Analysis of Metagenomes on the EGEE Grid

Gabriel Aparicio

Ignacio Blanquer

Vicente Hernández

Valencia University of Technology

(Universidad Politécnica de Valencia)

www.eu-egee.org

Information Society and Media

Topic summary

- **Introduction**
  - Definitions and objectives.
  - Case studies.
- **Metagenomic Analysis System**
  - Design and Deployment.
  - Automation.
- **Results and Performance.**
- **Conclusions and Future Plans.**

Introduction

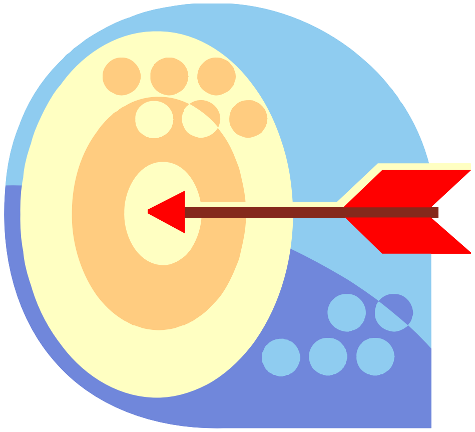
- **Definitions**
  - A metagenome is a collection of genes which can be studied as a single gene without isolating them.
  - A Metagenome Analysis is the group of necessary steps to transform a file of a coded metagenome into another file with some interest information.
  - This can include:
    - Database filtering.
    - BLAST alignments.
    - BLAST output filtering.
    - Creation of Phylogenetic Trees.

Introduction

- **Why Grid is a Good Solution?**
  - A Metagenome can be coded into several hundred of thousand sequences.
  - Sequential time can take more than a year.
  - Public databases are continuously changing.
  - Several steps can be done in parallel.
- **In a Grid, the global job can be divided into subjobs.**
  - A Metagenome Analysis can be processed in a few days with a Grid Infrastructure.

Introduction

- **Objectives**
  - Evaluate and validate the EGEE Grid infrastructure to develop the analysis of a large metagenome.
  - Develop a framework to perform multiple alignment and phylogenetic analysis for metagenomes.
    - Efficient and “infrastructure”-friendly.
    - Fault tolerant (jobs and output).
    - Semi-automatic.
  - Operate this framework for several large experiments

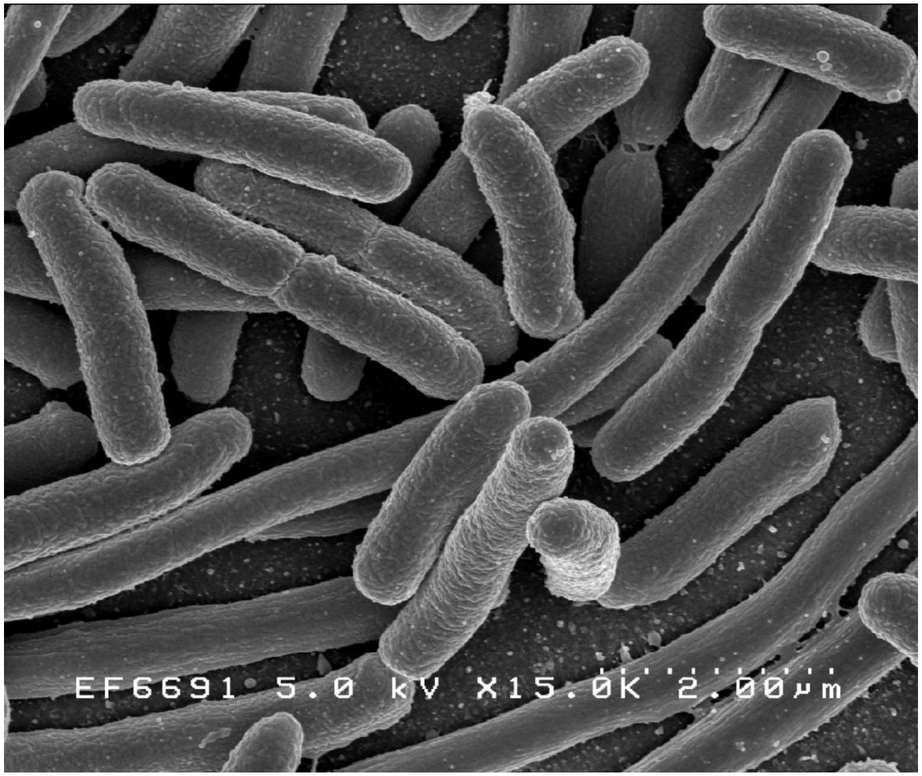


Case Studies - Biological

- **Farm Soil**
  - A sample from a nutrient-rich and moderately contaminated soil environment.
    - This community is very diverse and complex.
    - Many yet unknown enzymes are probably present there.
- **Whale Fall**
  - Sample from a whale carcass.
    - They are known to be a nutrient-rich environment in the bottom of the ocean.
    - A heterogeneous mixture of bacteria flourish there.
- **Sargasos’s Sea**
  - Oceanic samples taken from surface waters.
    - They represent the diversity of bacteria that live planktonically

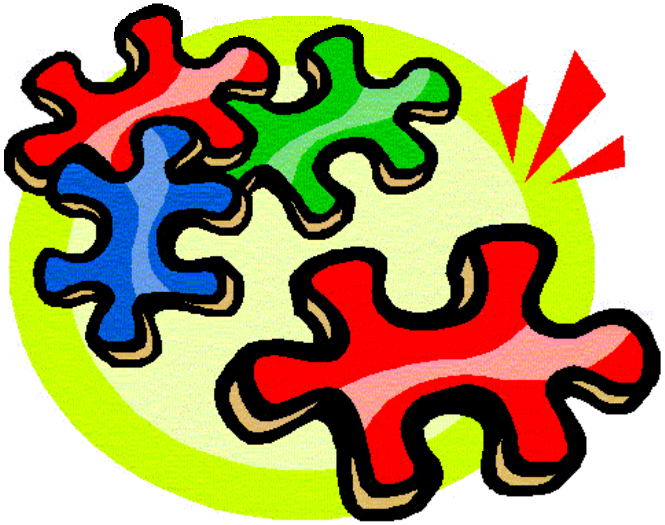
Case Studies – Medical

- **Gut Metagenomes**
  - Several metagenomes of the human intestinal microbiota.
  - A consortia of bacteria that helps its host to metabolize many nutrients that would be indigestible otherwise.
  - It is involved in other functions
    - Maturation and modulation of the immune response of the host.
    - Prevention of infection by bacterial pathogens.



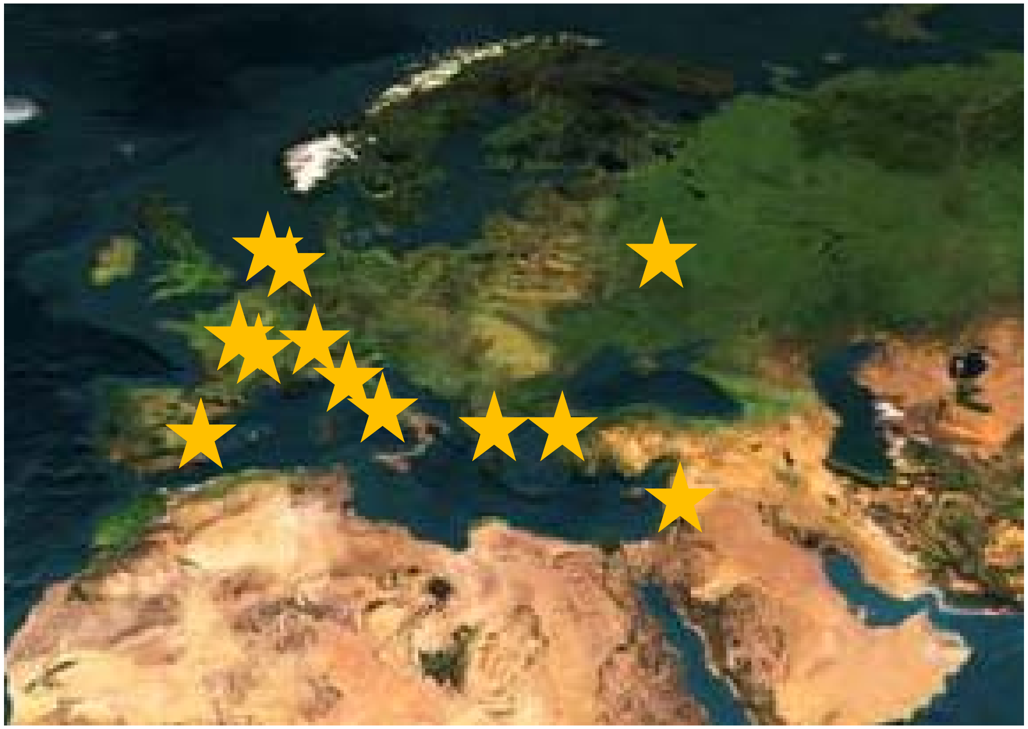
Deployment

- **Stages and Components**
  - Pre-processing
    - Data filtering, splitting and replication.
  - Submission and monitoring
    - Submission and re-submission components.
    - Parallel and sequential processing engines (BLAST, mpiBLAST).
  - Results retrieval
    - Output Transfer.
    - Post-processing.



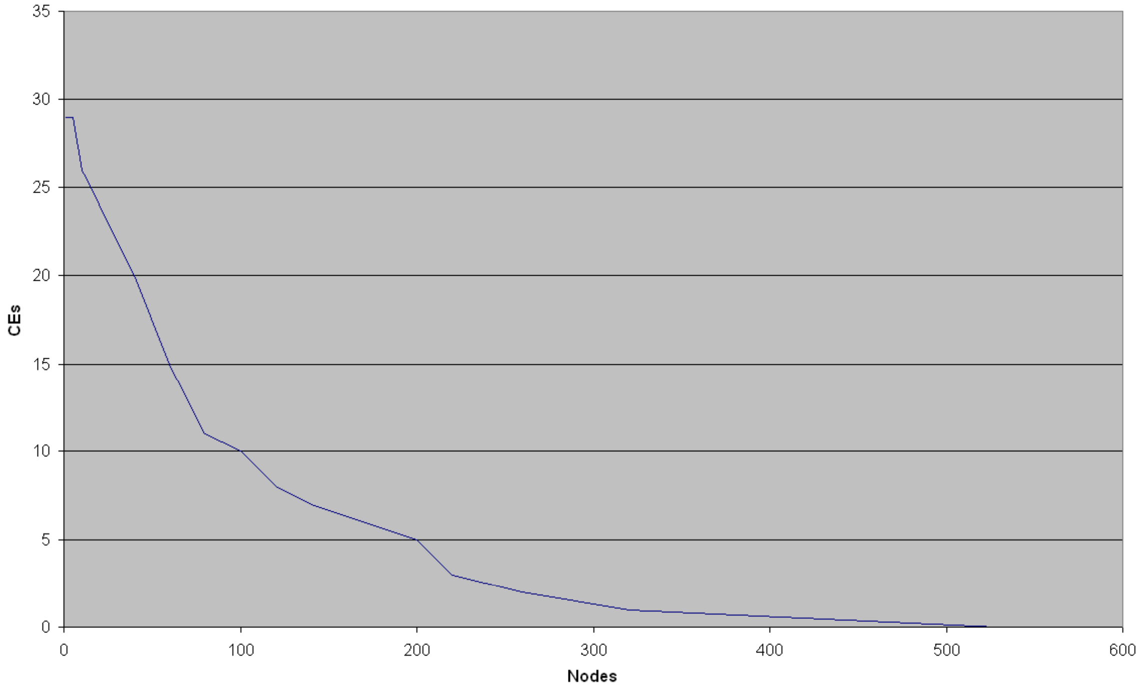
Deployment - Data

- **Selecting SEs and Replicating Files**
  - All jobs need certain common files.
    - A filtered nr database with sequences from procaryotic species.
  - These files have to be replicated to increase performance and to distribute network bandwidth.
  - SEs hosting is located according to their geographical and administrative closeness to the selected CEs, their performance and their configuration.
  - 12 Replicas have been made.



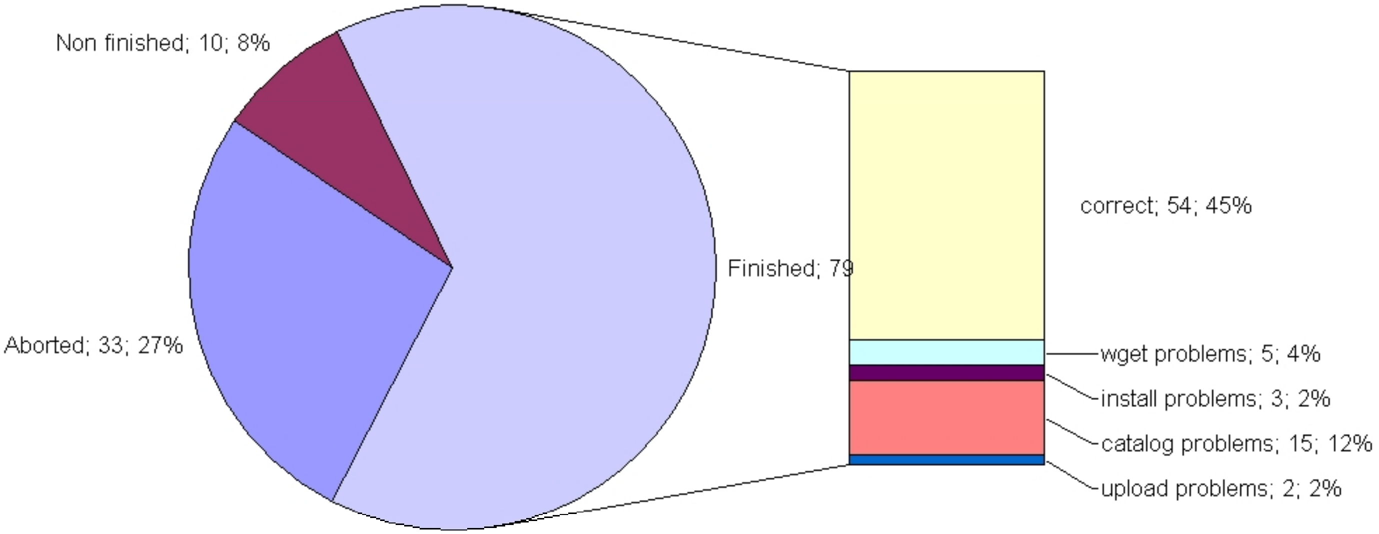
Deployment – CEs

- **Sequential or parallel BLAST kernels?**
  - There are around 122 CEs in BIOMED VO.
  - There are only around 30 CEs able to run MPICH jobs.
  - The number of CEs decreases when the number of required nodes increases.
  - Full efficiency in MPICH jobs is achieved occasionally.
  - About 1000 CPUs for MPI jobs and About 17000 for Sequential Jobs.



Deployment - CEs

- **Selecting CEs**
  - Not all available CEs are able to produce results.
  - Not all available CEs have the same performance.
  - CEs need to be selected to distribute jobs according to their performance.

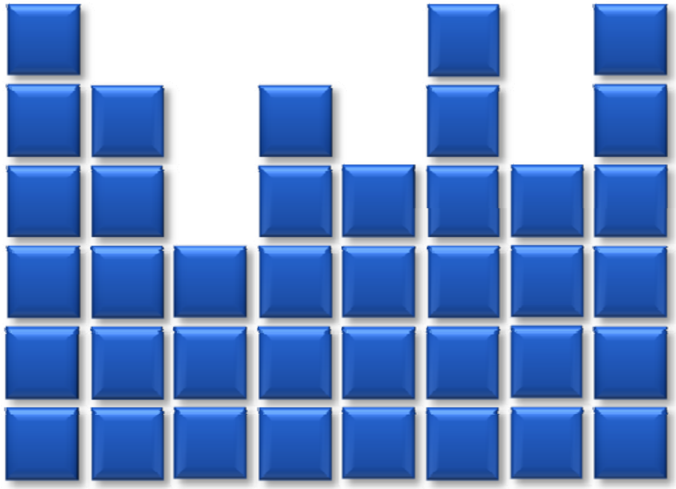


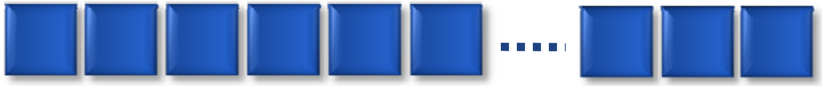
Deployment – Filtered CEs

- **Selecting CEs**
  - Even Removing the CEs Producing Errors, there are Temporary Errors that Affect Almost any CE.
  - In an Experiment Involving only those CEs, 41,2 % Were Successful in its first Execution and 58,8% of the Jobs Needed to be Resubmitted
    - 25,4% Were Aborted due to Unspecified Error (Error While In CondorG Queue).
    - 11,1% Were Aborted due to Errors with the Catalogue (lcg-cp Mainly, and Sometimes lcg-cr).
    - 9,4% Were Aborted due to the Expiration of the Proxy (VOMS Credentials are Limited to 168 hours).
    - 7,0 % Were Aborted due to Authentication Errors (Globus Error 7).
    - 3,7% Were Aborted due to JobWrapper Errors (Cannot read JobWrapper output, both from Condor and from Maradona).
    - 2,3% Were Cancelled due to Excessive Waiting Time.

Deployment – Job Splitting

- **Splitting global job**
  - The global job has to be broken down into subjobs
    - The number of jobs depends on the number of input sequences and the desired individual average duration.
  - The subjob lifetime will decrease
    - Increase interactivity.
    - Improve monitoring capabilities.





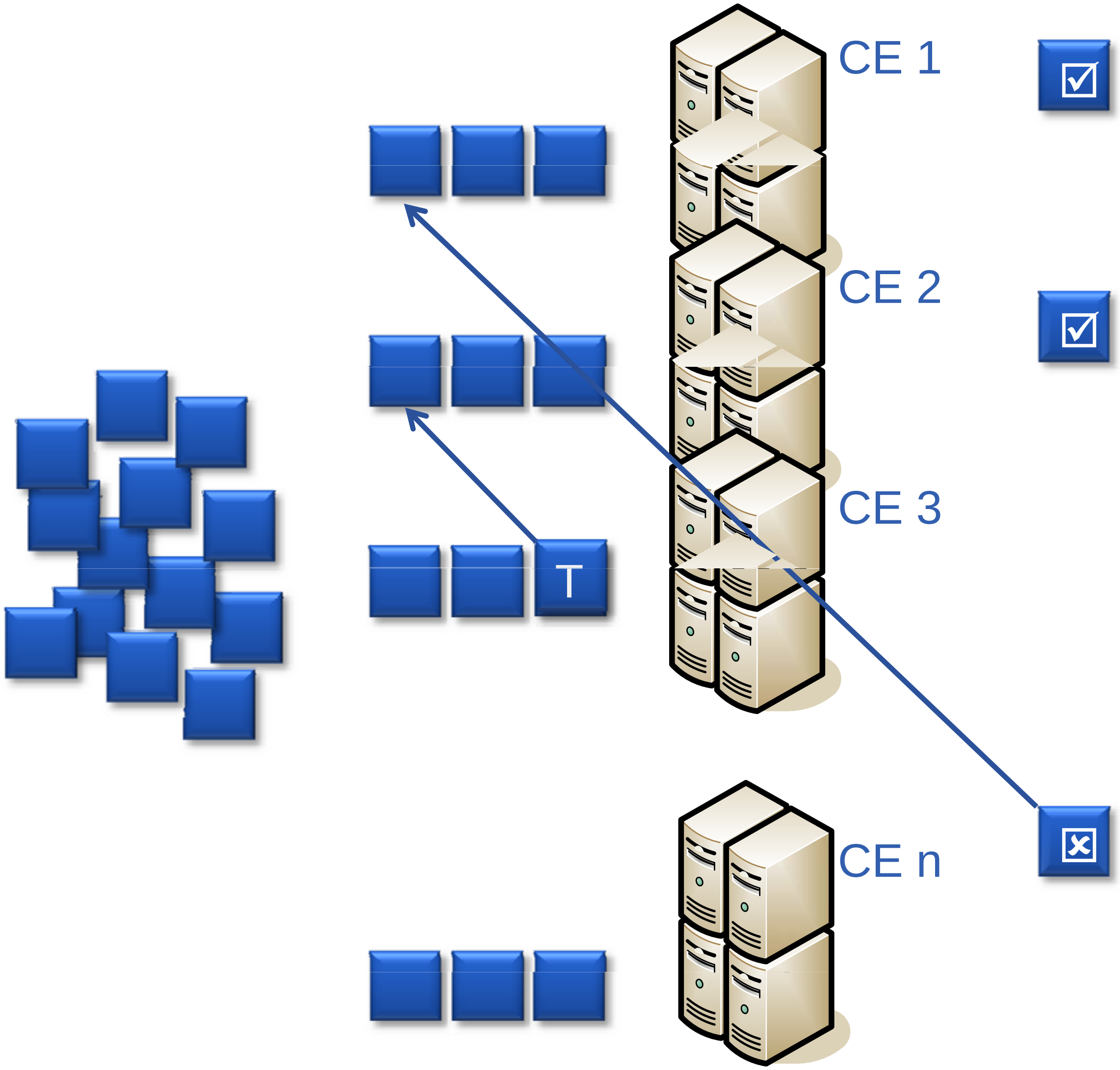
Automation – Job Submission

- **Submitting Jobs**
  - Subjobs are assigned to a list of CEs
  - These CEs have been tested.
  - Assignation is done according to obtained performances in previous experiments.
- **Monitoring**
  - Periodically, jobs status are monitored.
  - In case of errors (aborted job, bad results, etc.), the job is automatically resubmitted.
  - In case the job is running too long, the job is cancelled and resubmitted.
  - In case the job has finished successfully, its CEs is saved for later submissions.

Automation

- **Resubmitting Jobs**
  - Each correctly finished job saves its CEs and puts it into a list.
  - The jobs are resubmitted to a random CE of this list.
  - If the list does not exist, the job is submitted to the same CE.
- **Retrieving Results**
  - Once results are available, they are downloaded and the standard outputs are explored to find any error.
  - A retrieved job is no longer monitored.

Job Splitting and Automation



Re-submission List

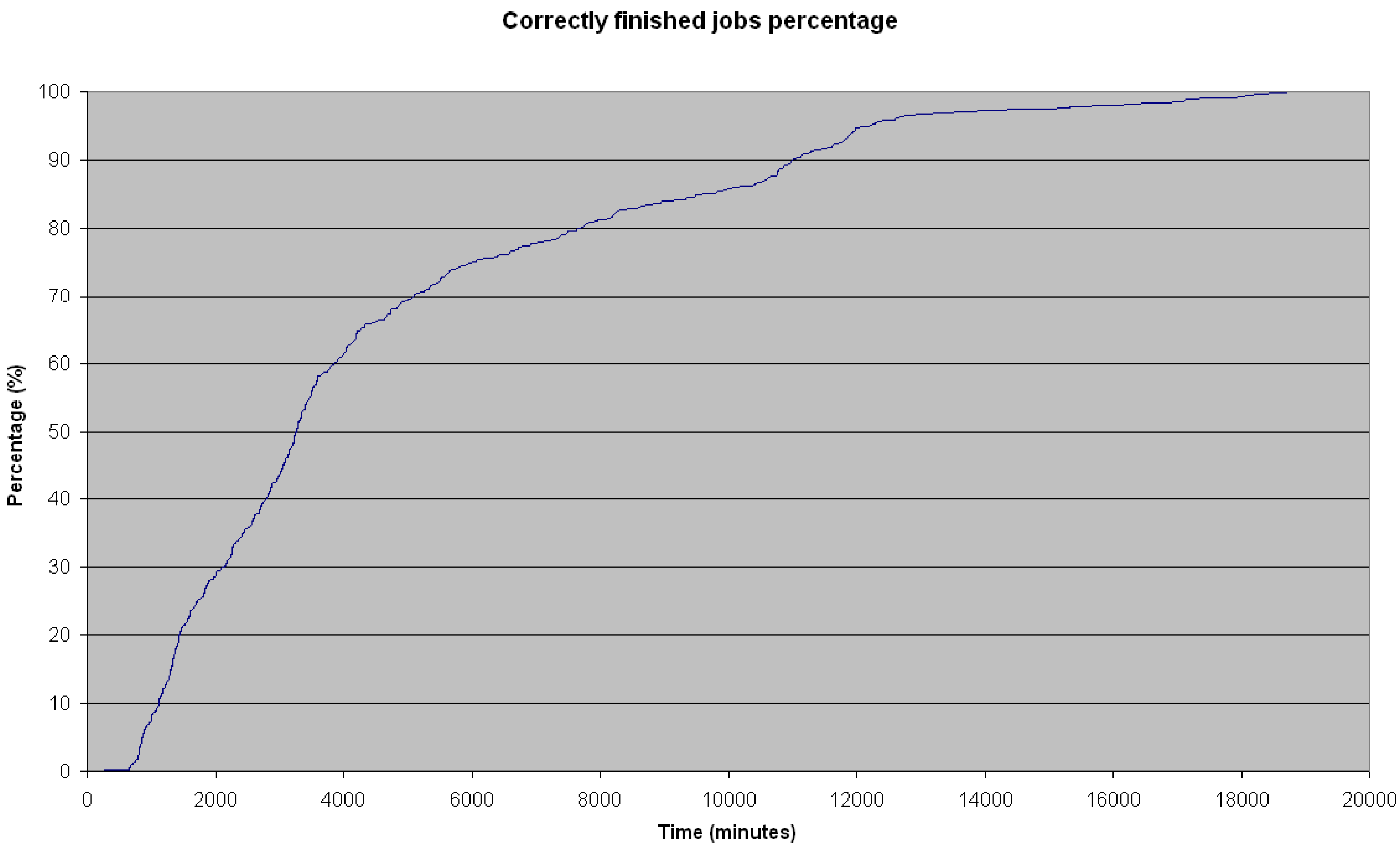
CE 1

CE 2

Results and Performances

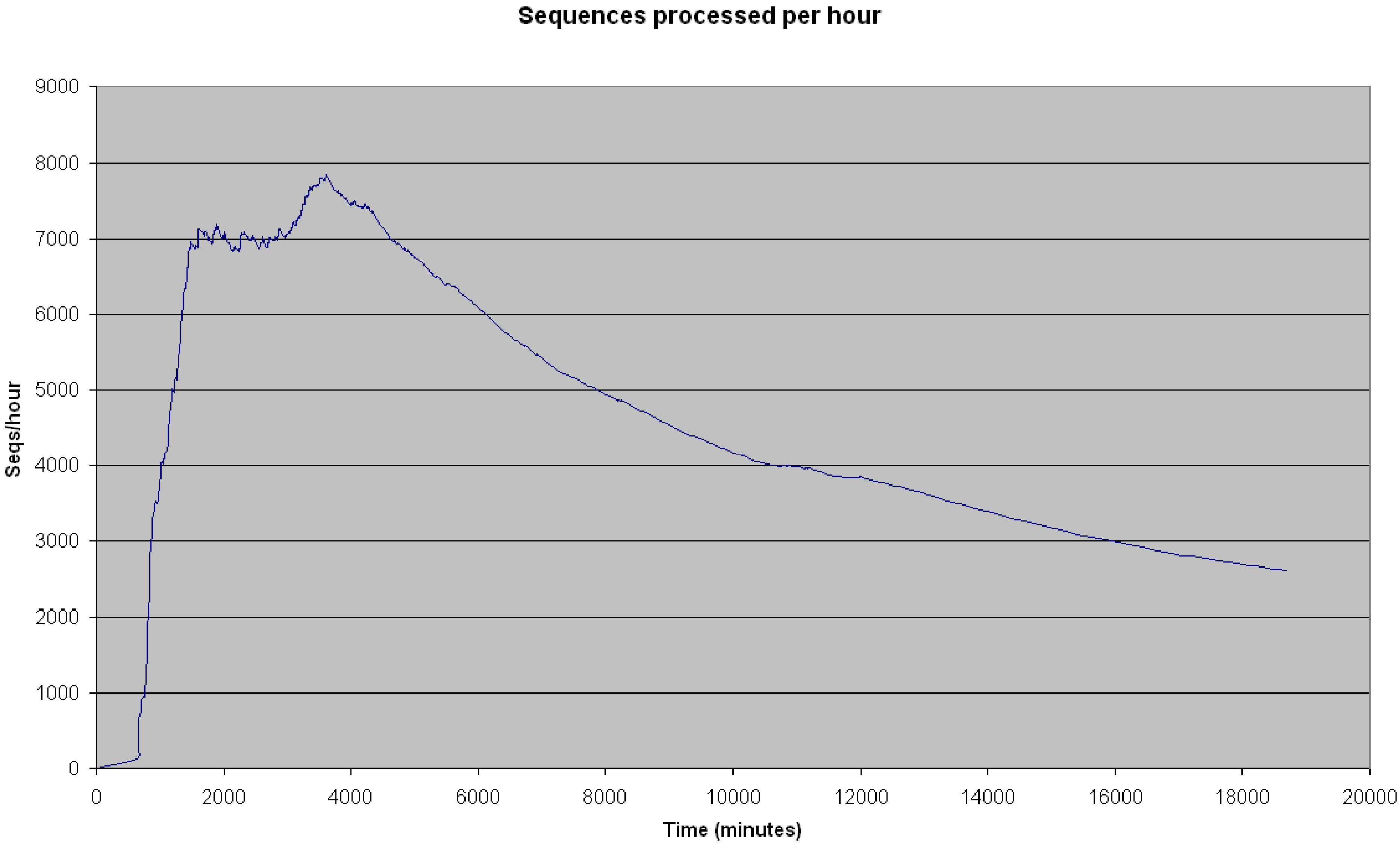
- **Jobs are too long to run sequentially**
  - Sargasso Sea Metagenome takes 512 days.
- **The same job in Grid takes 13 days to be fully finished.**
  - Speedup is around 40.
- **High speed for most jobs (90% in 7 days)**
  - Speedup is around 80.
  - No needed to finish all jobs to begin with new stages.

Results and Performances



Enabling Grids for E-sciencE

Results and Performances



Future plans

- **To create several shell-scripts with different stages depending on the desired results**
  - Cross-analysis of metagenomes, e.g.
- **To deal with new case studies**
  - 17 Metagenome Studies have been Processed so Far (About 10 CPU Years).
- **To improve automation performances**
  - Improve the selection of resources.
  - Improve the resubmission mechanism.

Conclusions

- **The EGEE grid has demonstrated to be a successful tool for metagenomic analysis.**
- **Metagenomic analysis involves several steps that require intensive computation**
  - There are many different experiments that can be defined and are not currently performed due to its cost.
- **The results obtained are successful and relevant from the users' point of view.**

Contact

Vicente Hernández / Ignacio Blanquer / Gabriel Aparicio

Universidad Politécnica de Valencia
Camino de Vera s/n
46022 Valencia, Spain
Tel: +34-963879743
Fax. +34-963877274
E-mail: vhernand@dsic.upv.es
iblanque@dsic.upv.es
gaparicio@itaca.upv.es