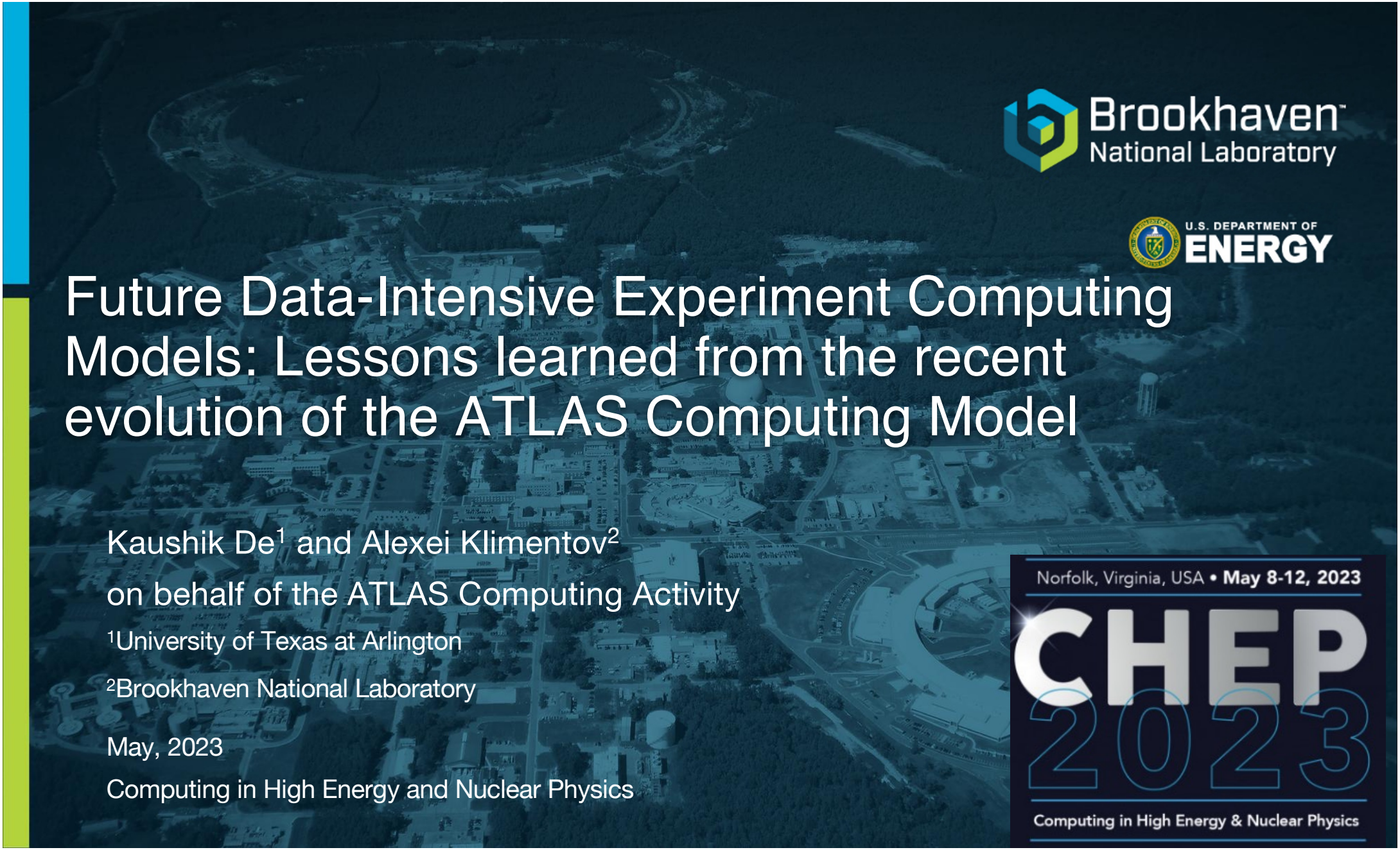
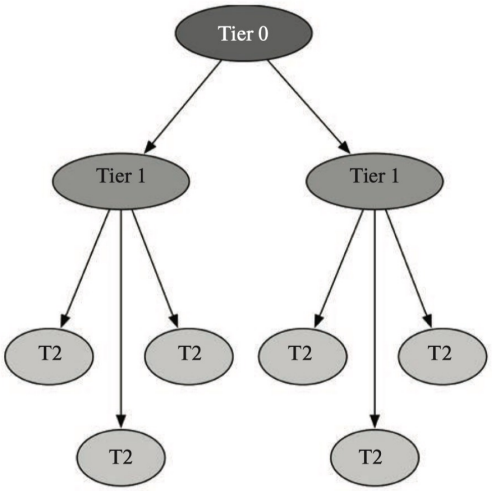

Brookhaven
National Laboratory
U.S. DEPARTMENT OF
ENERGY
Future Data-Intensive Experiment Computing Models: Lessons learned from the recent evolution of the ATLAS Computing Model
Kaushik De[1] and Alexei Klimentov[2]
on behalf of the ATLAS Computing Activity
[1]University of Texas at Arlington
[2]Brookhaven National Laboratory
May, 2023
Computing in High Energy and Nuclear Physics
Norfolk, Virginia, USA • May 8-12, 2023
CHEP 2023
Computing in High Energy & Nuclear Physics

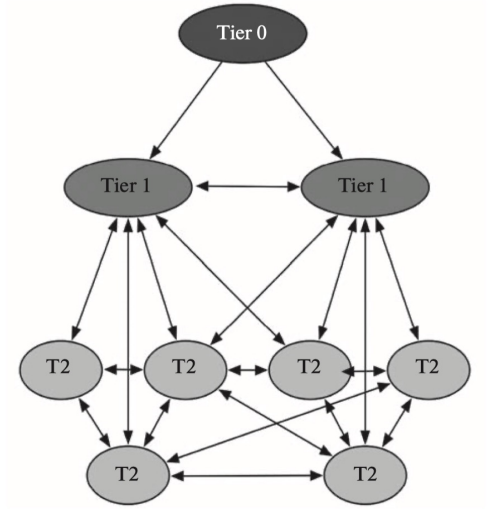
Outline

- HEP Computing Model Evolution
- Scientific workflows
- Challenges and R&D projects
- Computing Model for tomorrow
- Summary and Conclusions

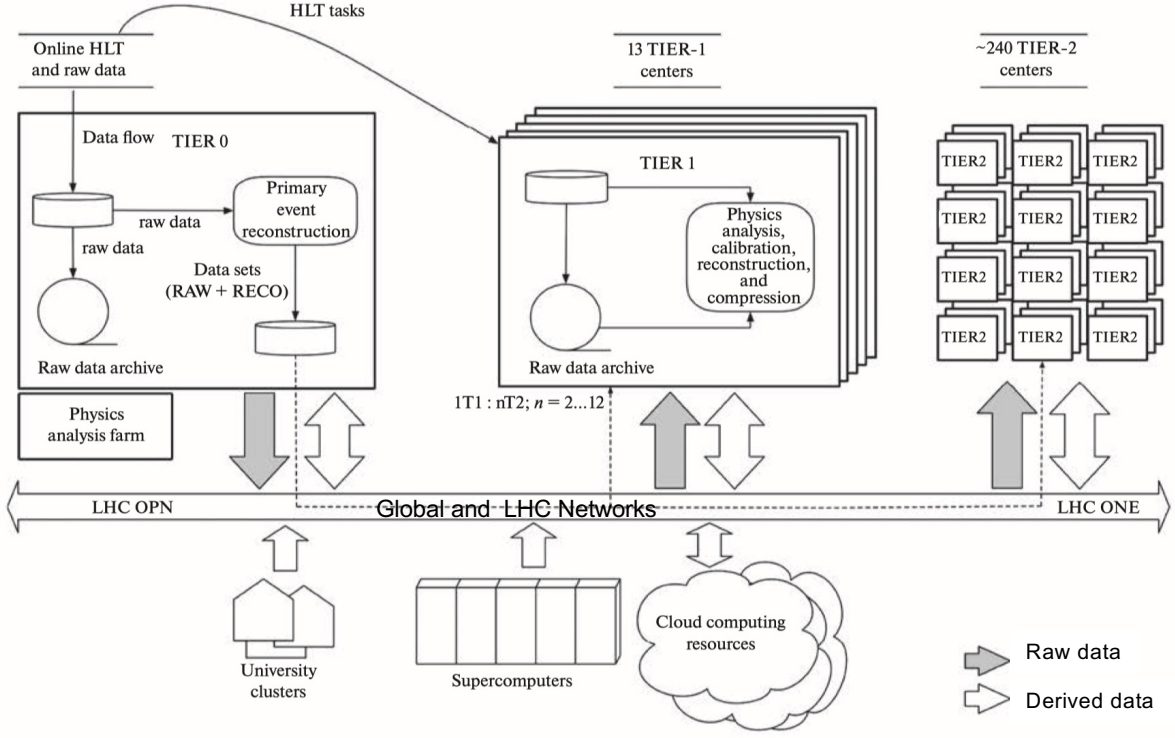
HEP Experiments Computing Model Evolution



Hierarchical Computing Model (LHC startup)



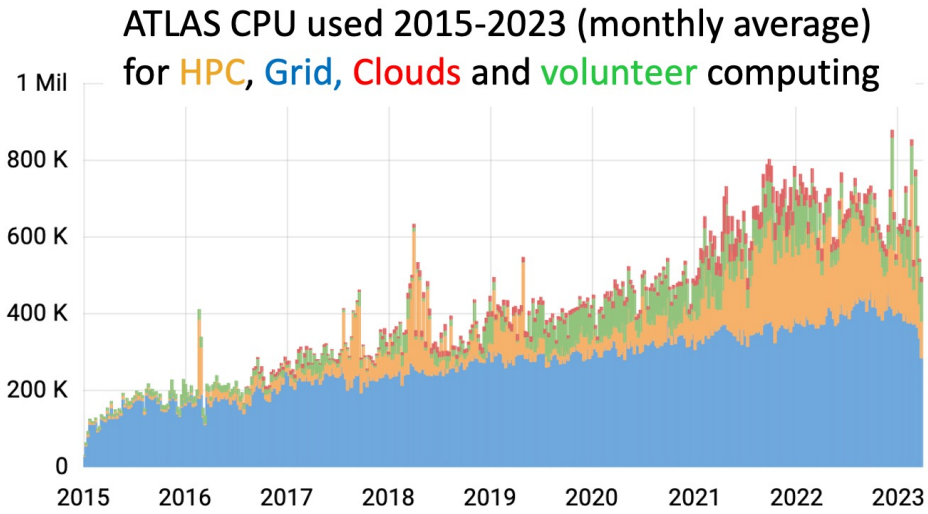
Mesh Computing Model (LHC Run 1)

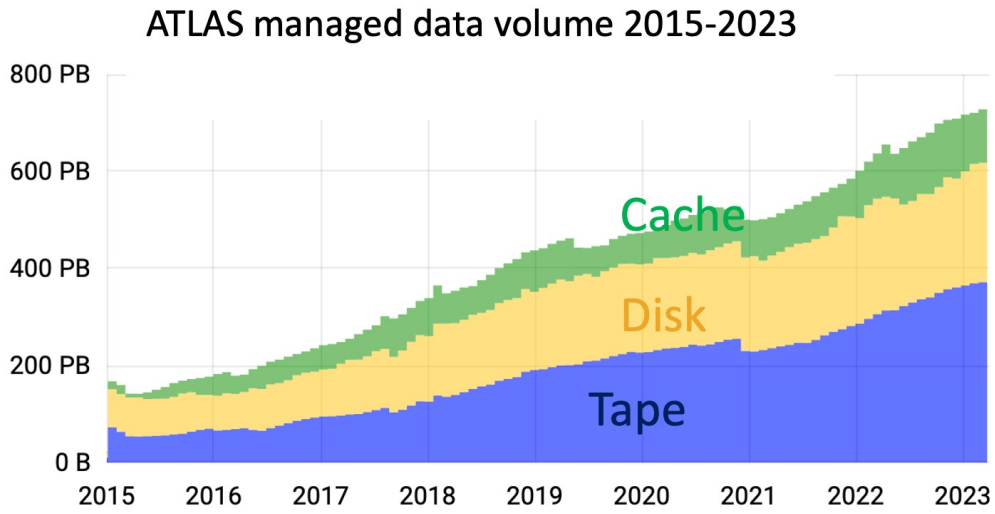


Computing model implemented for the LHC Run 2 and Run 3

LHC Run 1 : 2010-2012; Run2 : 2015-2018; Run3 started at 2022

Workflow and Data Management Challenges

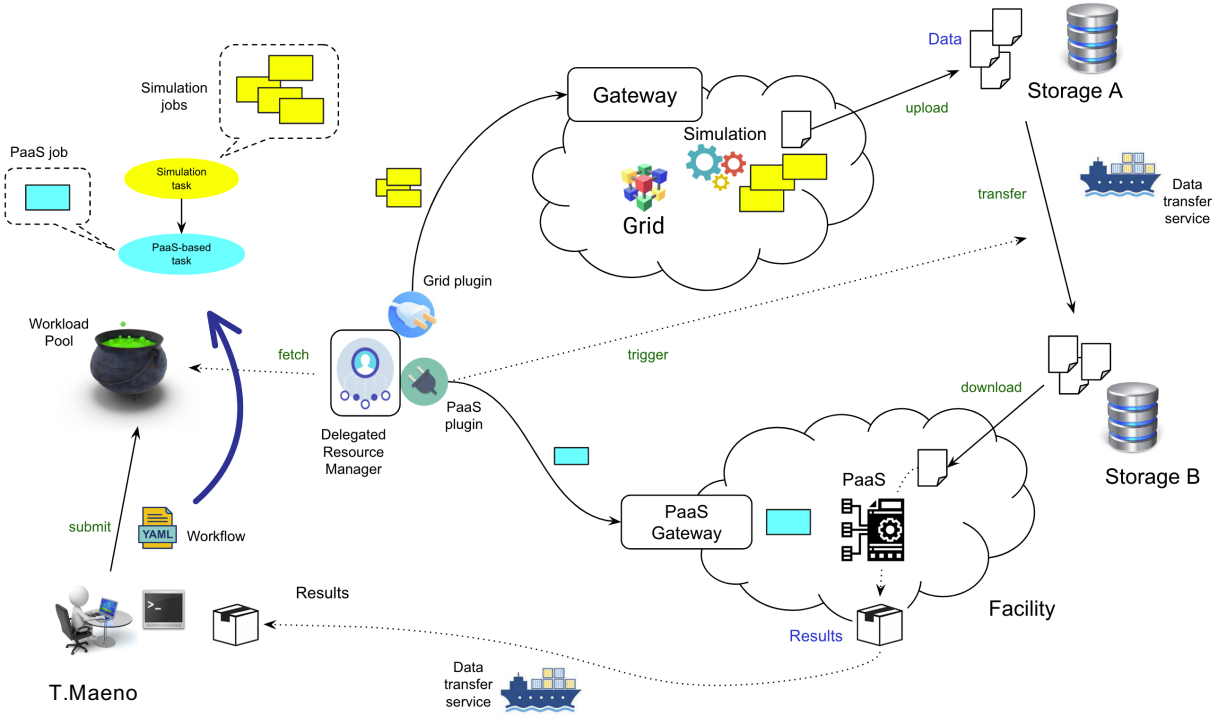




- ATLAS scale 2015-2023
  - Left plot : monthly average number of running cores for various resources
  - Right plot: managed data volume
  - Heterogeneity of computing resources increased dramatically after/during Run2 (2018)
- 1M+ payloads per day are executed via supercomputers, grids, and clouds and managed data volumes close to 0.8 EB (just for one of LHC experiments). This was not the case several years ago when grid computing was deemed a “universal solution” for HEP
- Data volume and number of payloads are not the only challenges we have today → Workflow complexity is growing

Workflow Complexity in Scientific Computing. High Throughput Complex Workflow

- Unlike a few years ago, scientific workflow is no longer presented in the form of a directed graph of operations statically mapped to available resources. Scientific workflows have become increasingly complex, time-varying, and more distributed, with AI/ML components, sophisticated iterative chains of MapReduce tasks, and platforms such as REANA
- Today, we are entering not only the exascale era, but also a new era of workflow complexity in scientific computing. A complex workflow addressing extraordinary scientific needs[1]



This sample complex scientific workflow demonstrates the challenges of the dynamic and elastic nature of data and workload management and optimization. WMS PanDA schedules each step to a suitable resource or service based on its requirements using relevant plugins. Once the first step is completed, a data transfer service transfers the output data of the first step to the facility where the second step is scheduled. The final results are delivered to the user as soon as the second step is completed.

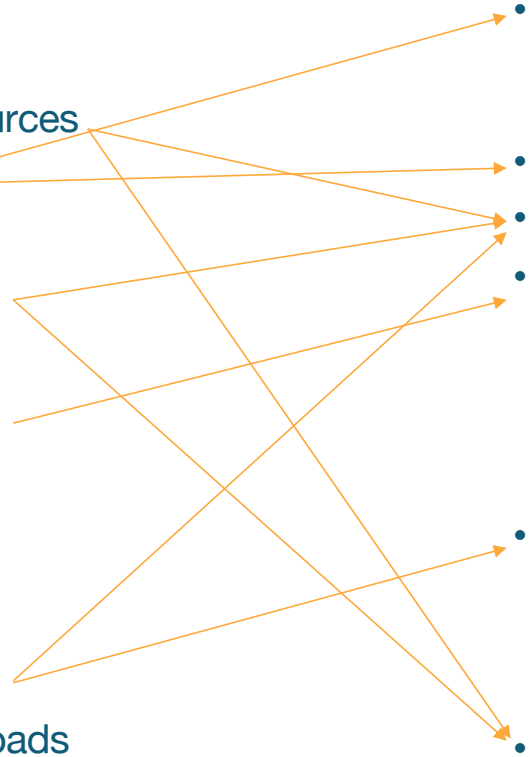
*Ref T.Maeno CHEP23 talk "*Distributed Heterogeneous Computing with PanDA in ATLAS"

B. Brown, "A Vision for the ASCR Facilities Enterprise," presented to the advanced scientific computing advisory committee September 2021.

Today's Landscape

Challenges

- Known challenges :
  - Heterogeneity of Computing Resources
  - Unprecedented data volume
  - Millions of payloads per day
  - Workflow complexity is increasing
- Unknown challenges
  - Computing infrastructure evolution
    - Analysis Facilities
    - Data storage and Data Lakes
    - Commercial Cloud providers
    - …
  - Physics Analysis Model
  - New architectures and hybrid payloads
  - New technologies (DB, BigTables,…)



Work on Scenarios and R&D projects

- Data Carousel
  - A new role for tapes. Use tape as 'cold storage' (in addition to archival storage)
- Data Popularity and data on demand
- Complex workflows brokering algorithms
- Evaluate different technologies and design technology agnostic software
  - Including database backends
  - Commercial clouds
  - Opportunistic computing resources
- Evaluate new tools for analysis
  - Dask
  - Jupyter notebook
  - Commercially available tools
- Resources, data and workload management modelling

R&D projects to address future challenges

- Workload and data placement strategy
  - Investigate/Model how data and workloads can be better distributed between computing sites
    - What is the optimal way to distribute large (EB) data samples based on data popularity information ?
    - What is the best way to match different types of resources and workflows to increase overall system resilience and performance ?
    - What is optimal data granularity ?
      - Dataset (collection of files) → File → Event
        - Adopt a streaming paradigm and optimize the granularity of data moved, stored, and processed by scientific workflows
- Workload and data placement algorithms
  - Ultimate goal
    - Better exploit the architectural features of modern (and future) heterogeneous computing infrastructure by developing payload-brokering and partitioning algorithms for complex workflows and exabyte data volume
  - Go from heuristic based approach to ML and more sophisticated Deep Reinforcement Learning algorithms
  - Data on-demand
    - Keep only popular data and be ready to delete and recreate on-demand unpopular ones
      - Trade in on time to recreate data vs cost to keep (including time to stage data from tapes)

Computing Model for Tomorrow (and day after tomorrow) I

- A fundamental question for the development of a computing model in the field of particle physics is how data will be processed, analyzed, and simulated in 10 years.
  - *This question is not new*
- When answering the question, it is necessary to consider budget constraints, in almost all countries, for increasing computing power for scientific experiments
- Until recently, the computing model was built under the assumption that experiments were the “owners” of the computing resource. The work of many groups in different countries in recent years has been aimed at showing how computing infrastructure not owned by experiments and/or associated computer centers can be efficiently used and integrated with the Grid computing system.
  - *Probably Vera C. Rubin and ATLAS are the most advanced with integrating commercial clouds with on-prem resources*
  - *Also National Agencies in several countries (Australia, Canada, CH, Japan, USA,…) successfully evaluated/deployed model where resources are shared by academic communities*
    - *Including commercial resources*
  - A lot of work is being done by the HENP community to adapt the distributed SW stack to be able to use non-grid resources - without ending up in commercial lock-in.

Ref F.Barreiro CHEP23 talk : “Accelerating science: the usage of commercial clouds in ATLAS Distributed Computing”

Computing Model for Tomorrow (and day after tomorrow) II

- The possible answers to this question are as follows:
  - experiments will continue to buy the necessary hardware and expand their computer infrastructure.
    - An obvious advantage is the advantage of the owner of the resource: the resource can be used and available at any time; and this advantage should be considered only if there is a sufficient resource at the time of maximum load (campaign of analysis and data processing); the rest of the time the computing resource will not be used in full.
  - experiments will buy computing power from those who provide it on a commercial basis:
    - An advantage of this approach is that the capital costs are borne by a third party.
    - A disadvantage is the lack of guarantees that the resource will be available in the required volume or available for use when required and the need to trust a third party and provide it with access to data from international collaboration.
  - A compromise is the option in which the basic resources belong to the experiments and, at the moment of maximum load, providers of computing accounts and services are used.

Summary and Conclusions

- Modern scientific facilities produce large volumes of data approaching exabytes. Data processing involves the ingestion of large data volumes using complex, distributed workflows executed across multiple systems.
- Heterogeneity of computing resources increased dramatically after LHC Run2
- Many HEP and astronomy experiments, as well as National Agencies, demonstrated how commercial facilities and supercomputers could be used by academic community and transparently integrated with Grid resources.
- The future computing model will address that the basic resources belong to the experiments (research community) and, at the moment of maximum load, providers of computing accounts and services are used.
- To address future challenges
  - New data and workload management algorithms should be developed to optimize workflow resilience
  - Set of R&D projects will help us to answer many questions, but we need modelling tools for our distributed computing and software stack

Acknowledgements

This talk drew on presentations, proposals, discussions, comments and input from many. Thanks to all, including those we've missed.

Thanks to members of the PanDA, Rucio and BigPanDAmon projects, ATLAS Distributed Computing colleagues, ATLAS colleagues R.Jones and I.Trigger for reviewing and valuable comments, colleagues from BNL Computational Science Initiative : A.Hoisie and SJ Yoo, KT Lim (SLAC) and S.Klasky (ORNL)