CHEP 2024
Distributed computing (Track 4) Summary

Daniela Bauer, Fabio Hernandez,
Panos Paparrigopoulos, Gianfranco Sciacca

October 19 - 25, 2024

CHEP
2024

Themes

- Tokens, Tokens, Tokens
    - Tokens being widely used and lessons are being learned
    - Current production setups
    - Development of best practice
    - Indigo-IAM development (also in response to “lessons learned”)
    - Adoption of tokens outside WLCG - early stages
- Operations
    - Grid computing adapts to changing circumstances
        - Operations: Optimizing use of available resources
        - Monitoring
        - New architectures and modern resources: ARM, HPCs and Clouds!
    - Security: It’s not just technology, people matter, too
- Distributed computing as part of non-WLCG computing models:
    - Gaining popularity especially in Astronomy: SKA, LSST, Einstein Telescope, CTA, HERD, but also DUNE (not astronomy)
        - Predicted SKA data volumes easily comparable to WLCG, building on WLCG experience for large scale operations!

30 Talks
19 Posters

October 19 - 25, 2024 CHEP 2024

Link to all track 4 talks: https://indico.cern.ch/event/1338689/sessions/553987/#all

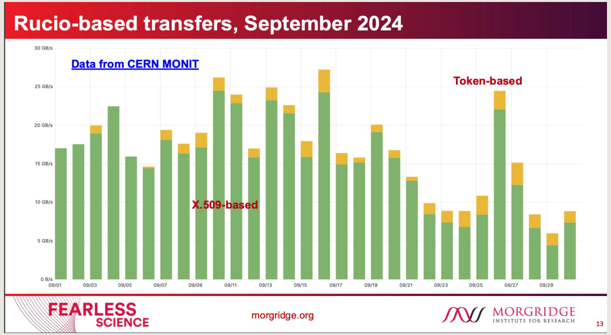
Tokens

- Tokens are now a reality! The infrastructure is almost token ready- time to focus on the operational models!
- DC24, a major milestone: millions of transfers with tokens! Lessons learned:
  - Token implementations of middleware need to improve: FTS/Rucio/Dirac workflows/IAM: all doing a lot of work:
    - Since August ATLAS has been running tokens to 15 sites: 1-2Hz with 5Hz spikes!
  - Performance must be stress-tested
- IAM is being improved:
  - Moving from OpenShift to K8S
  - Better token lifecycle management and storage
  - OIDC/OAuth 2.0: from MitreID Connect to Spring Authorization Server
  - Open Policy Agent (OPA) to speed up policy evaluation and a move to 2FA

October 19 - 25, 2024
CHEP 2024

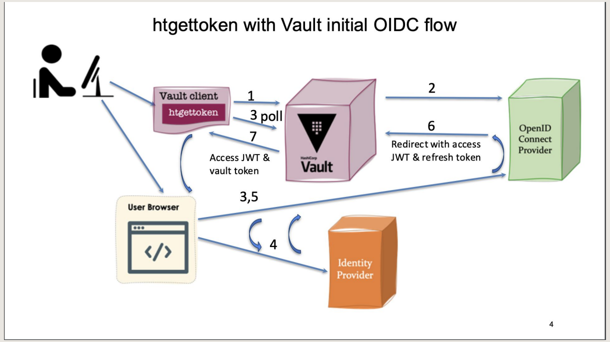
Tokens

- CMS has made strides in token usage:
  - Every CE but 3 use tokens for pilot submissions, analysis to follow
  - CMS is using tokens in production (via Rucio) since early September (in over 30 sites)
    - 1 token per dataset, IAM can handle just fine
  - By the end of 2025 all services should be able to handle both x509 and tokens!
- Fermilabs' Vault instance is used to manage tokens:
  - All FNAL hosted experiments (minus CMS) have been using tokens with all grid jobs for over a year now, with the CILogon OIDC Provider.
  - Vault is hiding the token complexity from users:
    - Vault is paid but there is a very promising open source version.
  - CMS will do the same, possibly via a CERN instance.


Rucio-based transfers, September 2024

Data from CERN MONIT

Brian Bockelman - CMS Token Transition

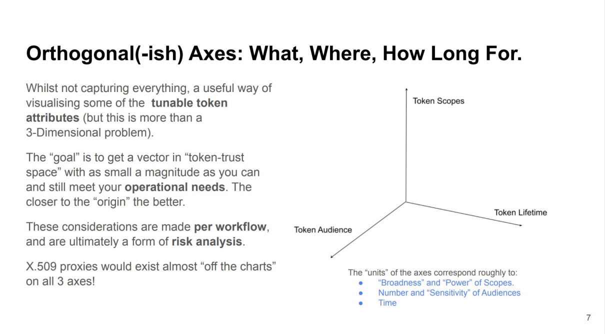

htgettoken with Vault initial OIDC flow

Nick Smith - Fermilab's Transition to Token Authentication

Tokens

- The balance between operability, security and performance needs to be found:
  - The Token Trust and Traceability WG Aims to form best practices, for users, devs, service providers + issuers.
  - Audience, lifetime, scopes are the three orthogonal parameters that one needs to tweak to meet the operational needs without compromising too much security.

- Next steps:
  - Use tokens on all grid jobs for stageout and reading
  - Users no longer need to issue x509/proxies
  - Accounting needs to be figured out



Orthogonal(-ish) Axes: What, Where, How Long For.

Whilst not capturing everything, a useful way of visualising some of the **tunable token attributes** (but this is more than a 3-Dimensional problem).

The “goal” is to get a vector in “token-trust space” with as small a magnitude as you can and still meet your **operational needs**. The closer to the “origin” the better.

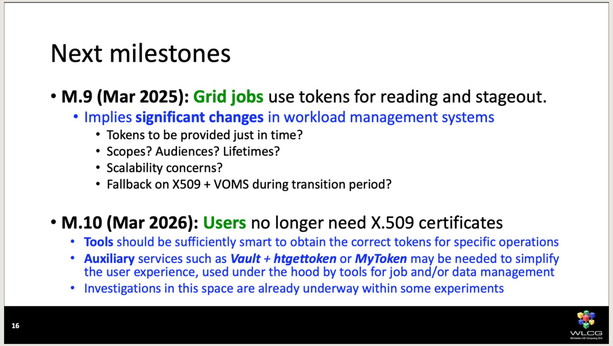
These considerations are made **per workflow**, and are ultimately a form of **risk analysis**.

X.509 proxies would exist almost “off the charts” on all 3 axes!

The “units” of the axes correspond roughly to:
- “Broadness” and “Power” of Scopes.
- Number and “Sensitivity” of Audiences
- Time

7

Matt Doidge - Early recommendations from the Token Trust and Traceability WG



Next milestones

- **M.9 (Mar 2025): Grid jobs** use tokens for reading and stageout.
  - Implies **significant changes** in workload management systems
    - Tokens to be provided just in time?
    - Scopes? Audiences? Lifetimes?
    - Scalability concerns?
    - Fallback on X509 + VOMS during transition period?
- **M.10 (Mar 2026): Users** no longer need X.509 certificates
  - **Tools** should be sufficiently smart to obtain the correct tokens for specific operations
  - **Auxiliary** services such as ***Vault + htgettoken*** or ***MyToken*** may be needed to simplify the user experience, used under the hood by tools for job and/or data management
  - Investigations in this space are already underway within some experiments

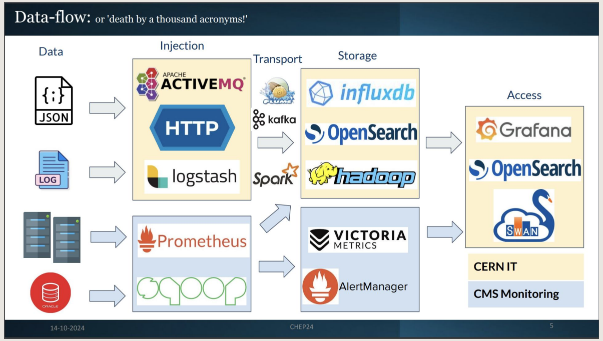
16

Tom Dack - WLCG transition from X.509 to Tokens: Progress and Outlook

October 19 - 25, 2024
CHEP
2024

Operations - Monitoring

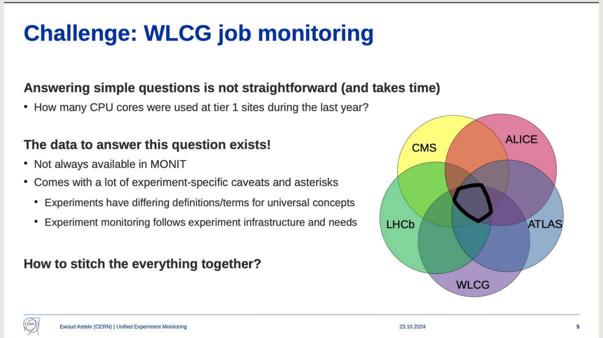
- WLCG central ops:
  - Size and complexity of infrastructure grows: new architectures, non-grid sites person-power doesn't.
  - Focus is on integrating those heterogeneous resources, while keeping the grid operating (while pushing for common tools and approaches!)
- ATLAS Hammercloud
  - Automatic exclusion/recovery of sites
- Monitoring:
  - Unified Experiment Monitoring
  - Experiment specific: CMS
    - Common tools and technologies to minimize maintenance and operations



Data-flow: or 'death by a thousand acronyms!'

14-10-2024 CHEP24 5

Brij Kishor Jashal - Advanced monitoring capabilities of the CMS Experiment for LHC Run3 and beyond



Challenge: WLCG job monitoring

Answering simple questions is not straightforward (and takes time)

- How many CPU cores were used at tier 1 sites during the last year?

The data to answer this question exists!

- Not always available in MONIT
- Comes with a lot of experiment-specific caveats and asterisks
  - Experiments have differing definitions/terms for universal concepts
  - Experiment monitoring follows experiment infrastructure and needs

How to stitch the everything together?

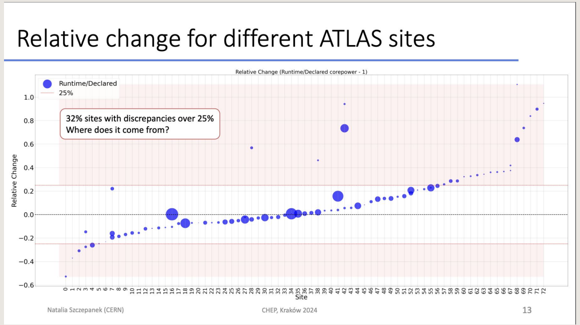
Ewoud Ketele (CERN) | Unified Experiment Monitoring 23.10.2024 5

Ewoud Ketele - Unified Experiment Monitoring

Operations - Optimisations

- Optimising the use of available resources:
  - ATLAS: HEP benchmark
    - Distinguishing fact/reality from fiction/ideal case
  - ATLAS: Results of the review of the ATLAS workflow management system (PanDa)
  - ALICE: Job Optimizers
    - Submit jobs faster and to the correct sites
  - ALICE: Whole node scheduling
    - Better exploit node resources with tuned payloads
  - ALICE: Unprivileged subdivision of job resources within the ALICE grid
  - CMS pilot overloading

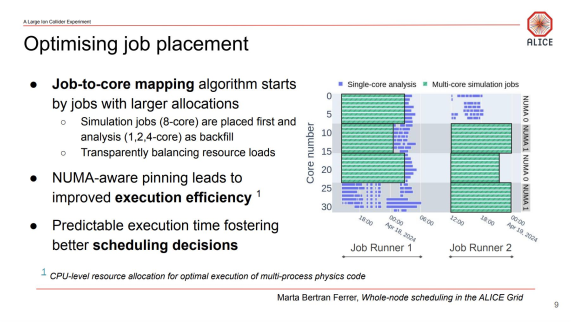
October 19 - 25, 2024
CHEP
2024



Relative change for different ATLAS sites

Natalia Szczepanek (CERN) CHEP, Kraków 2024 13

Natalia Szczepanek - Optimization of ATLAS computing resource usage through a modern HEP Benchmark Suite via HammerCloud and PanDA



A Large Ion Collider Experiment

Optimising job placement

- **Job-to-core mapping** algorithm starts by jobs with larger allocations
  - Simulation jobs (8-core) are placed first and analysis (1,2,4-core) as backfill
  - Transparently balancing resource loads
- NUMA-aware pinning leads to improved **execution efficiency** [1]
- Predictable execution time fostering better **scheduling decisions**

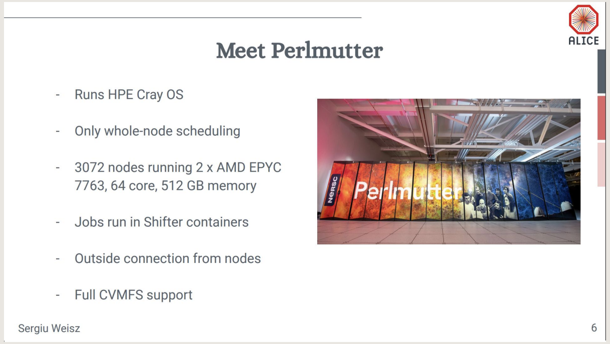
[1] *CPU-level resource allocation for optimal execution of multi-process physics code*

Marta Bertran Ferrer, *Whole-node scheduling in the ALICE Grid* 9

Marta Bertran Ferrer- Whole-node scheduling in the ALICE Grid: Initial experiences and evolution opportunities

Operations - Alternative architectures

- LHAASO trying to integrate Chengdu Supercomputing Centre
  - Dedicated link to avoid firewall issues and SLURM/HTCondor/XrootD to the rescue
- ALICE integrating the Perlmutter HPC
  - Integrated successfully getting the resources equivalent of a T2
- JAliEn (ALICE's Grid Framework) evolved to support ARM!
  - Also riscv64 architecture, as a proof of concept
- CVMFS performance upgrades:
  - New cache manager to open fewer files and improvements on parallel downloads!
- HEPCloud, after 6 years and a lot of problem solving is now a mature provisioning system which provides access to compute resources (HPC + clouds) similar to the size of the US CMS Tier-1 facility at Fermilab!
- SPECTRUM
  - EU funded project: focus on strategy, but also technical blueprint for data-intensive science and infrastructures



ALICE

Meet Perlmutter

- Runs HPE Cray OS
- Only whole-node scheduling
- 3072 nodes running 2 x AMD EPYC 7763, 64 core, 512 GB memory
- Jobs run in Shifter containers
- Outside connection from nodes
- Full CVMFS support

Sergiu Weisz 6

Sergiu Weisz - Integrating the Perlmutter HPC system in the ALICE Grid

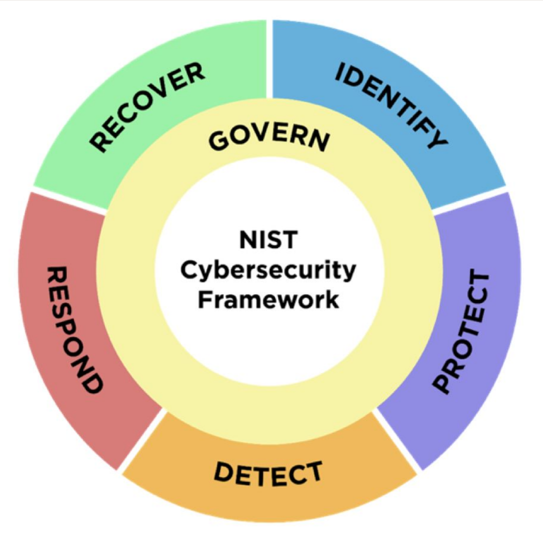
Security

- Threat from cyber attacks is persistent: strategy and a plan are needed

- People are the key: Communicate, collaborate, share

- Security operations centre (SOC) fits with an overall cybersecurity plan such as the Trusted CI Framework.
    - Be proactive to prevent cybersecurity incidents: monitor, detect, respond

- The pDNSSOC package was suggested as a lightweight way for smaller sites to get the benefits of a SOC; more volunteers/testers of this would be welcome.

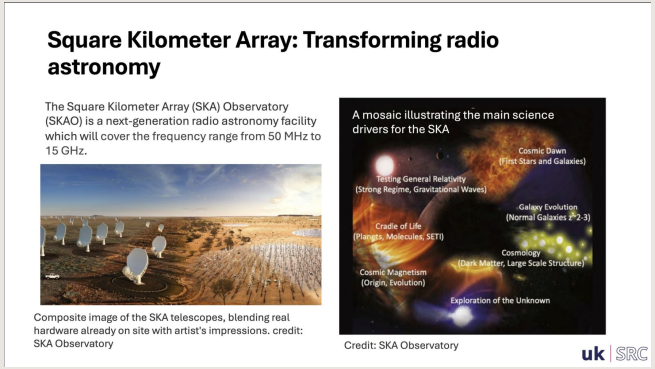


David Crooks - Designing Operational Security systems: People, Processes and Technology

October 19 - 25, 2024 CHEP 2024

Life outside the WLCG - 1

- SKA (all purpose radio telescope):
  - At full operations expects rates up to 400 PB/year by 2030 - easily comparable to LHC experiments
  - Construction is planned in stages and data from the very first stage is available
  - Computing organized in ~9 SRCNet resource centres, using common tools like IAM
- CTA (gamma ray astronomy):
  - Observations planned to start at 2030. Simulations are running since 2011
  - Production system centred around DIRAC (to move to DiracX)* with a CTA specific extension and soon Rucio
- Einstein Telescope (gravitational wave observatory):
  - (Data) challenge driven iterative development for computing model
  - Still multiple iterations expected



Square Kilometer Array: Transforming radio astronomy

The Square Kilometer Array (SKA) Observatory (SKAO) is a next-generation radio astronomy facility which will cover the frequency range from 50 MHz to 15 GHz.

Composite image of the SKA telescopes, blending real hardware already on site with artist's impressions. credit: SKA Observatory

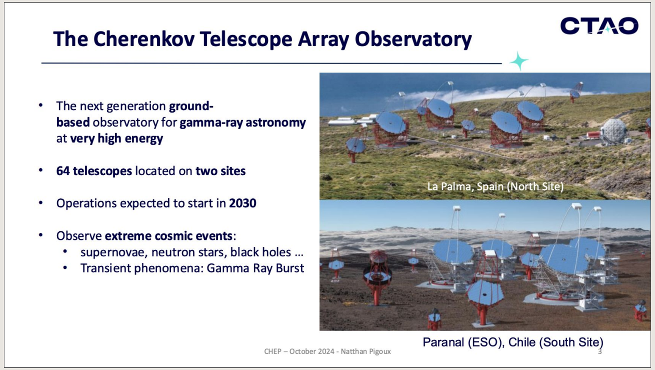
A mosaic illustrating the main science drivers for the SKA

Cosmic Dawn
(First Stars and Galaxies)
Testing General Relativity
(Strong Regime, Gravitational Waves)
Galaxy Evolution
(Normal Galaxies z~2-3)
Cradle of Life
(Planets, Molecules, SETI)
Cosmology
(Dark Matter, Large Scale Structure)
Cosmic Magnetism
(Origin, Evolution)
Exploration of the Unknown

Credit: SKA Observatory

uk | SRC

Ian Collier -The path to exabyte astronomy: SRCNet v0.1 for the Square Kilometre Array



The Cherenkov Telescope Array Observatory

CTAO

- The next generation **ground-based** observatory for **gamma-ray astronomy** at **very high energy**
- **64 telescopes** located on **two sites**
- Operations expected to start in **2030**
- Observe **extreme cosmic events**:
  - supernovae, neutron stars, black holes …
  - Transient phenomena: Gamma Ray Burst

La Palma, Spain (North Site)

Paranal (ESO), Chile (South Site)

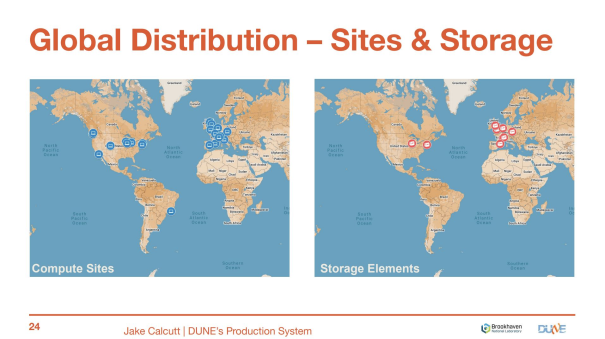
CHEP – October 2024 - Natthan Pigoux 3

Natthan Pigoux - The Cherenkov Telescope Array Observatory Production System Status and Development

* see plenary: https://indi.to/DHhVG

Life outside the WLCG - 2

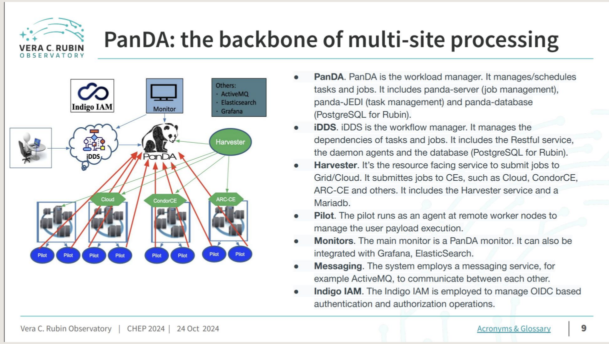
- DUNE (neutrinos)
  - Worldwide distribution of data and compute
  - FNAL legacy systems were used in the first prototypes: these have now been replaced
  - Uses HTCondor + GlideinWMS (from CMS) for job submission, Rucio for data management.
  - Participated in DC24, validating the new stack
- Vera C Rubin (sky survey)
  - Final phases of construction. Planned to start at 2025
  - Dealing with monitoring and log keeping / analysis challenges.
  - Uses PanDa (from ATLAS)/RUCIO/FTS/IAM
- HERD (High Energy cosmic-Radiation Detection facility - in space !)
  - > 90 PB, ~16000 CPU cores, in 10 years
  - Uses DIRAC/Rucio from WLCG
  - T0/T1/T2 distributed model, with T2 doing simulations only



Global Distribution – Sites & Storage

24 Jake Calcutt | DUNE's Production System

Jacob Michael Calcutt - Evolution of DUNE's Production System



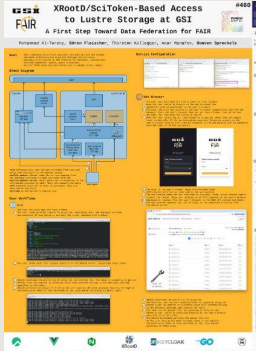
PanDA: the backbone of multi-site processing

- **PanDA**. PanDA is the workload manager. It manages/schedules tasks and jobs. It includes panda-server (job management), panda-JEDI (task management) and panda-database (PostgreSQL for Rubin).
- **iDDS**. iDDS is the workflow manager. It manages the dependencies of tasks and jobs. It includes the Restful service, the daemon agents and the database (PostgreSQL for Rubin).
- **Harvester**. It's the resource facing service to submit jobs to Grid/Cloud. It submittes jobs to CEs, such as Cloud, CondorCE, ARC-CE and others. It includes the Harvester service and a Mariadb.
- **Pilot**. The pilot runs as an agent at remote worker nodes to manage the user payload execution.
- **Monitors**. The main monitor is a PanDA monitor. It can also be integrated with Grafana, ElasticSearch.
- **Messaging**. The system employs a messaging service, for example ActiveMQ, to communicate between each other.
- **Indigo IAM**. The Indigo IAM is employed to manage OIDC based authentication and authorization operations.

Vera C. Rubin Observatory | CHEP 2024 | 24 Oct 2024 Acronyms & Glossary | 9

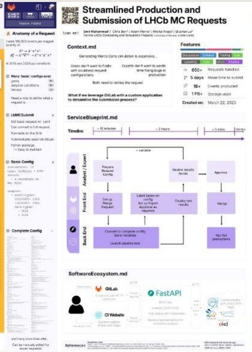
Fabio Hernandez - Preparation of Multi-Site Data Processing at the Vera C. Rubin Observatory

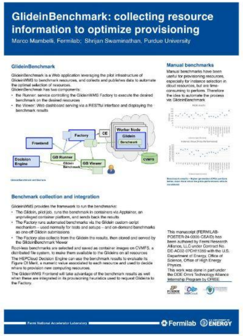
October 19 - 25, 2024 CHEP 2024


XRootD/SciToken-Based Access to Lustre Storage at GSI

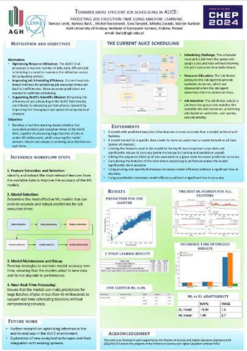
A First Step Toward Data Federation for FAIR

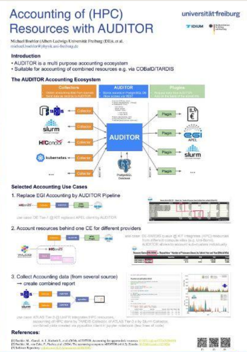
Mohammad Al-Turany, Sören Fleischer, Thorsten Kollegger, Anar Manafov, Rouven Spreckels


Streamlined Production and Submission of LHCb MC Requests

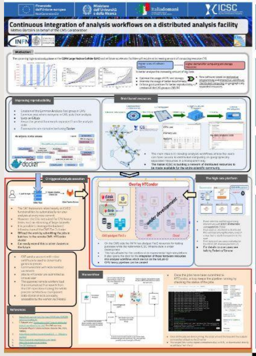

GlideinBenchmark: collecting resource information to optimize provisioning

Marco Mambelli, Fermilab; Shrijan Swaminathan, Purdue University

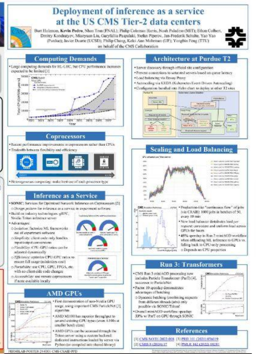

Towards more efficient job scheduling in ALICE: Predicting job execution time using machine learning

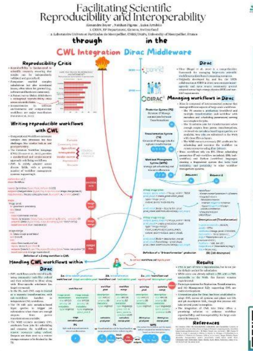

Accounting of (HPC) Resources with AUDITOR

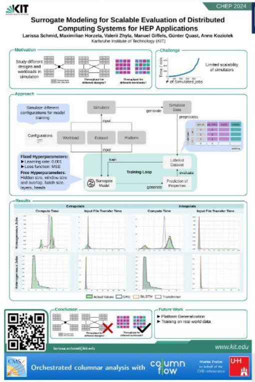
Michael Boehler (Albert-Ludwigs Universität Freiburg (DE)), et al.


Continuous integration of analysis workflows on a distributed analysis facility

Matteo Bartolini on behalf of the CMS Collaboration

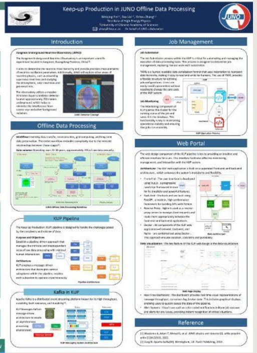

Deployment of inference as a service at the US CMS Tier-2 data centers

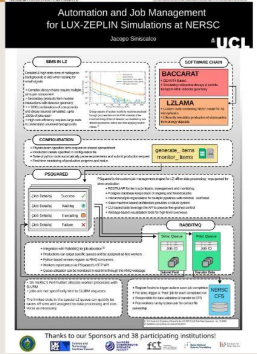

Facilitating Scientific Reproducibility and Interoperability through CWL Integration in the Dirac Middleware


Surrogate Modeling for Scalable Evaluation of Distributed Computing Systems for HEP Applications

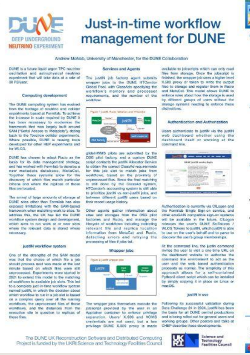
Larissa Schmid, Maximilian Horzela, Valerii Zhyla, Manuel Giffels, Günter Quast, Anne Koziolek

Karlsruhe Institute of Technology (KIT)

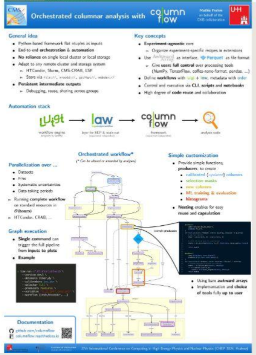

Keep-up Production in JUNO Offline Data Processing


Automation and Job Management for LUX-ZEPLIN Simulations at NERSC

Jacopo Siniscalco


Just-in-time workflow management for DUNE

Andrew McNab, University of Manchester, for the DUNE Collaboration


Orchestrated columnar analysis with columnflow

October 19 - 25, 2024

CHEP 2024

Fri 25th Oct - Track 4 summary: Distributed Computing - 12

Thank you!

- Thanks to all the Track 4 speakers and poster presenters!
- To the organization: the program committee and the team of track convenors for their outstanding work in making this a success!
- To all of you, making this conference a wonderful experience!
- Looking forward to seeing you all again soon!