Other Lab and Facilities R&D Projects - BNL

Eric Lancon, Doug Benjamin, Vincent Garonne, Chris Hollowell, Qiulan Huang, Jerome Lauret, Tejas Rao, Ofer Rind, Alex Zaytsev

Nov 7, 2022

A Coordinated Ecosystem for HL-LHC Computing R&D (7-9 Nov 2022)

https://indico.cern.ch/event/1203733/

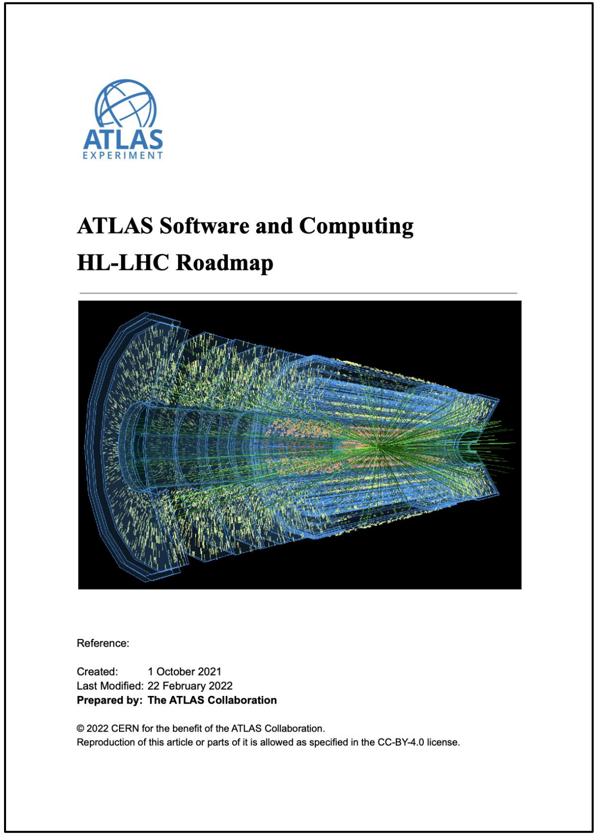
Challenges for Efficient Facility Operation into HL-LHC Era

- Managing anticipated hardware volume for HL-LHC is going to be challenging for facilities, in particular (disk) storage
- HEP solutions fall behind current trends and may come with additional costs in a multi-program environment (ex: Python ecosystem not widely adopted, Grid technology, etc…)
- Requirements for Federated Identity and compliance with cyber regulations may be challenging

Brookhaven National Laboratory

Hardware volume and budget

- Budget exercise for US ATLAS Tier-1 into the HL-LHC era
  - Internal costing model applied to ATLAS hardware forecast
  - Costing model provides qualitative budgetary assessments into Run4, derived from hardware requirements
  - Not-surprisingly, costs at Tier-1 facility driven by storage


ATLAS EXPERIMENT

ATLAS Software and Computing
HL-LHC Roadmap

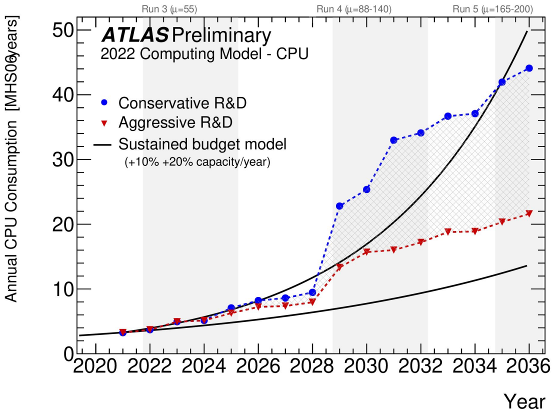
Reference:

Created: 1 October 2021
Last Modified: 22 February 2022
Prepared by: The ATLAS Collaboration

© 2022 CERN for the benefit of the ATLAS Collaboration.
Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.

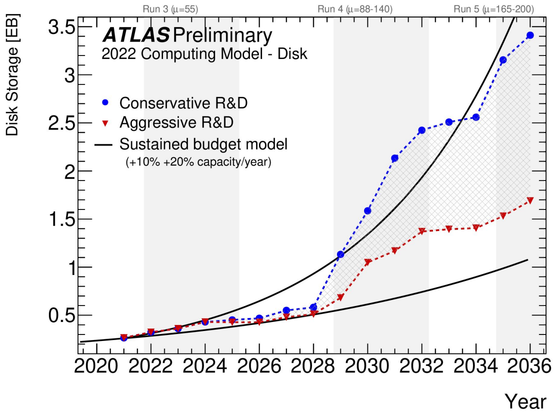
Hardware volume profile into HL-LHC era

CPU



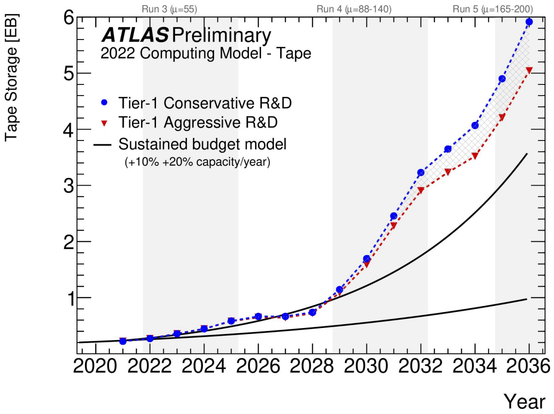
2030: **3** x 2023

Disk



2030: **3** x 2023

Tape

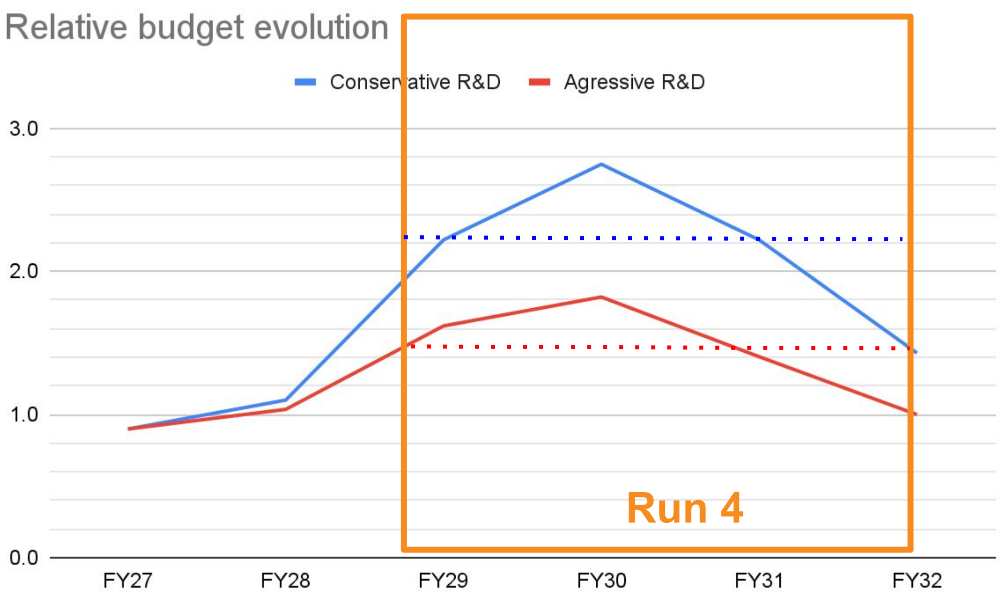


2030: **4** x 2023

Analysis not included

E. Lancon et al., A Coordinated Ecosystem for HL-LHC Computing R&D (7-9 Nov 2022)

Budget profile



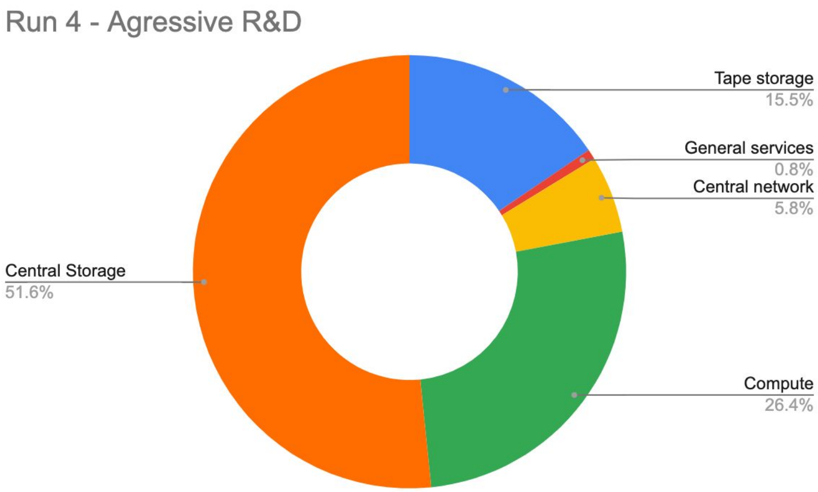
Flat = average FY27 & FY28
Conservative : 2.2 x Flat
Aggressive : 1.5 x Flat

Storage is the most costly resource

How to reduce budget requirement for (disk) storage?

- **Store less** (requirement is 3x RAW data volume)
  - Address event size (content and improved compression)
  - Versioning,
  - Replication policies.

- **Store differently**
  - Use of different storage technologies tailored for each usage,
  - Currently one class of storage for all types of data and usages



Extreme Compression for Large Scale Data Store

Jérôme Lauret[1*], Juan Gonzalez[2], Gene Van Buren[1], Rafael Nuñez[2], Philippe Canal[3] and Axel Naumann[4]

Brookhaven National Laboratory

Store differently

- Current disk storage:
    - Filled with warm/cold data
    - All data types are treated the same, even if they have very different values (DAOD have much higher value than logs, Experimental Data has more value than Simulation, …)
    - All data types are expected to be available immediately everywhere
    - Designed for IO while most applications are not IO limited or critical
    - Not even optimized for IO intensive applications like interactive analysis
- More optimal foundation for supporting HL-LHC activities would be:
    - Bulk storage : Object store (better scaling, operational benefits, globally accessible, …)
    - IO intensive: dedicated POSIX storage - high IOPS design
    - Archive/Cold storage: backup/frozen data
    - And a tiered storage solution to effectively leverage storage “classes”

Storage matching workflows

- Different workflows have different storage requirements
  - Production workflows typically spend more time on processing than IO operations
    - Capacity is a more important criteria than IOPS
    - Entire events are read into memory and processed. The IO access pattern is different from user analysis workflows
  - User analysis workflows tend to require more IOPS
    - The IO access pattern is different from reconstruction or simulation. Users use only part of the event record and more random access pattern.
    - IOPS instead of Bulk capacity is the most important optimization criteria.
- Columnar Analysis workflows should benefit from High IOPS flash storage (SSD/nvme)
- New storage architectures <-> new access methods

Brookhaven National Laboratory

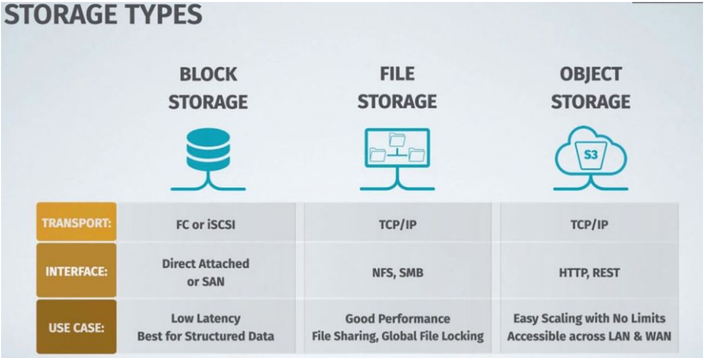
Takeaway

- One type of storage for all is not optimal and likely will not scale into the HL-LHC era (3 x today's disk space)
- Operational costs need to be considered as well… not done today.

Brookhaven National Laboratory

Object Storage at SDCC

- EIC, CFN & NSLS II using Object storage and accessing it via S3 using MinIO.
  - 5 PB of usable storage allocated.
  - Millions of objects. Size varying from few bytes to GBs.
- Advantages of Object storage
  - Massive scalability - Can scale to 100’s of billions of files.
  - Reduced cost compared to traditional RAID filesystems.
  - Can be accessed from everywhere i.e. Ease of sharing of data, high data security using Federated access to storage.
  - Loose coupling of clients.
- Disadvantages -
  - IO interface is the primary drawback.
  - IO throughput performance lower compared to traditional filesystems like GPFS/Lustre.
  - Data reorganization may be needed but modifying data is tedious,



Brookhaven National Laboratory

Storage Usage Effectiveness

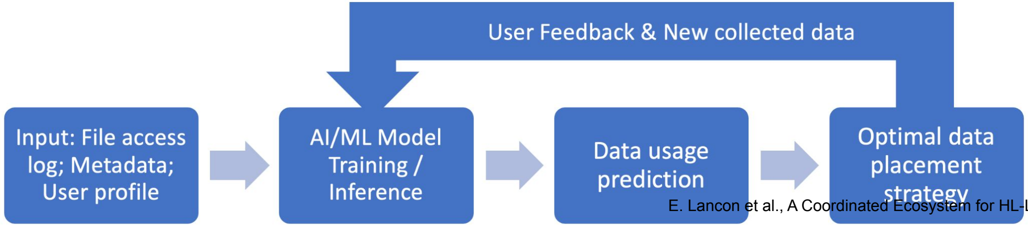
LDRD 2022-2024: Qiulan Huang (PI, SDCC), V. Garonne (SDCC), Al Kagawa (CSI), Xin Dai (CSI)

Motivation

- In the current multi-tier storage "class” system at the Data Center:
  - Unused data is stored on expensive storage
  - Fast IO storage is not currently used

Goals

- Design an efficient monitoring platform
- Develop an optimal data management system for the data center to maximize usable space while minimizing access latency, within budget, hardware, and compliance constraints
  - Heavy use of storage, metadata and data popularity information
  - Detect early failures and pathological usage pattern
  - Develop a precise AI/ML prediction model to possibly forecast the future usage of the data
  - Orchestration of data for optimal movement and placement



The new ecosystem – and user tools

- Jupyter / Python
  - Jupyter initially deployed at BNL for non-LHC projects
  - Light source, Belle II, ‘long tail’ of science

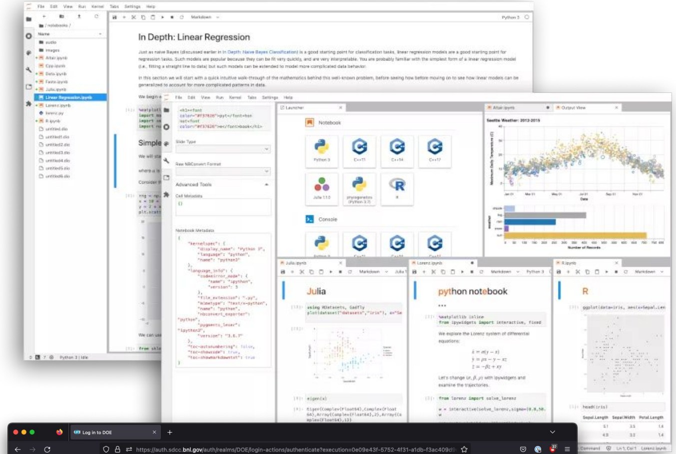
- Containers
  - Non-LHC projects are the drivers
  - For HEP/NP: Reana, ServiceX deployed at BNL

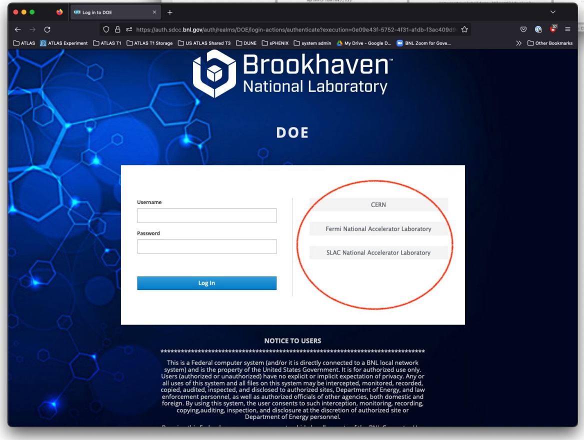
- Federated Identity
  - A requirement today
  - BNL’s Jupyter instance accessible with non-BNL credentials (exception to DOE O142.3B)

Evolution of User Analysis Tools

US ATLAS, NP, Belle II, NSLS II…

- Pythonic Big Data tools being used increasingly at Data centers
  - JupyterLab allows users to access compute resources from within a web browser, instead of via traditional ssh command line interface (CLI)
- Federated ID Jupyter Hub at SDCC
  - Allows ATLAS users to use their CERN/FNAL/SLAC credentials as well as local credentials
- Our users can access storage and compute farm through this mechanism.
  - Leverage tools developed and maintained by a larger community outside of HEP

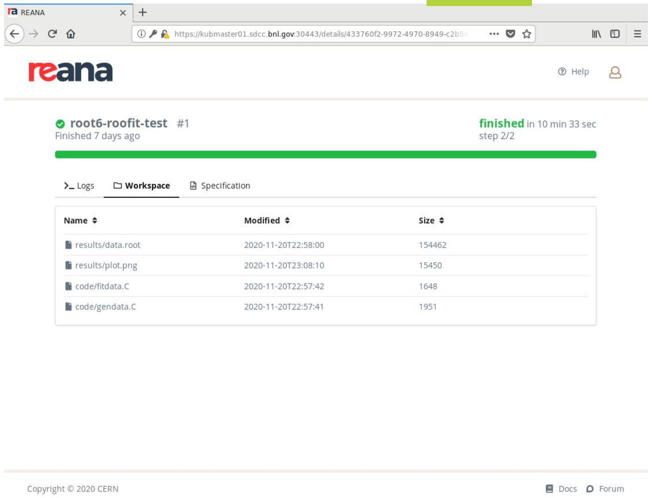




Brookhaven National Laboratory

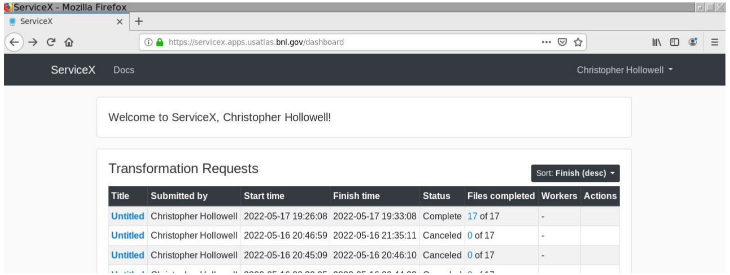
New ecosystem at SDCC

NP



- **REANA**
  - Work with CVMFS
  - Users can interface and submit container jobs to SLURM on the SDCC IC cluster
  - Successfully ported REANA to OKD - required numerous changes to REANA service containers and helm/pod YAML
- **ServiceX**
  - Deployed an ATLAS XAOD transformer instance in our production OKD cluster,
  - Modified helm/POD YAML and containers to function in OKD,
  - Successfully used from within our Jupyter deployment by users, including an IRIS-HEP developer
- **FuncX**
  - NSLS II evaluation (together with Airflow)

US ATLAS



Brookhaven National Laboratory

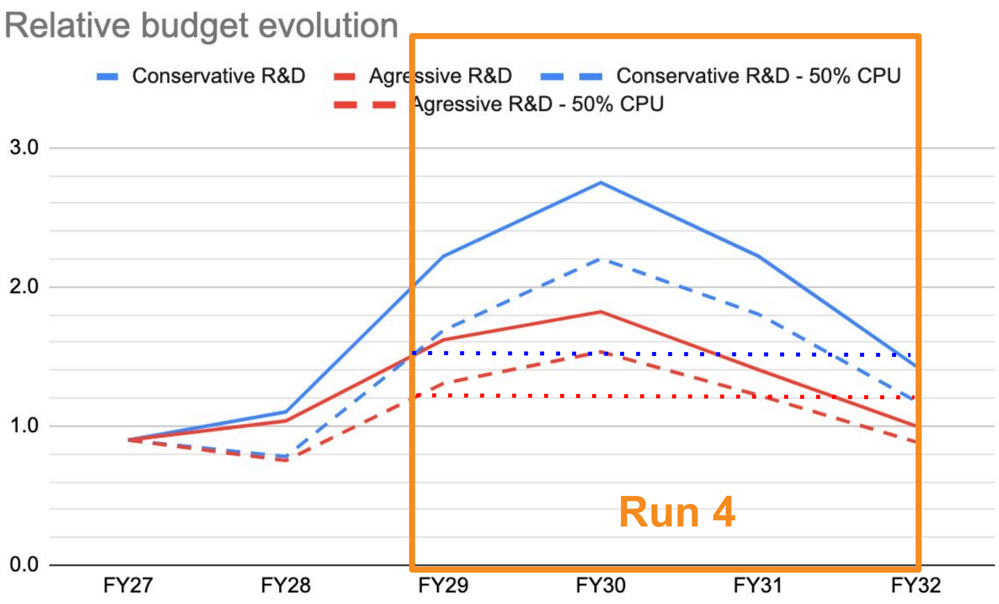
Summary

- R&D is required to address storage challenges in the HL-LHC era
  - Effort needed for R&D at facilities,
  - R&D must include various actors (storage experts, middleware, analysis design, ...),
  - A co-design concept is required for success.

- LHC is at risk of falling behind on the new user oriented software ecosystems (containers, python, …)
  - Dedicated LHC solutions may become less effective to maintain,
  - Migration to new ecosystem will be more costly as time goes on,
  - Prototyping, education should be strongly supported.

Brookhaven National Laboratory

Offloading 50% of CPU

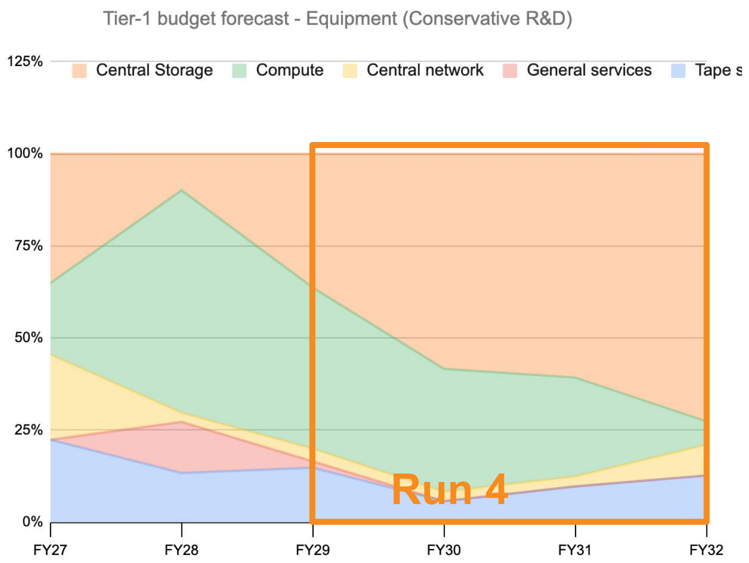


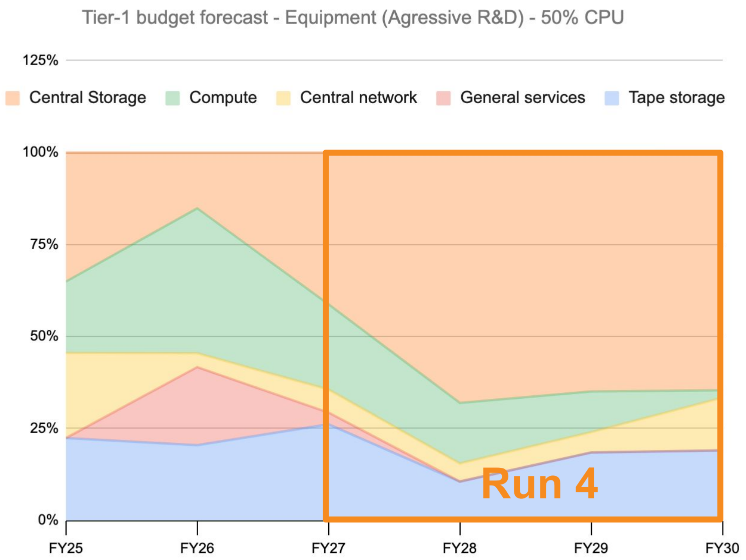
If software allows offloading 50% for CPU requirement to other facilities (like HPCs)

Conservative - 50% CPU : 1.5 x Flat
Aggressive - 50% CPU : 1.2 x Flat

E. Lancon et al., A Coordinated Ecosystem for HL-LHC Computing R&D (7-9 Nov 2022)

Budget decomposition - 2 extreme scenarios





In all scenarios disk storage is > 50% of the required equipment investment
Tape storage can be above 20% depending on performance requirements

E. Lancon et al., A Coordinated Ecosystem for HL-LHC Computing R&D (7-9 Nov 2022)