EGI Workload Manager environmental impact optimization

A.Tsaregorodtsev,
CPPM-IN2P3-CNRS, Marseille,
GreenDIGIT Technology Design Workshop,
23 January 2025

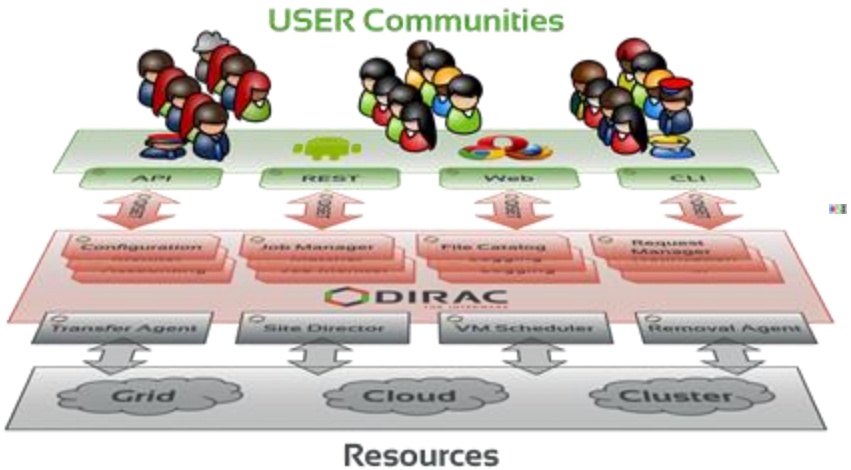
DIRAC
THE INTERWARE

Outline

- DIRAC quick overview
- WMS architecture
- Main WMS components and optimization possibilities
- Conclusion

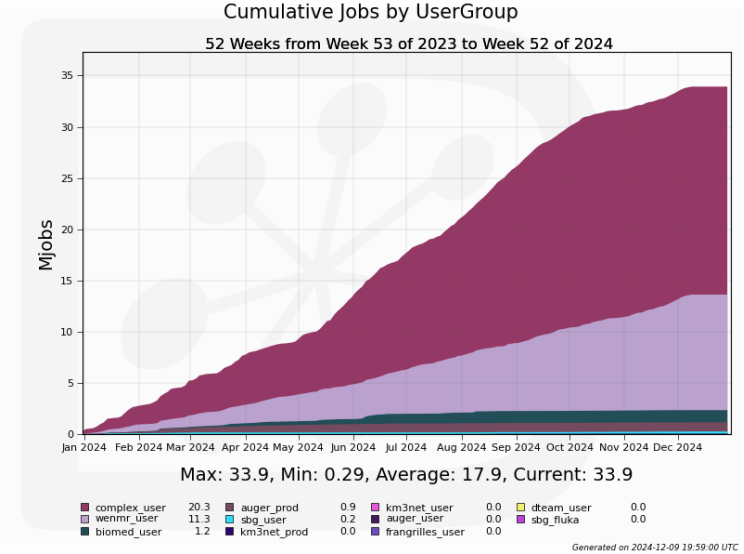
The interware

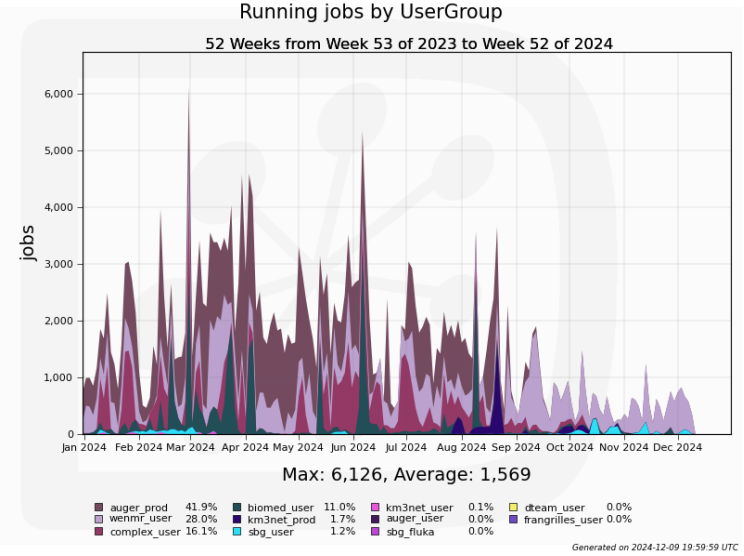
- A software framework for distributed computing
- A **complete** solution to one (or more) user community
- Builds a layer between users and resources
- A *framework* shared by multiple experiments, both inside HEP, astronomy, and life sciences



EGI Workload Manager

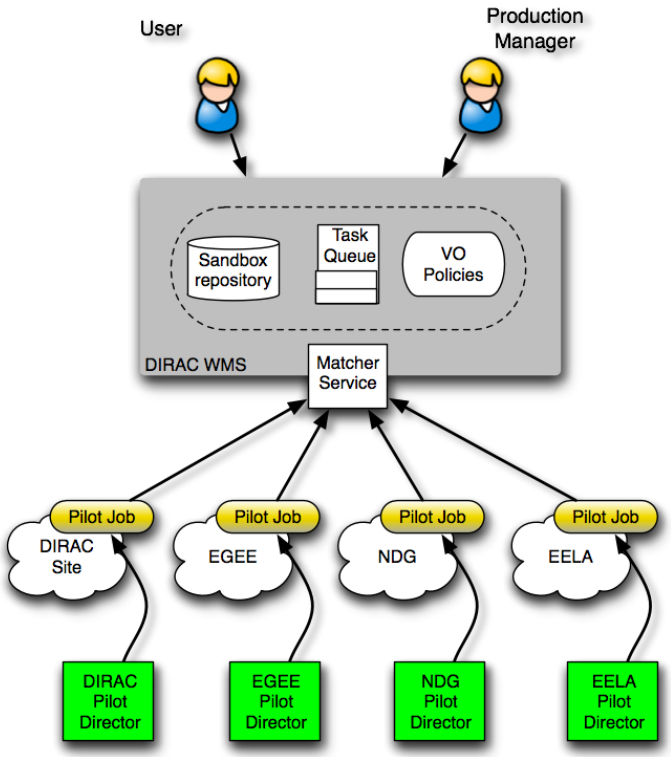
- Hosted at CC/IN2P3 in Lyon
  - Maintained by a joint team from several CNRS laboratories
  - Cluster of 10+3 medium sized virtual servers + MySQL/ElasticSearch host service
- ~25 VOs
- > 200 registered users
  - Some of them represent large communities
- Running smoothly
  - ~35M jobs in 2024
    - Up to 6K concurrent jobs
      - ~1600 on average last month
    - Up to 200K jobs per day in spikes





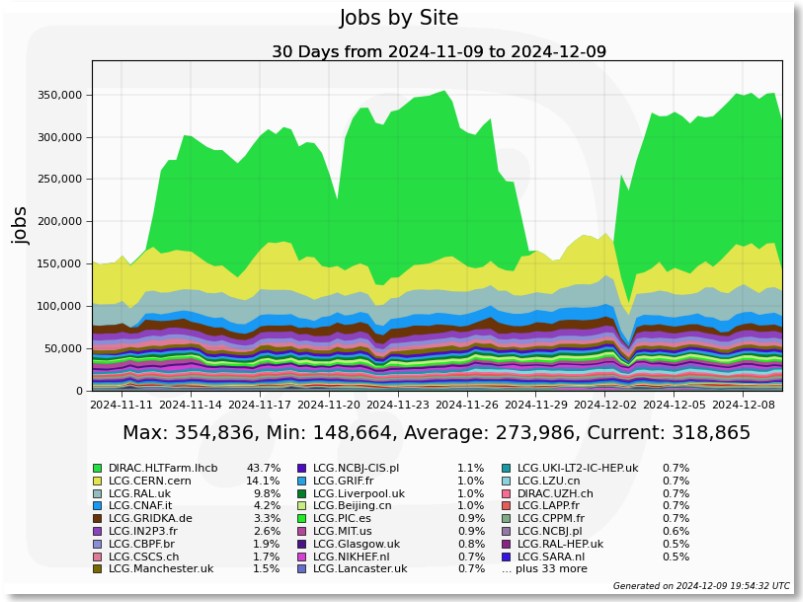
WMS Architecture

- DIRAC WMS architecture is based on the ***PULL*** paradigm and the concept of pilot jobs
- User jobs are submitted to the central Task queue
  - Global view of user tasks allows to apply efficiently scheduling policies
- Pilots are submitted to various computing resources :
  - With specific submission mechanism in each case
- Pilots match (pull) the user payloads and steer their execution
  - Late binding of user payloads to the computing resources
  - Ensure precision of policies application
  - Reduce payload failure rates



DIRAC in LHCb

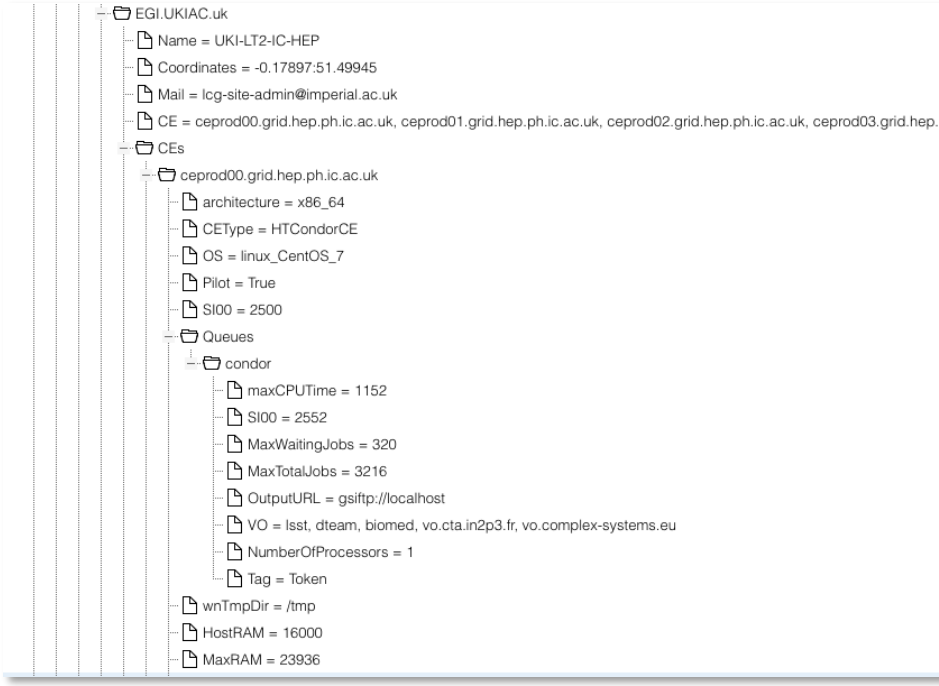
- DIRAC was originally developed for the LHCb experiment at LHC, CERN
  - The largest DIRAC user community
  - Initially focusing on HTC/grid computing resources for processing large data volumes (O(10PBs))
  - Later Cloud, HPC and special clusters were included
- DIRAC now went Open Source
  - GPL v3 license, open code in Github https://github.com/DIRACGrid
- Ongoing study in LHCb on the environmental sustainability
  - Similar goals as for the GreenDIGIT
    - WMS optimizations in the DiracX next generation software
  - We will share the developments



More than 300K concurrent jobs

Resources description

- Each site, CE, queue is described in DIRAC in details necessary for job scheduling
  - OS, CPU, memory limits, etc
  - Quantified values or tags
  - Extra parameters will be added as necessary
    - Site « eco-metrics », Green Rank based on WP3/WP4 outcome
- Sources of information
  - BDII, GocDB, GreenDIGIT
  - in place measurements, private communications, etc
- Static resources description now
  - Submitting pilots to predefined queues
- Possible improvements
  - more dynamic queue definitions reflecting the needs of particular job groups
  - Time dependent site ranking (dayly/weekly cycles)



Job submission

- Users submit jobs with precise description
  - Applications
  - Input/output
  - Resources requirements
- User payloads must be characterized to apply optimization algorithms
  - CPU or I/O bound
  - Application efficiency
  - Urgency (turnaround), priority
- The job characterization should be done in a real execution enviroment
  - Groups of jobs with the same characteristics/requirements
  - Trial jobs measurements applied to all the jobs in the group

Pilot submission

- Algorithms for pilot submission to sites
  - Now:
    - Random site choice – giving all the sites a chance
    - Goal: maximizing the user job turnaround
    - relaxed resources requirements: pilots can run different user jobs
  - Optimized:
    - Prioritizing sites to send pilots based on their environmental ranking
    - More specific pilot resources requirements limiting to particular job groups with same properties
    - Goal: reasonable compromise between turnaround and environmental impact

Job matching

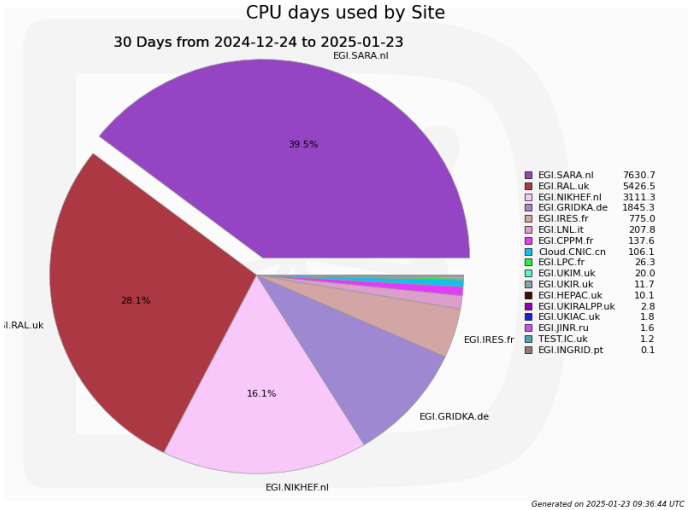
- Pilots match user jobs based on requirements and resource capacity
  - Jobs are picked up with a probability proportional to their priorities
    - Based on policies of VOs
  - Job priorities can be modified based on the site parameters at the matching time
    - Limited flexibility at matching time if pilots are submited with very specific requirements (for specific jobs)
- Pilots can match and execute several jobs in parallel
  - Resource reservation by whole nodes rather than by single-core slots
  - Running a small batch system on a worker node
  - Possibility to optimize a set of jobs in a given slot
    - E.g. combination of CPU and I/O bound jobs
- Overall pilot efficiency contributes well to the environmental sustainability in general !

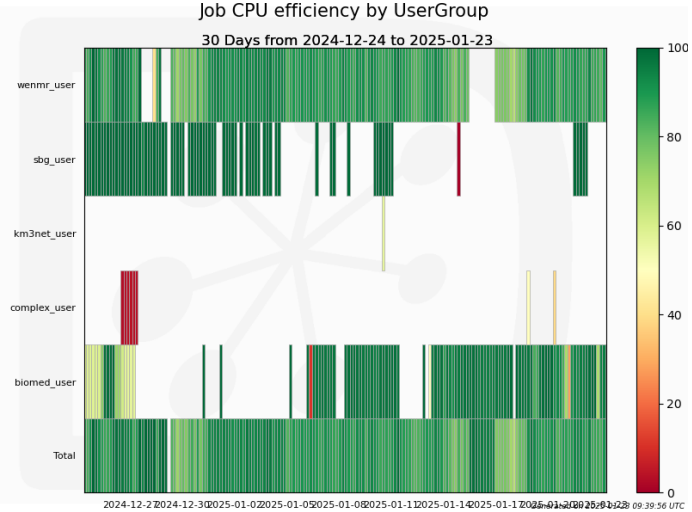
Job Wrapper

- Pilots create job wrappers to steer user jobs execution in the worker nodes
- Job wrapper – a watchdog process running in parallel with the user application
  - Initializes user application
    - Downloading input data, downloading software (CVMFS cache)
      I/O bound phase
  - Watches resources consumption by application
    - CPU, Disk, Memory, I/O operations
    - CPU bound phase
  - Sends job heartbeats and receives WMS commands
    - E.g. killing stalled jobs
  - Uploading results
    - Output data
    - Accumulated job parameters (resources usage, efficiency, etc)
    - I/O bound phase
- Job parameters as measured by the job wrapper are available to users
  - Allow for better description of subsequent jobs
- Job wrapper is part of the WMS
  - Can interact with the execution environment, e.g. CPU throttling while in I/O phase

Monitoring/Accounting

- DIRAC has extensive accounting of consumed resources
  - Integral and time-dependent fine granularity data
    - Per site, per community, per activity, per user, etc
    - CPU, jobs, data operations, etc
  - MySQL and ElasticSearch databases
- Accounting is extremely important to measure the effect of the WMS optimizations
- Extra accounted quantities will be added – to be defined, e.g.
  - Pilot efficiency
  - CPU « eco-normalized »
  - Energy usage, CO2 emissions if the corresponding measurements will be
  - available





Status

- EGI Workload Manager is in production since 2014

- The DIRAC software is now being rewritten to comply with modern standards and practicies
  - DiracX (neXt generation DIRAC) to be first released in Q1 2025
  - The base WMS architecture is retained and all of the above still holds
  - Continuity of the DIRAC -> DiracX transition will be ensured
  - GreenDIGIT related developments can go in parallel
  -
- Updated scheduling algorithms will be tested in the certification setup and applied in production
  - For selected communities, activities first to minimize interference

DIRAC
THE INTERWARE

Summary

- DIRAC WMS manages user payloads running on distributed heterogeneous computing resources at a large scale
- The WMS architecture based on the pilot jobs paradigm offers multple tools for collecting job and resource parameters, orchestration of the job execution, measurement of the overall system performance
- Enviromental requirements can be plugged into the scheduling process on various stages to minimize the environmental impact