# Science Gateway on GARUDA GRID for Open Source Drug Discovery community

**Presented by**

**Santhosh J**

**Authored by**

**Karuna Prasad, Mangala N, Janaki Ch**

***Centre for Development of Advanced Computing (C-DAC)***

Bangalore, India

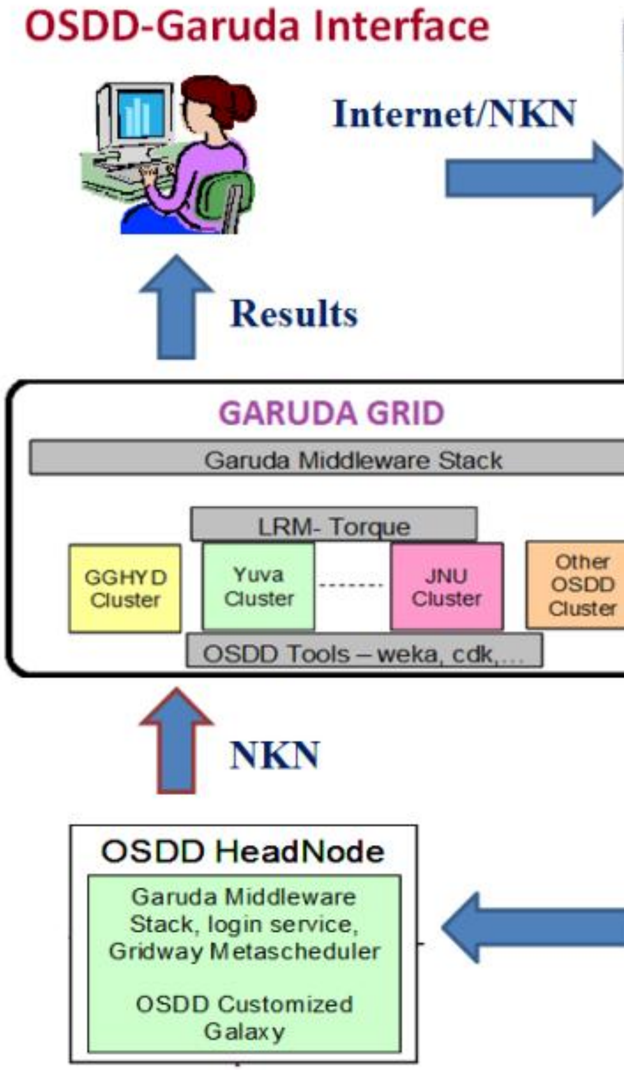
# Outline

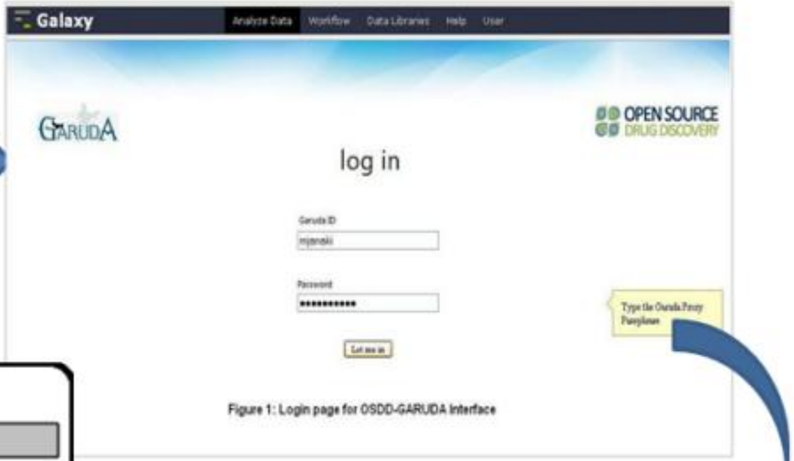
- Motivation
- Science Gateway for OSDD
- Garuda Grid
- OSDD-GARUDA Collaboration
- Galaxy-Garuda Architecture
- Gridway Job Runner
- Results and Achievements

# Motivation

- A pipeline of computational chemistry methods was used to discover drugs for malaria and thalassemia, by the CSIR Open Source  Drug Discovery initiative
- This involved several scientist working on different phases of the pipeline and where each task was computation and data intensive.
- To solve the problem, the GARUDA grid was enabled with specia science gateway to enable collaboration between the scientists and provide a seamless pipeline for computational discoveries.
- This paper describes the components of the system used – i)large compute resource of Garuda Grid, ii) secure remote access to the scientists to collaborate for problem solving, iii) provision of suitable workflow on Garuda.

# Science Gateway for OSDD

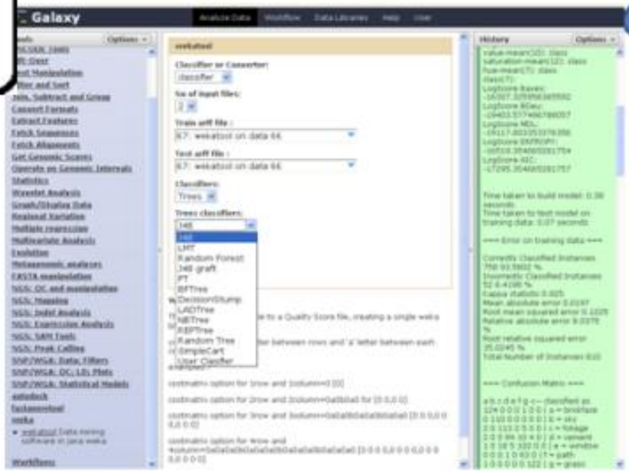




Garuda ID: mjanaki

Password: ●●●●●●●●●●●

Figure 1: Login page for OSDD-GARUDA Interface



Classifier or Converter: classifier

No of input files: 2

Train arff file : 67: wekatool on data 64

Test arff file : 67: wekatool on data 64

Classifiers: Trees

Trees classifiers: J48

# GARUDA-OSDD user community

- User wants a simple access for all the research and experimental activities
- Results of their experiments can be shared for analysis
- Domain expert users can’t understand all these middleware layers
- Interface which can enable the complex computational analysis for experimental biologists

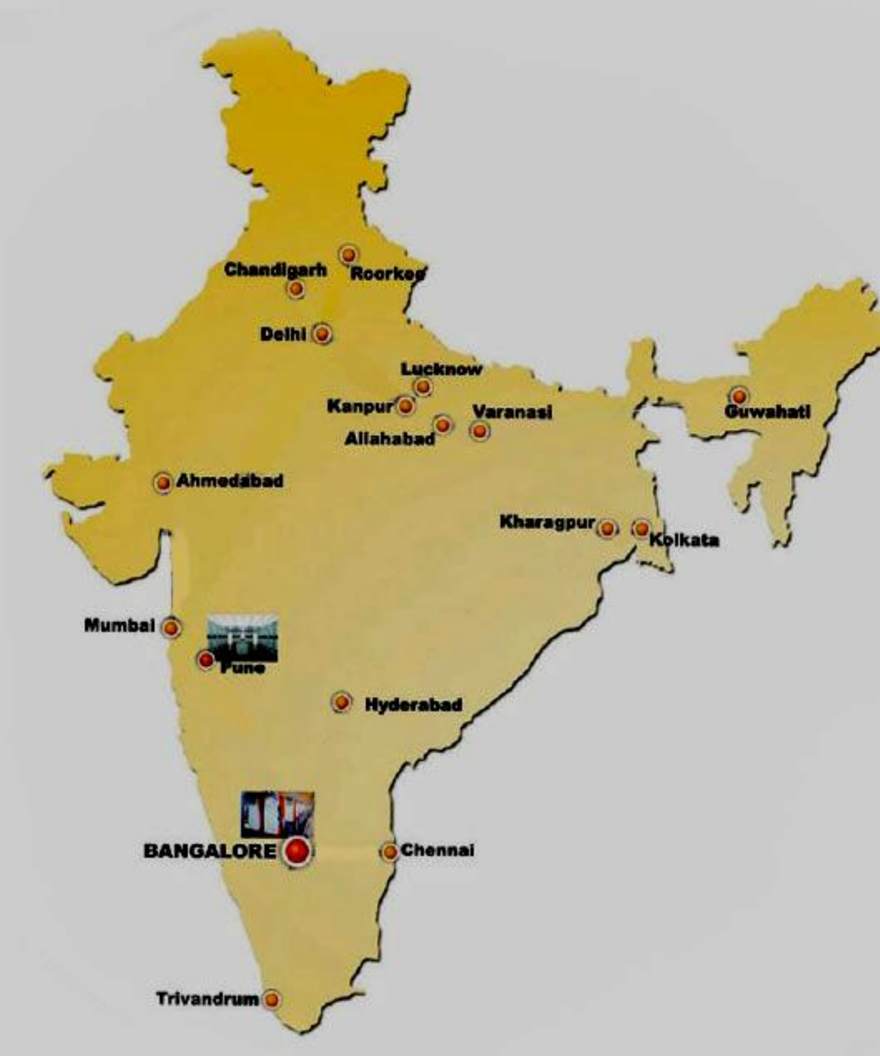
# GAURDA Grid

## GARUDA - Global Access to Resources Using Distributed Architecture

- **Resources** : GARUDA is heterogeneous resource distributed across India. These resource are aggregated from C-DAC and GARUDA partners like IISc, PRL, IITG, IITD and others. Total computational power is nearly 6000 cpus (~ 70TF of compute power) and about 17TB of storage has been aggregated on Garuda

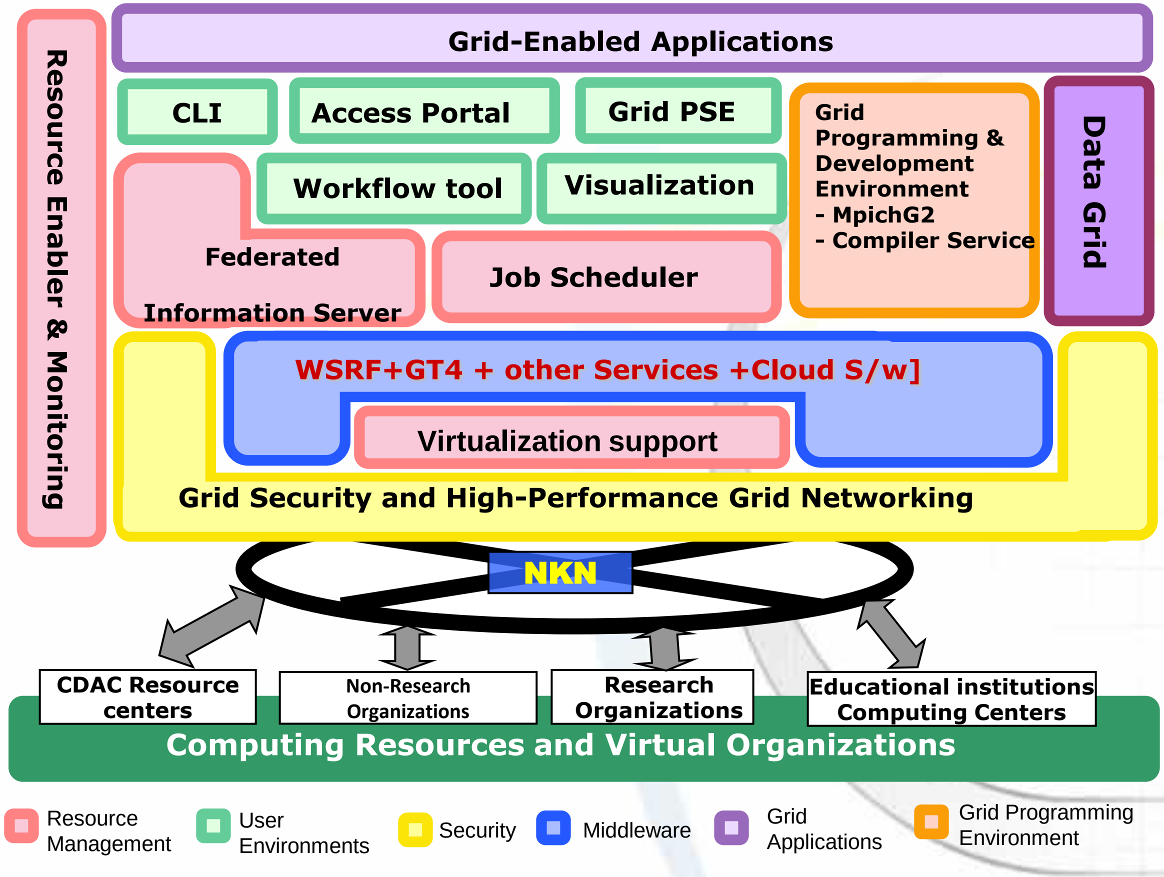
- **Network** : The National Knowledge Network (NKN) backbone, a Pan-Indian communication fabric to provide seamless and high-speed access to resources. NKN is an initiative by the Ministry of Information Technology, Government of India, to provide ultra high speed connectivity across the entire country. Academic institutes and R&D organizations can leverage this network for their applications. NKN currently supports 1Gbps and shall scale upto 10Gbps.

- GARUDA Grid middleware stack, tools and services which provide an integrated infrastructure to applications and higher-level layers



GARUDA Project is funded by Ministry of Communication and Information Technology (MCIT), Govt of India.

# High level GARUDA Architecture

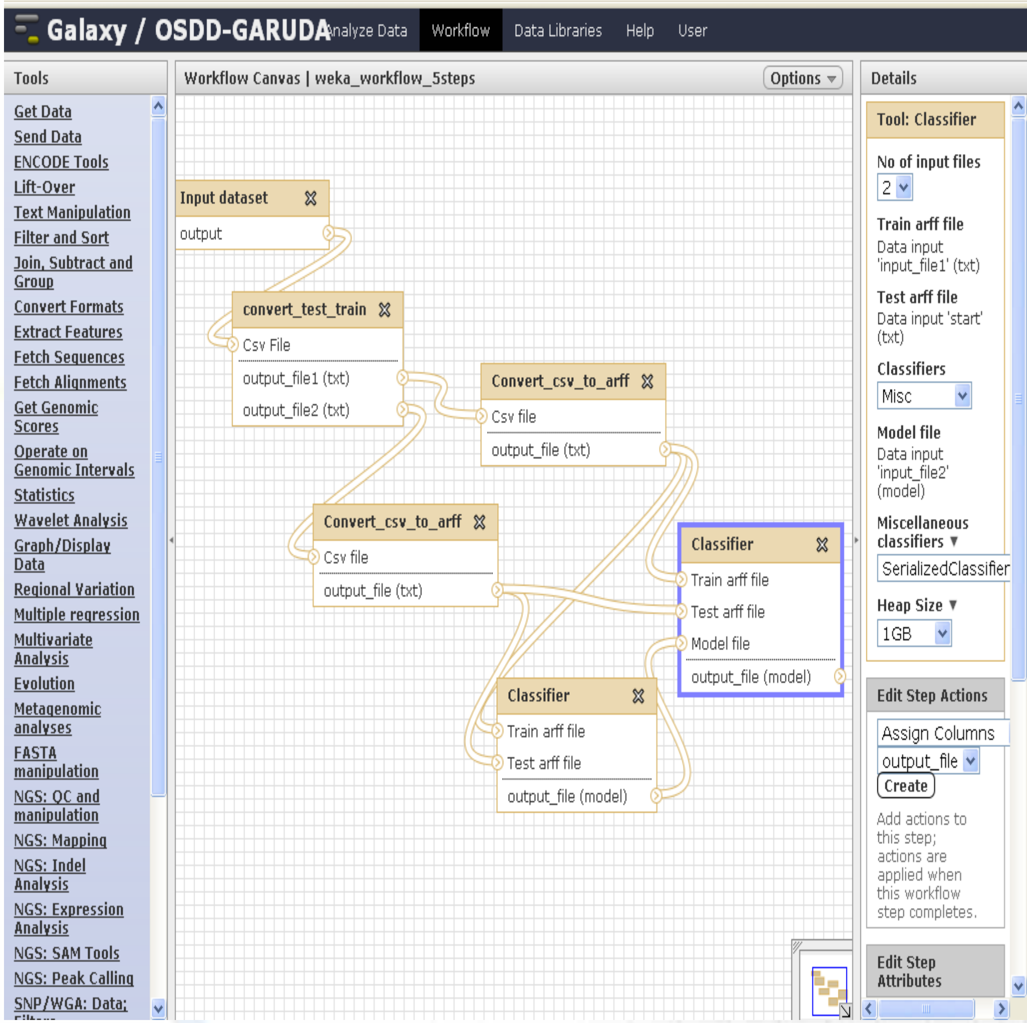

Grid-Enabled Applications
Resource Enabler & Monitoring
CLI
Access Portal
Grid PSE
Grid Programming & Development Environment
- MpichG2
- Compiler Service
Data Grid
Workflow tool
Visualization
Federated Information Server
Job Scheduler
WSRF+GT4 + other Services +Cloud S/w]
Virtualization support
Grid Security and High-Performance Grid Networking
NKN
CDAC Resource centers
Non-Research Organizations
Research Organizations
Educational institutions Computing Centers
Computing Resources and Virtual Organizations
Resource Management
User Environments
Security
Middleware
Grid Applications
Grid Programming Environment

# Galaxy workflow

- Galaxy is a popular workflow in the bioinformatics community due to ease of use, sharing results and workflows and persisting analysis makes it more valuable for research in the community.
- Galaxy can be run on clusters supporting SGE , PBS as local resource manager.
- Many popular tools like weka, gromacs, Namd etc can exploit the grid resources efficiently through the workflow.

# Galaxy Workflow

- Simplified GUI design.
- Ease of integrating modules.
- Fewer components for creating workflows.
- Sharable workflows for better collaboration



Tool: Classifier

No of input files: 2

Train arff file: Data input 'input_file1' (txt)

Test arff file: Data input 'start' (txt)

Classifiers: Misc

Model file: Data input 'input_file2' (model)

Miscellaneous classifiers: SerializedClassifier

Heap Size: 1GB

Assign Columns: output_file

Add actions to this step; actions are applied when this workflow step completes.

# Science Gateway

- Science Gateways provide a mechanism to user for accessing distributed shared compute resources for domain-specific applications
- It also provides an interface for visualizing simulated output through a collaborative visualization gateway.
- Specific community get benefitted science gateway as it comes with integrated, web-based data and knowledge management, secure data access, simulation capability, and analysis/visualization capabilities
- In order to synchronize efforts by various members of the group, it is important to provide a common platform like science gateway that facilitates data exchange and interaction among community members.

# OSDD- GARUDA Collaboration

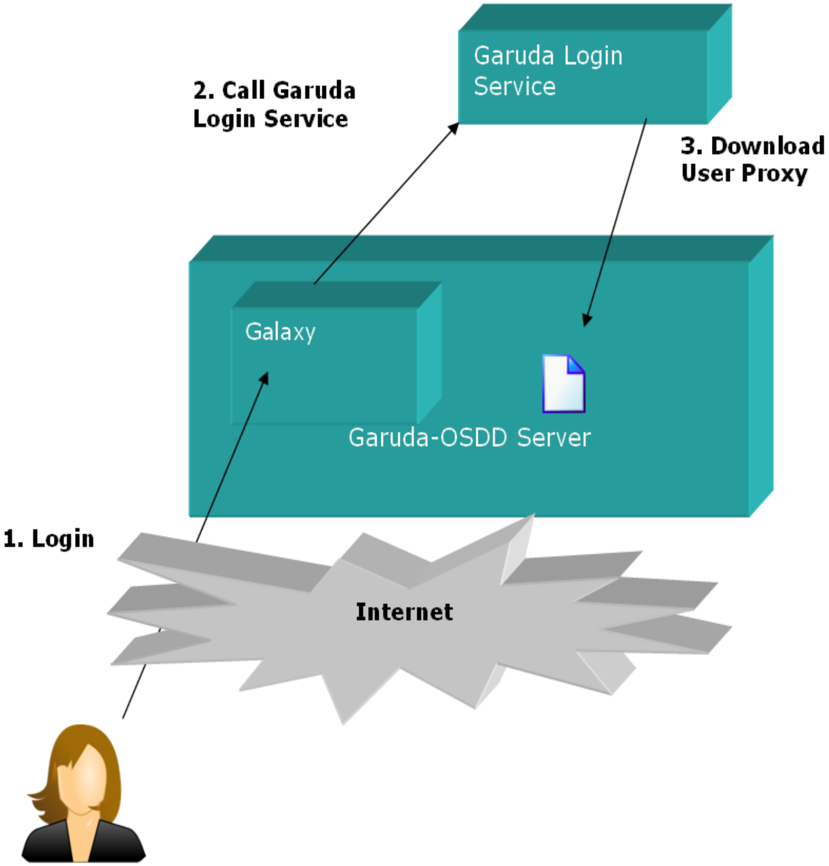
- GARUDA grid provides an unprecedented e-Infrastructure for OSDD applications.
- It provided access to the HPC clusters provided to run drug discovery problems through the NKN connectivity to OSDD centers.
- Secure access was enabled to high-end resources for scientists and students even from remote locations.
- Open source Science Gateway is enabled for genomics and proteomics applications.

# Trust and Security for Science Gateway

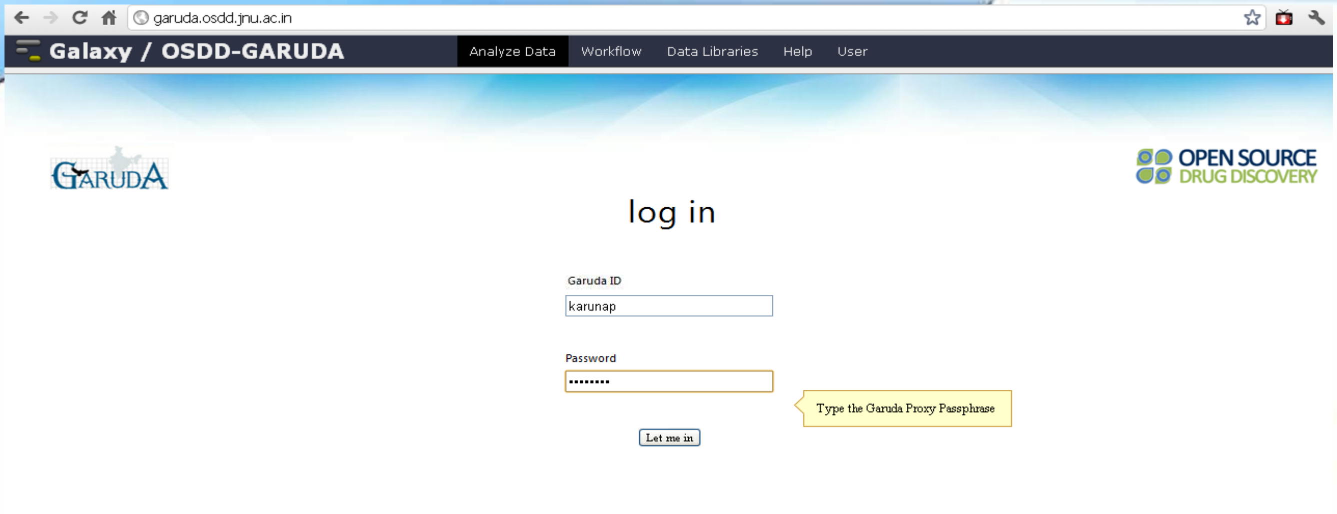
- **Digital certificates**: an electronic document issued by a trusted party or a certificate authority that binds the physical identity of an entity that is user or a machine (hardware) to their public key. This identity that is the digital certificate is then used to authenticate the parties involved in the transaction.
- **Proxy certificate**: These are the short-lived certificates that can be issued locally where the user is known but can have a global scope. They contain information about the roles and privileges of the user.
- **Indian Grid Certification Authority (IGCA):** IGCA is a Certification Authority that issues certificates to bind the physical identity of the entity(user, application or host) to the public key.
- **Registration Authority:** The IGCA delegates the authentication of individual identity to Registration Authorities. RA authenticates the identities of entities and requests the IGCA to issue a certificate for that entity. RA’s must sign an agreement with the IGCA, stating their adherence to the procedures. RA’s act as a user interface of IGCA to verify the end entities identity. RA must meet the end user face to face.

# Science Gateway Login Flow

- Users registers with IGCA, face-to-face meeting with RA
- Every user and a service on Garuda grid is identified by a certificate, which contains information vital to identifying and authenticating the user or service.
- The user can thus use that certificate to establish his/her identity and login to the web-based scientific workflow and access the remote computational clusters over internet.



- Web based Garuda – OSDD Science Gateway uses digital certificates to validate user's identity and grant them access.
- Each user of the grid needs to be registered in the specific Virtual Organization, which is role based access.
- Public key is used for user authentication and the proxy certificate is used for single sign-on and rights delegation.
- The use of proxy certificate limits the exposure of long-term credentials
- During job execution, to access various other services like data services, libraries etc separate authentication is not required.
- The proxy certificates will have the right to do authentication for the period of job execution time.



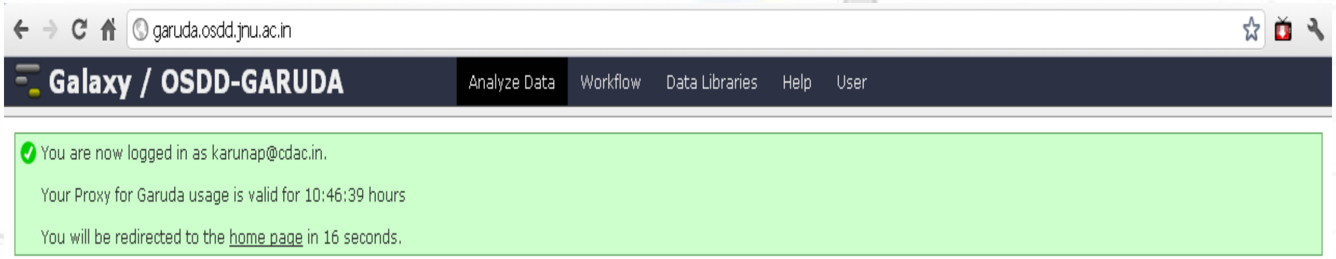
log in

Garuda ID: karunap

Password: ••••••••

Type the Garuda Proxy Passphrase

## Login page of Customized Galaxy Interface



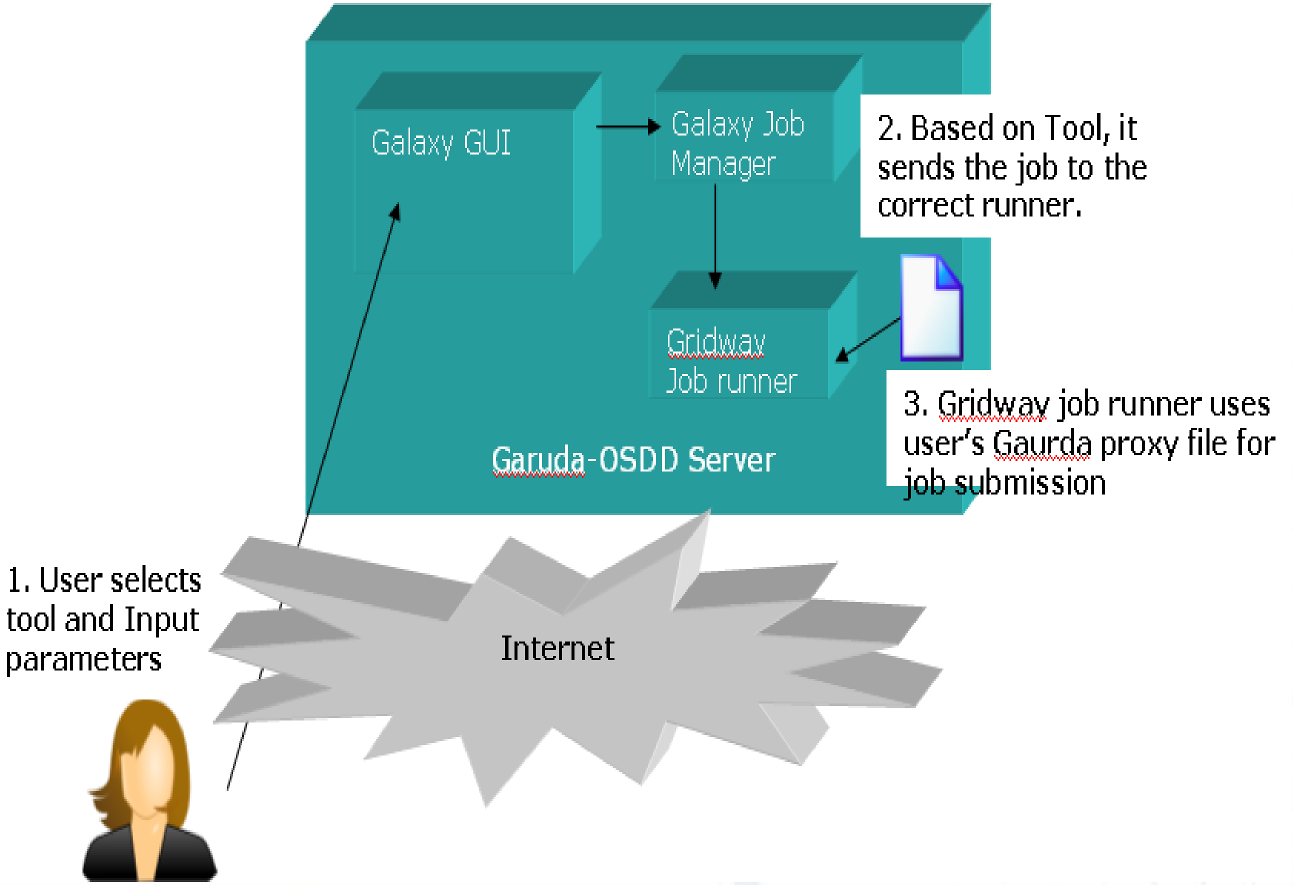
✓ You are now logged in as karunap@cdac.in.

Your Proxy for Garuda usage is valid for 10:46:39 hours

You will be redirected to the home page in 16 seconds.

## Page showing proxy validity

# Garuda-Galaxy Job –submission Flow



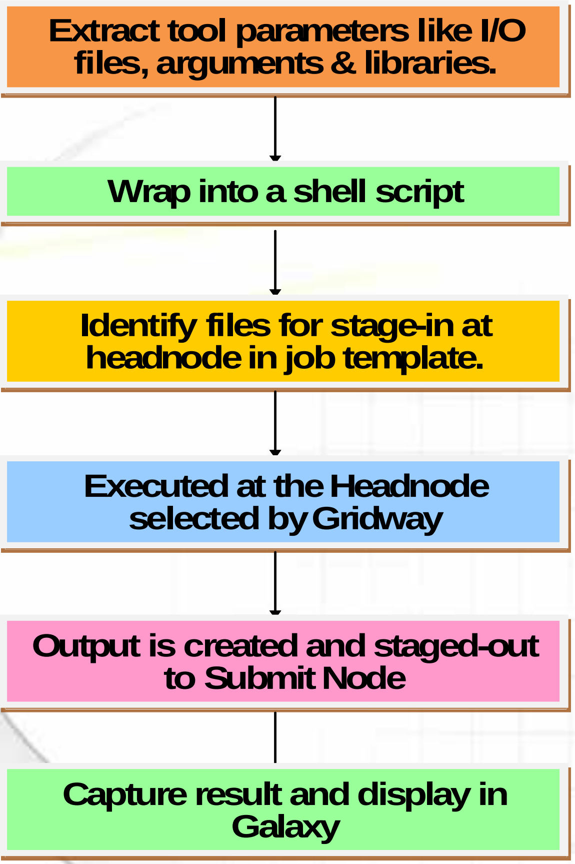
2. Based on Tool, it sends the job to the correct runner.

3. Gridway job runner uses user's Gaurda proxy file for job submission

1. User selects tool and Input parameters

# Gridway Job Runner

- Extracting the tools parameters
- Wrap in shell script
- Identifies files to be staged in at headnode and describe in the job template file
- The job template file will define all the job specific parameters
- Executed at the headenode scheduled by the gridway
- Output files staged out at the submit node
- Capture the result and display it in galaxy frontend.


Extract tool parameters like I/O files, arguments & libraries.

Wrap into a shell script

Identify files for stage-in at headnode in job template.

Executed at the Headnode selected by Gridway

Output is created and staged-out to Submit Node

Capture result and display in Galaxy

# Gridway Job Runner

- The gridway runner will be managing the execution of jobs submitted to the grid.
- Preparing the jobs for submission and creating a job wrapper
- Putting it in a Gridway queue to be submitted
- Monitoring the Job Id – watches the jobs currently in the queue and deals with the state change (queued to running and job completion), and
- Finishing the job
- Delete and recovery of jobs.

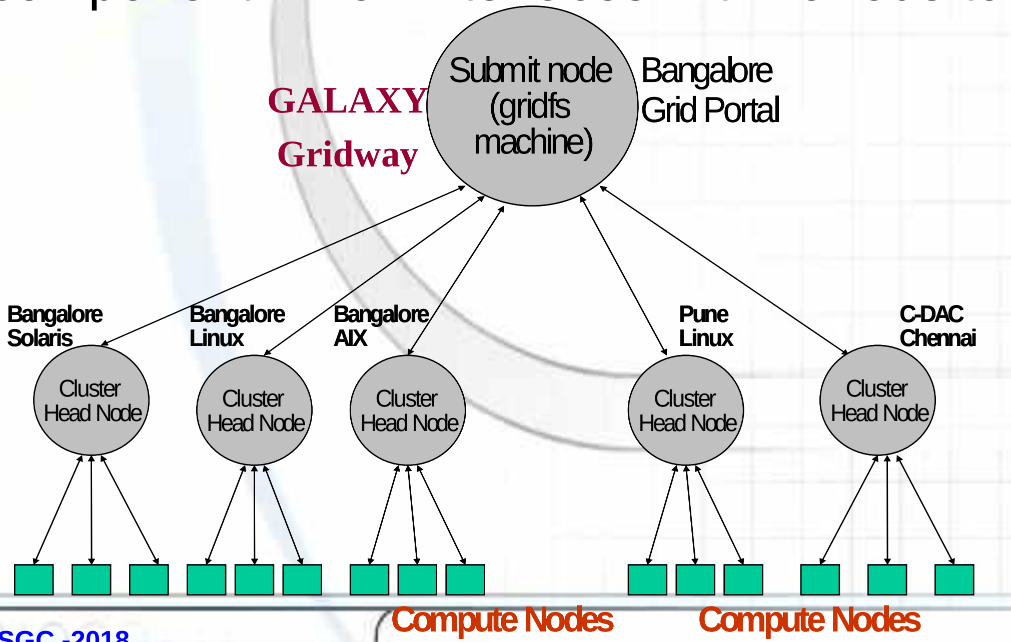
# Galaxy Workflow Architecture

The core components of the Galaxy Framework are *the toolbox*, *the job manager*, *the model*, and *the web interface*

- *Toolbox* - manages all of the details of working with command-line and web-based computational tools.
- *Job manager* - deals with the details of executing tools. It manages dependencies between jobs (invocations of tools) to ensure that required datasets have been produced without errors before a job is run.
- *Model* - provides an abstract interface for working with datasets. It provides an object-oriented interface for working with dataset content.
- *Web interface* - provides support for interacting with a Galaxy instance through a web browser.

# Garuda – Galaxy Architecture

- Galaxy has been deployed on GARUDA Grid Headnode and can be accessed by the user.
- This Grid Headnode is connected to several compute cluster resources.
- At the Grid Headnode Gridway meta-scheduler is present which interacts with LRMs on each of the clusters’ headnodes.
- Execution of a  tool (or workflow) from Galaxy happens based on the load scheduling by Gridway.
- Galaxy has a job manager component which interfaces with various tools’ parameters for execution.

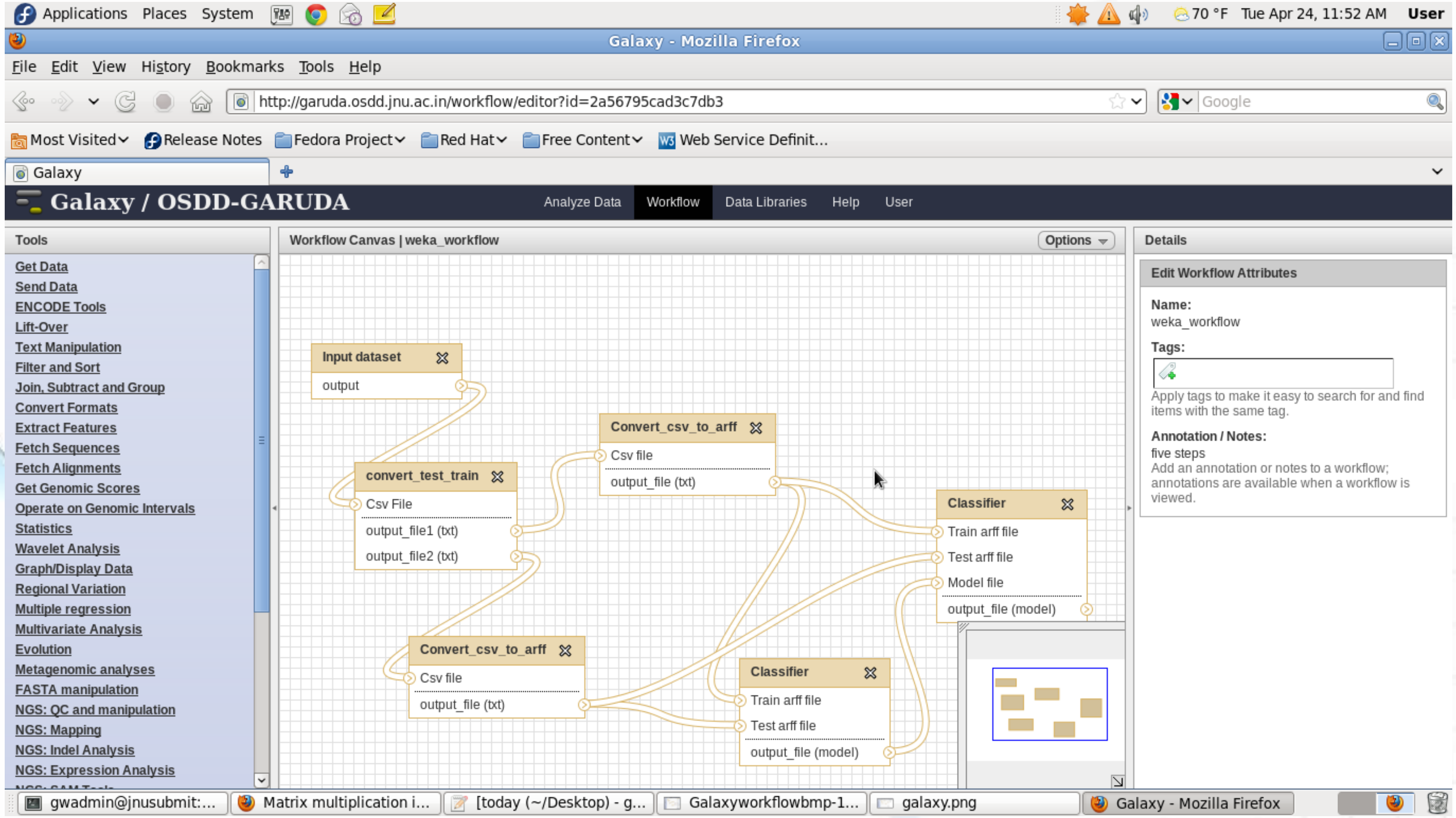


# Features: Garuda – Galaxy Gateway

- Integrated with Grid Authentication mechanism- Indian grid certificate Authority
- Integrated with Gridway Metascheduler to provide Job control
  - Job status change message displayed
  - Recovery of already running jobs, if the galaxy server restarts in between.
  - Whenever any job id is deleted from the users history on click of close button, the job is also deleted from the gwps.
- Integrated tools- Weka(for data mining) and Autodock(Virtual screening)
- Remote download of output/results with user defined names
- Bug report feature enabled with Job id in the subject.
- Data Log

# Results

- Galaxy workflow has the provision to visualize the output and errors files in the browser.
- These output and error files can also be downloaded at the user's desktop
- Various tools like Autodock, Namd, weka, gromacs has been added in this instance of galaxy tool shed.



**Name:**
weka_workflow

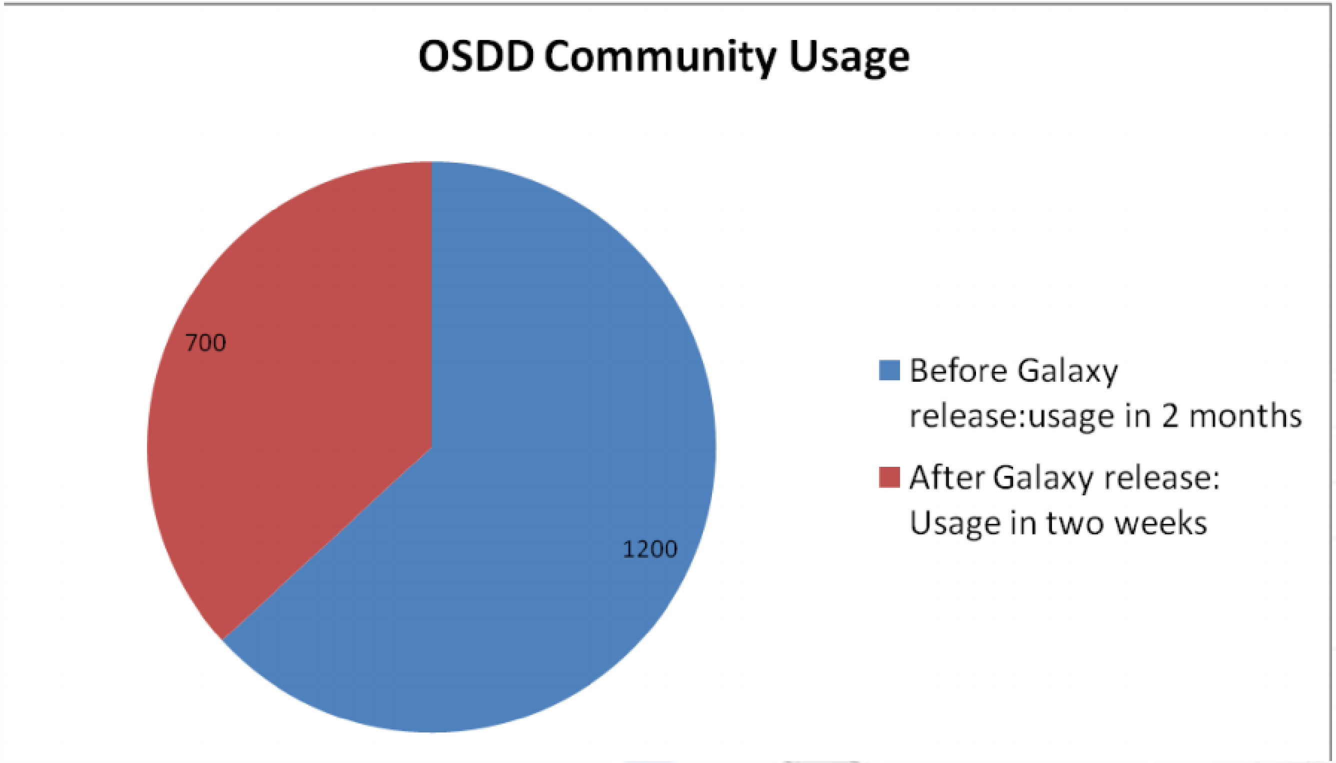
**Tags:**

Apply tags to make it easy to search for and find items with the same tag.

**Annotation / Notes:**
five steps
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Galaxy Workflow using Weka

# GARUDA Usage by OSDD Community



- OSDD is Open Source Drug Discovery Community initiated by CSIR, Govt of India
- >70 OSDD users became members of Garuda
- Galaxy is being used by OSDD members for Insilico Screening in Drug discovery pipeline

# Conclusions

- Galaxy is an open, web-based platform for data intensive biomedical research.

- It is been successfully demonstrated that Galaxy can be extended to the various environments like grid to exploit its computational power.

- Galaxy has been designed in a modular fashion making it easy to integrate with different schedulers and making any feature enhancements.

- The web based tool deployed on the grid headnode is accessible via a browser from individual researchers’ desktop.

# Acknowledgements

- Authors acknowledge Dr. Anshu Bhardwaj and Dr. Abdul U C Jaleel and other members of the Open Source Drug Discovery (OSDD) Community for collaborative work on GARUDA OSDD scientific workflow.

- Authors acknowledge the support of Department of Electronics and Information Technology (DeitY), Ministry of Communication and Information Technology (MCIT), India for GARUDA Grid Project.

- Authors also acknowledge and thank the support provided by Executive director of C-DAC, Bangalore, Chief-Investigator of the Garuda, Middleware & Operations Teams of Garuda and all the colleagues who have helped in accomplishment of the work.

# Thank You