLCG DOMA ACCESS Working Group

Input for the HL-LHC review

(please add your name here if it's missing): S. Jezequel, F. Wuerthwein, I. Vukotic, X. Espinal, O. Smirnova, T. Boccali, N. Hartmann, Y. Wei, D. Ciangottini, D. Spiga, T. Mkrtchyan

Lund Workshop preparations

- Present main aspects of the future white paper:
  - Experiences from the different caching infrastructures deployments
  - Foreseen evolution of data analysis: “analyzers” needs and impact on infrastructures
  - Foreseen evolution of the computing models, specially on compact data objects
  - Caching technologies and strategies
- Consolidate collaboration effort between HSF-Analysis and DOMA-ACCESS
  - Session on evolution of analysis and infrastructures for analysis

Q42020: re-scoping the ACCESS Working Group

- Start discussing new focus of the working group to also tackle the **WLCG-datalake**
  - The WG was created to address the data access for the analysis use case
    - We based our data access studies on a datalake-like model as the source of the data. Nevertheless this datalake-like was never discussed nor addressed in detail (strawman model was written to have a model to reference)

- The scope of data access studies should be enlarged now to cover **full picture** of data storage, data distribution and data access Investigating the combined impact of workloads and their requirements on the infrastructure.
  - Data orchestration:
    - Policies, data replication, redundancy,...
  - Data access:
    - Analysis, reprocessing, simulation, HPC, clouds,...
    - Caching infrastructures, Analysis facilities,...