UK CYVERSE™

CyVerse UK: widening the scope to the UK and beyond

- Bringing CyVerse to the UK
- UK CyberInfrastructure
- Collaborations

Alice Minotto
Alice.Minotto@earlham.ac.uk, @rosysnake

Robert Davey
robert.davey@earlham.ac.uk, @froggleston

Earlham Institute

Decoding Living Systems

UK CYVERSE™

CyVerse UK is the first European instance of CyVerse

Funded by the BBSRC funding agency as a pilot project in 2015

The project hardware and devops staff are hosted at the Earlham Institute

BBSRC bioscience for the future

Earlham Institute

UK CYVERSE™

Main purposes of the project are:

- give research groups access to life science cyberinfrastructure in the UK
- allow the analysis of data on UK-provisioned research HPC
- ensure reproducibility of analysis through application versioning / containerisation
- support data sharing
- distribute documentation and code as open source

CyVerse UK aims to give UK and EU users a geographical advantage, though it is available to users worldwide.

BBSRC
bioscience for the future

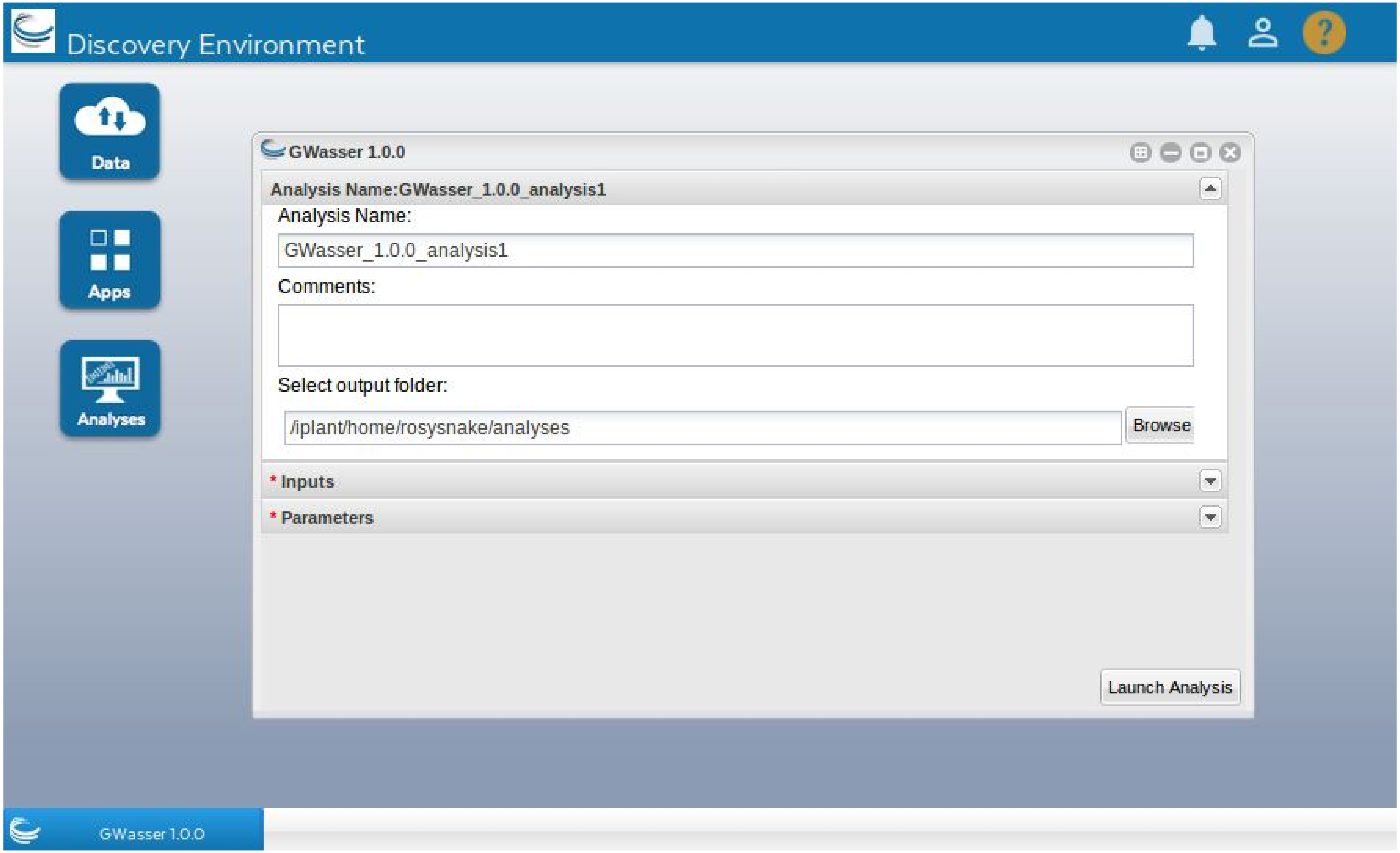
Earlham Institute

CyVerseUK Apps

- BAM to Fastq
- Beast 2.4.3 (with Beagle) on CyverseUK
- Blastn
- Blastp
- Bowtie2
- Btrim 0.3.0
- Concatenate multiple files
- CoverageBED
- Create Blast Database
- Cuffdiff up to 16 conditions
- Cufflinks
- Cuffmerge
- Cut columns
- Cutadapt SE
- GWasser-1.0.0
- Index BAM and get stats
- Index BAM file
- Index FASTA file (faidx)
- Kallisto v0.43.0
- MUSCLE
- Mikado-1.0.0
- MikadoCompare-1.0.0
- MikadoUtil-1.0.0
- MrBayes 3.2.6 (parallel) on CyverseUK
- Ontologizer 2.1
- Portcullis-1.0.0
- RAxML on CyverseUK
- Retrieve-Brassica-Phenotypes
- SAM to sorted BAM
- Sickle PE
- Sickle SE
- Sort BAM
- TopHat2 PE
- TopHat2 SE
- Tuxedo suite PE up to 4 conditions
- Uncompress files with gunzip

BUT… applications need to be duplicated to be run both in the US and in the UK.

User interfaces



Analysis Name:GWasser_1.0.0_analysis1

Analysis Name: GWasser_1.0.0_analysis1

Comments:

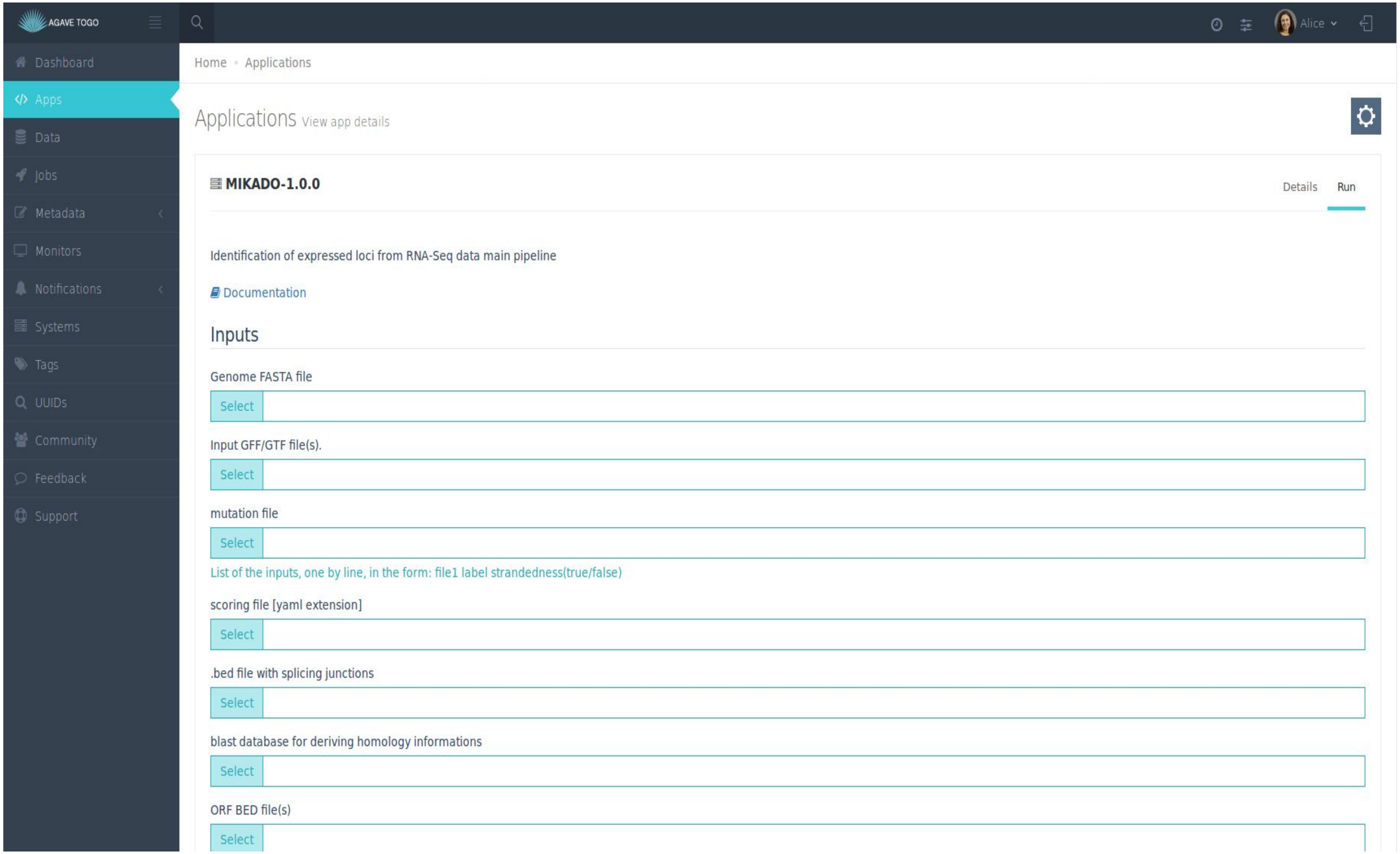
Select output folder: /iplant/home/rosysnake/analyses

* Inputs

* Parameters

Earlham Institute

User interfaces



Applications View app details

MIKADO-1.0.0

Identification of expressed loci from RNA-Seq data main pipeline

Documentation

Inputs

Genome FASTA file

Select

Input GFF/GTF file(s).

Select

mutation file

Select

List of the inputs, one by line, in the form: file1 label strandedness(true/false)

scoring file [yaml extension]

Select

.bed file with splicing junctions

Select

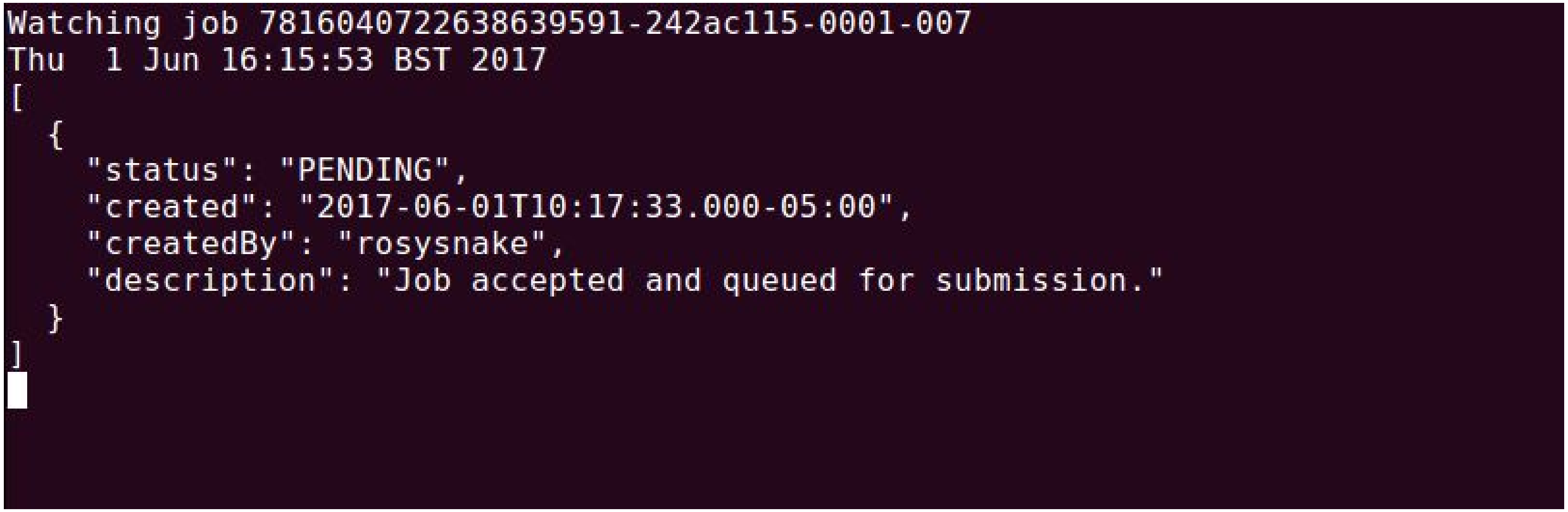
blast database for deriving homology informations

Select

ORF BED file(s)

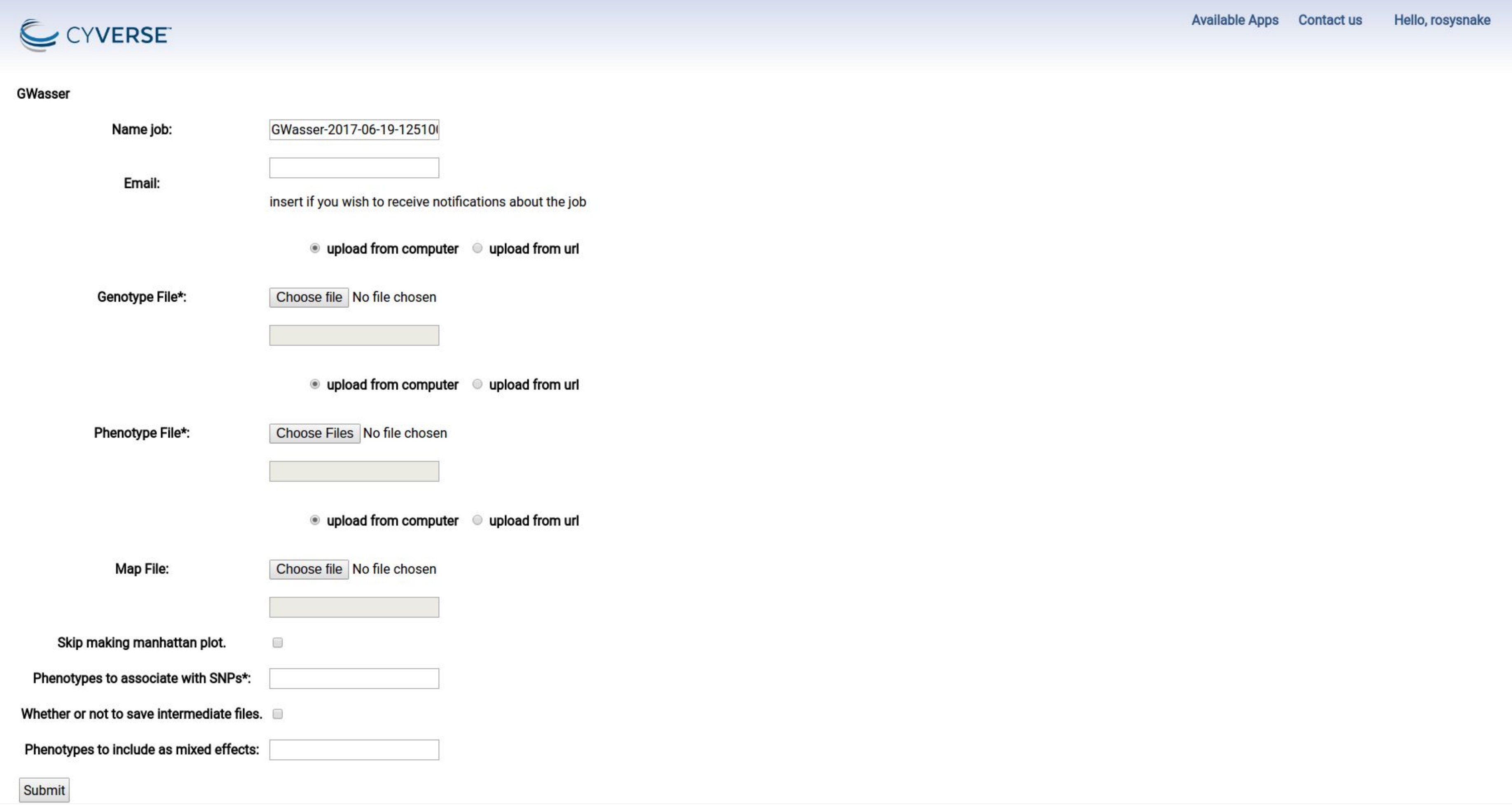
Select

User interfaces



```
Watching job 7816040722638639591-242ac115-0001-007
Thu  1 Jun 16:15:53 BST 2017
[
  {
    "status": "PENDING",
    "created": "2017-06-01T10:17:33.000-05:00",
    "createdBy": "rosysnake",
    "description": "Job accepted and queued for submission."
  }
]
```

User interfaces

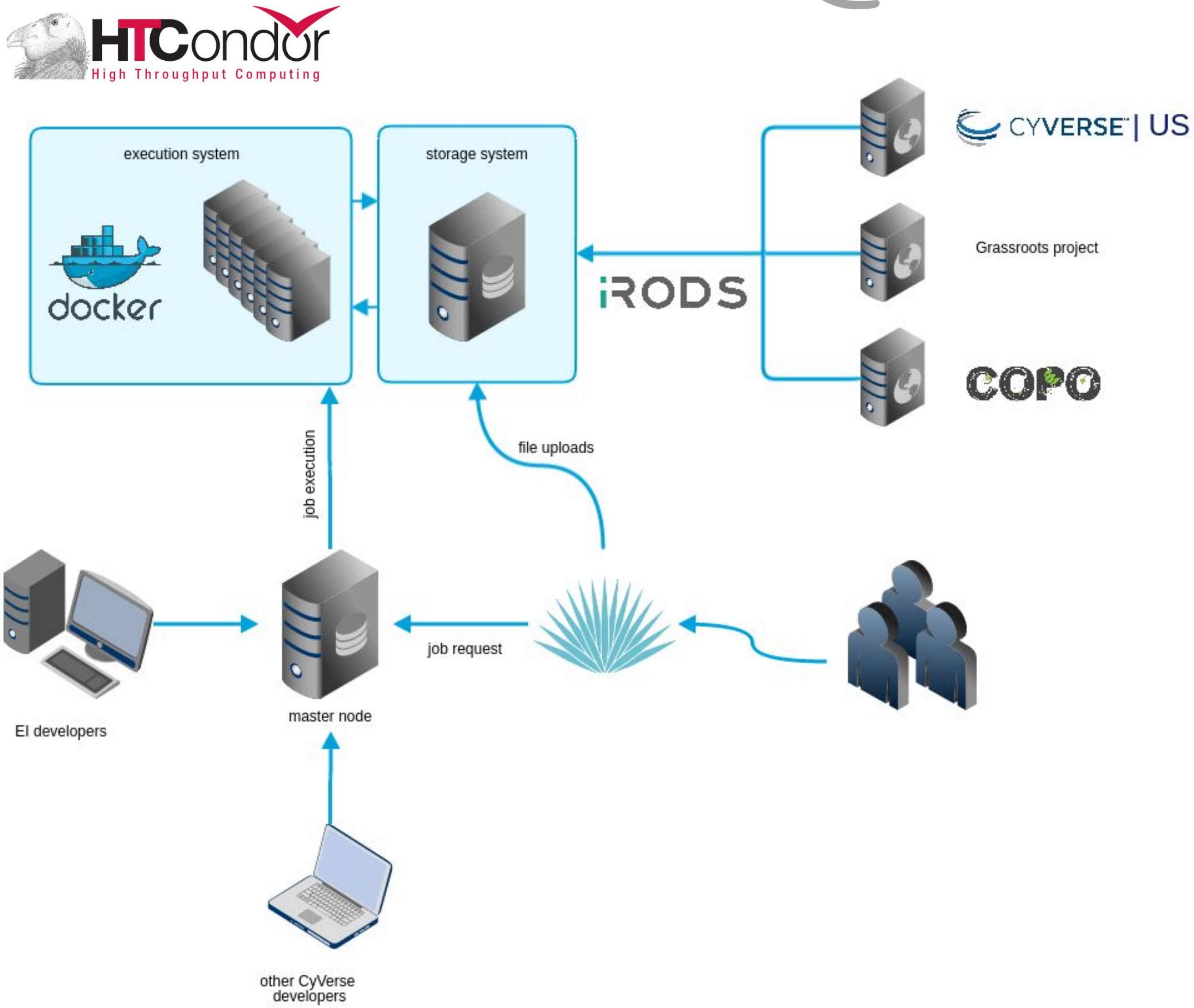


GWasser

Name job:	GWasser-2017-06-19-12510(
Email:	insert if you wish to receive notifications about the job
	● upload from computer ○ upload from url
Genotype File*:	Choose file No file chosen
	● upload from computer ○ upload from url
Phenotype File*:	Choose Files No file chosen
	● upload from computer ○ upload from url
Map File:	Choose file No file chosen
Skip making manhattan plot.	☐
Phenotypes to associate with SNPs*:	
Whether or not to save intermediate files.	☐
Phenotypes to include as mixed effects:	

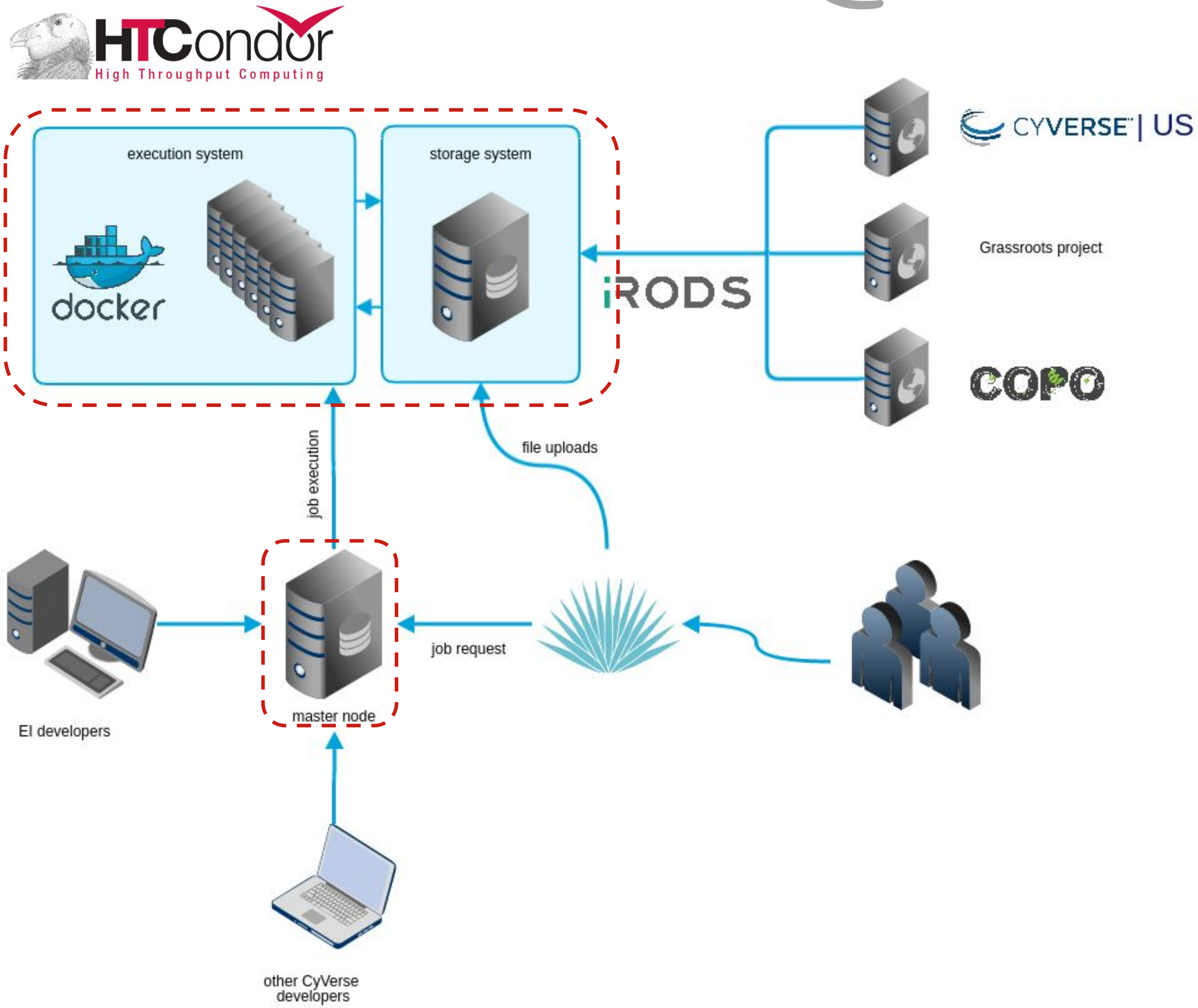
The infrastructure

UK CYVERSE™



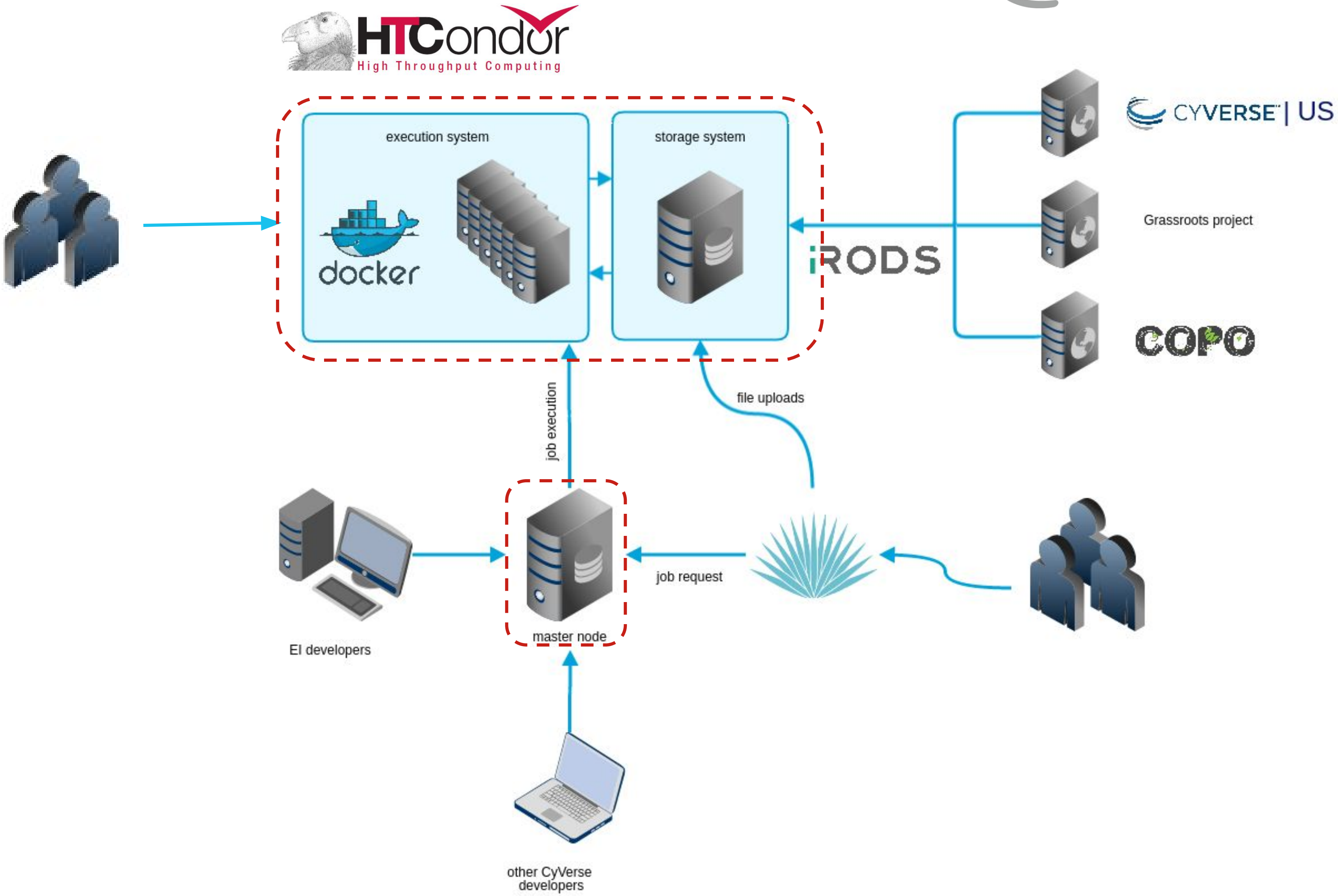
Earlham Institute

UK CYVERSE™



Earlham Institute

UK CYVERSE™



Earlham Institute

“Docker is the world’s leading software containerisation platform.”

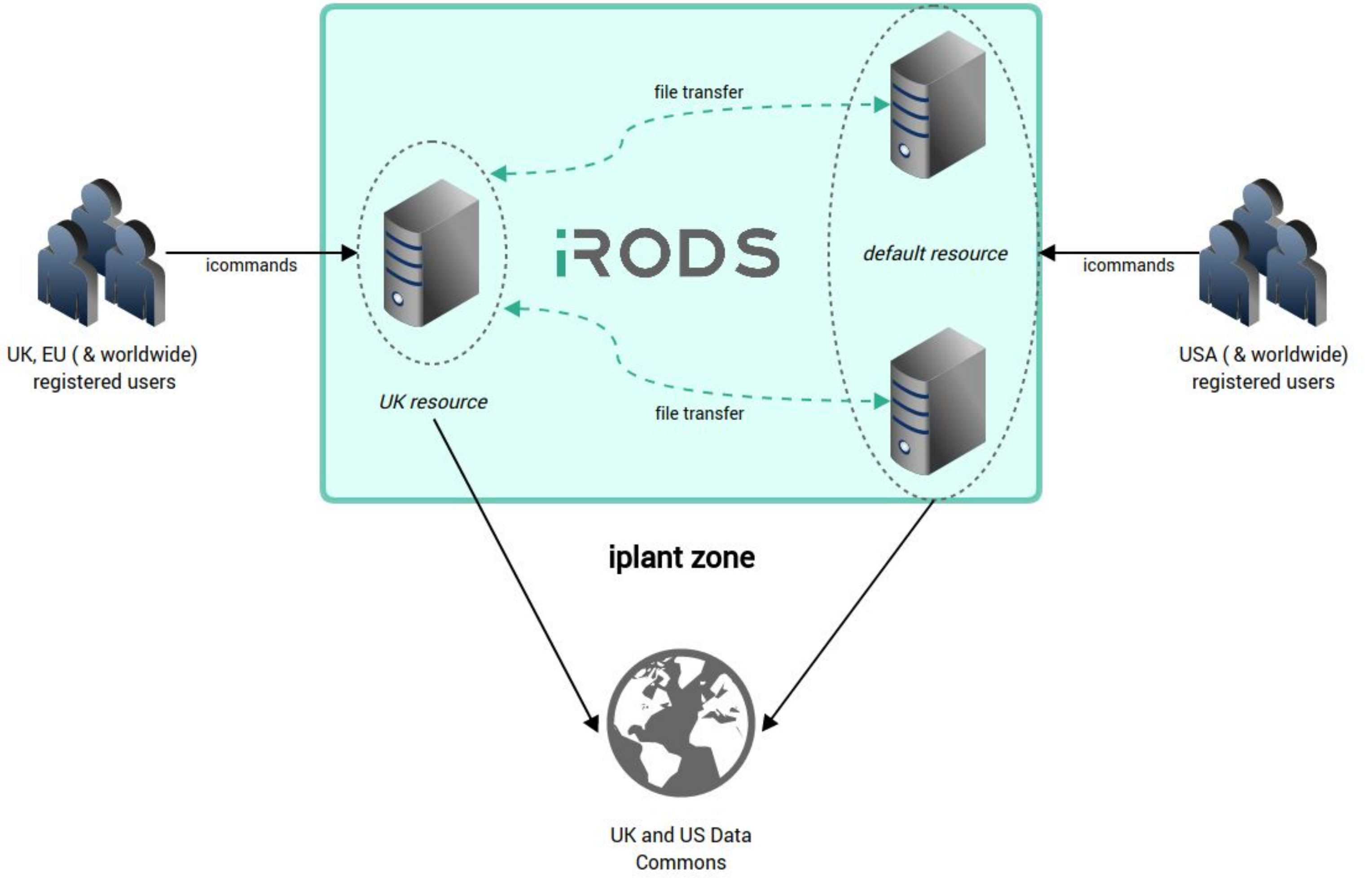
- Wrap a piece of software in a self-contained package
- Contains everything needed to run: code, runtime, system tools, system libraries
- Small portable server
- Software will always run the same, regardless of where it is run

The infrastructure

Work in Progress and Future Developments

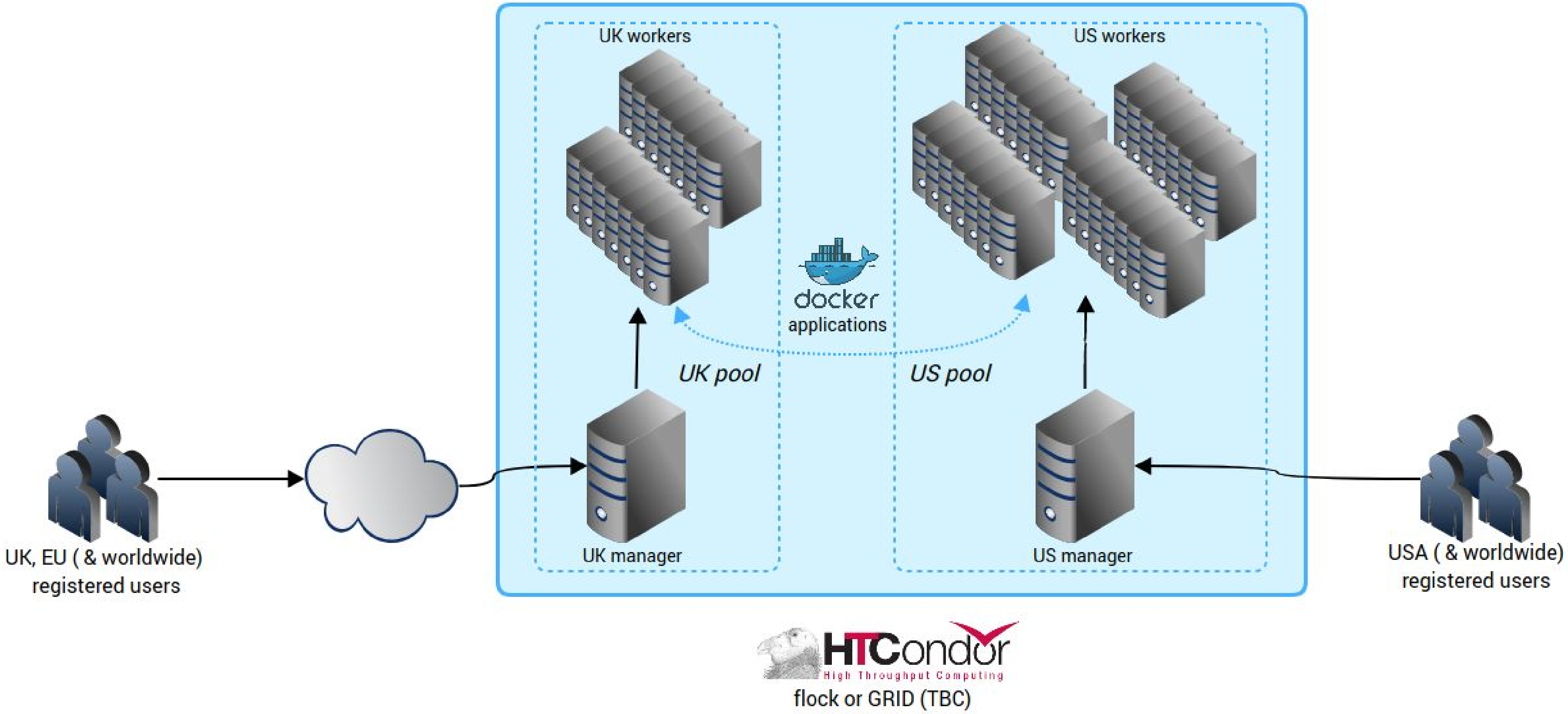
Earlham Institute

iRODS Storage Federation



HTCondor Federation

- Federation between US/UK HTCondor pools
  - Users grouped by location
  - User interface stays the same

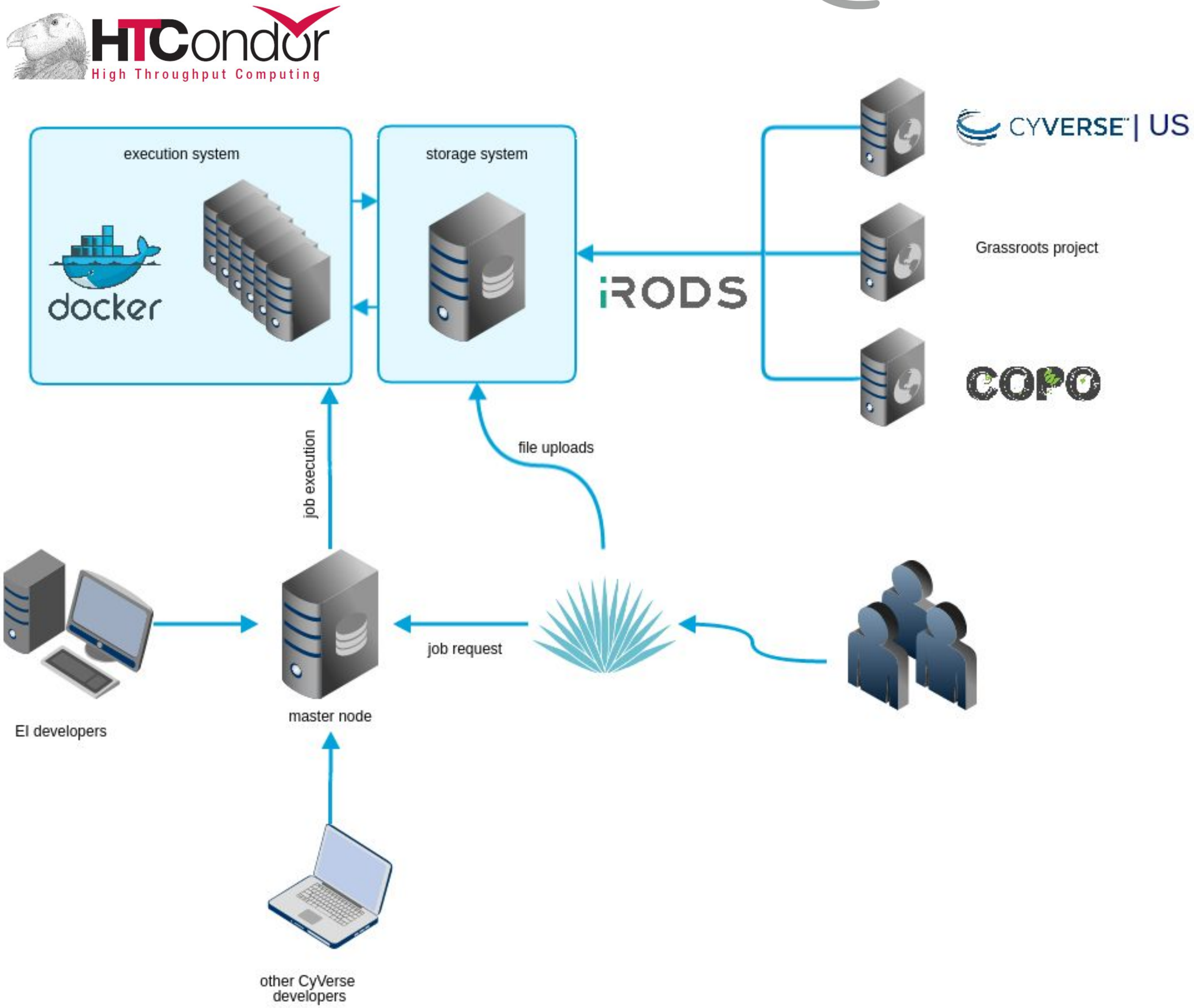


- Agave application to become DE Docker application
  - **Enable shared applications between UK and US**

Earlham Institute

Collaborations

UK CYVERSE™



HTCondor
High Throughput Computing

execution system

docker

storage system

iRODS

CYVERSE™ | US

Grassroots project

COPO

job execution

file uploads

EI developers

master node

job request

other CyVerse developers

Earlham Institute

Expansion and Community Support

Won a £400k BBSRC 16ALERT grant to expand the compute capacity of CyVerse UK

- Larger core count and RAM
- Better networking and switching
- Project-initiated VM requests to provide specific groups with cloud access
- Planning to include flexible analysis frontends such as the Genomics Virtual Laboratory

?

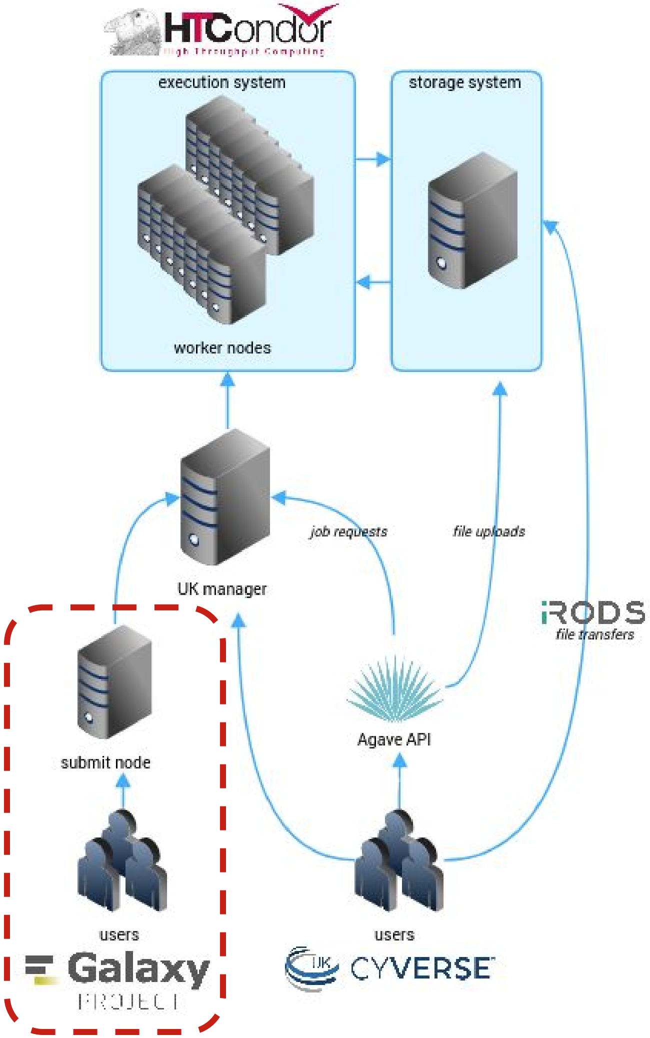
- Now providing webservice VMs to:
    - SignaLink
    - COPO
    - Grassroots

Galaxy Integration

Set up a pilot Galaxy integration project to demonstrate service provision

Users at BecA-ILRI in Nairobi, Kenya can submit Galaxy jobs to the CyVerse UK execution system:

- taking advantage of the CyVerse UK infrastructure
- Researchers can share and analyse data already available in the storage system (upload once)
- Reduces need for hardware on-site, and can promote collaboration
- Helps us understand limitations of intercontinental internet



CYVERSE™

Acknowledgements:

Erik Van Den Bergh
Robert Davey
Annemarie Eckes
Luis Yanes
Nicola Soranzo
Anthony Etuk
Xingdong Bian
Simon Tyrrell

Ben Ward
Ricardo Ramirez-Gonzalez
Luca Venturini
Daniel Mapleson

Sam Gallop
Adam Carrgilson
Chris Bridson
etal.

Tony Edgin
Matt Vaughn
John Fonner
Rion Dooley
Ramona Walls
Joe Stubbs
Ian McEwen
Nirav Merchant
etal.

CyVerseUK is supported by the BBSRC through strategic funding and the capital grant BB/M018431/1.

CyVerse is supported by the National Science Foundation under Award Numbers DBI-0735191 and DBI-1265383. URL: www.cyverse.org

BBSRC bioscience for the future
EXCELLENCE WITH IMPACT COMMENDATION WINNER
norwich research park
NATIONAL INSTITUTES OF BIOSCIENCE

Earlham Institute
Decoding Living Systems