(GENOMICS) WORKLOAD ORCHESTRATION WITH NEXTFLOW

Paolo Di Tommaso, CRG

22 June 2017, ISC HPC, Frankfurt

nextflow

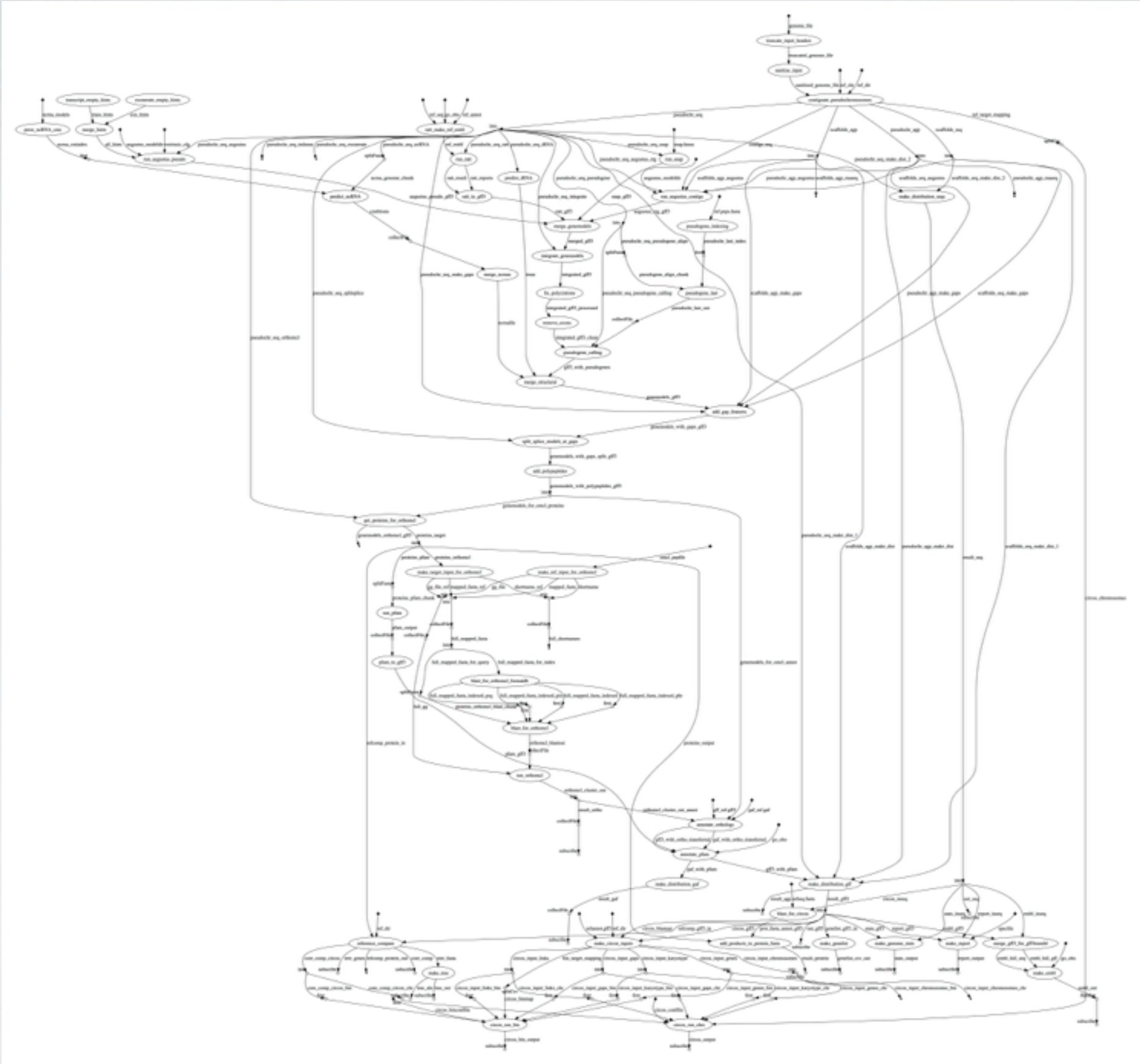
CRG Centre for Genomic Regulation

WHO IS THIS CHAP?

@PaoloDiTommaso
Research software engineer
Comparative Bioinformatics,
Notredame Lab
Center for Genomic Regulation (CRG)

GENOMIC WORKFLOWS

- Data analysis application to extract information from genomic datasets
- Mash-up of many different tools and scripts
- Embarrassingly parallelisation, can spawn 100-100k jobs over distributed cluster
- Complex dependency trees and configuration → very fragile ecosystem



* Companion parassite genome annotation pipeline, Steinbiss et al., DOI: 10.1093/nar/gkw292

OPEN ACCESS Freely available online

PLOS | ONE

Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome

Daniel Garijo[1], Sarah Kinnings[2], Li Xie[3], Lei Xie[4], Yinliang Zhang[5], Philip E. Bourne[3]*, Yolanda Gil[6]*

1 Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain, **2** Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, United States of America, **3** Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, United States of America, **4** Department of Computer Science, Hunter College, The City University of New York, New York, New York, United States of America, **5** School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China, **6** Information Sciences Institute and Department of Computer Science, University of Southern California, LosAngeles, California, United States of America

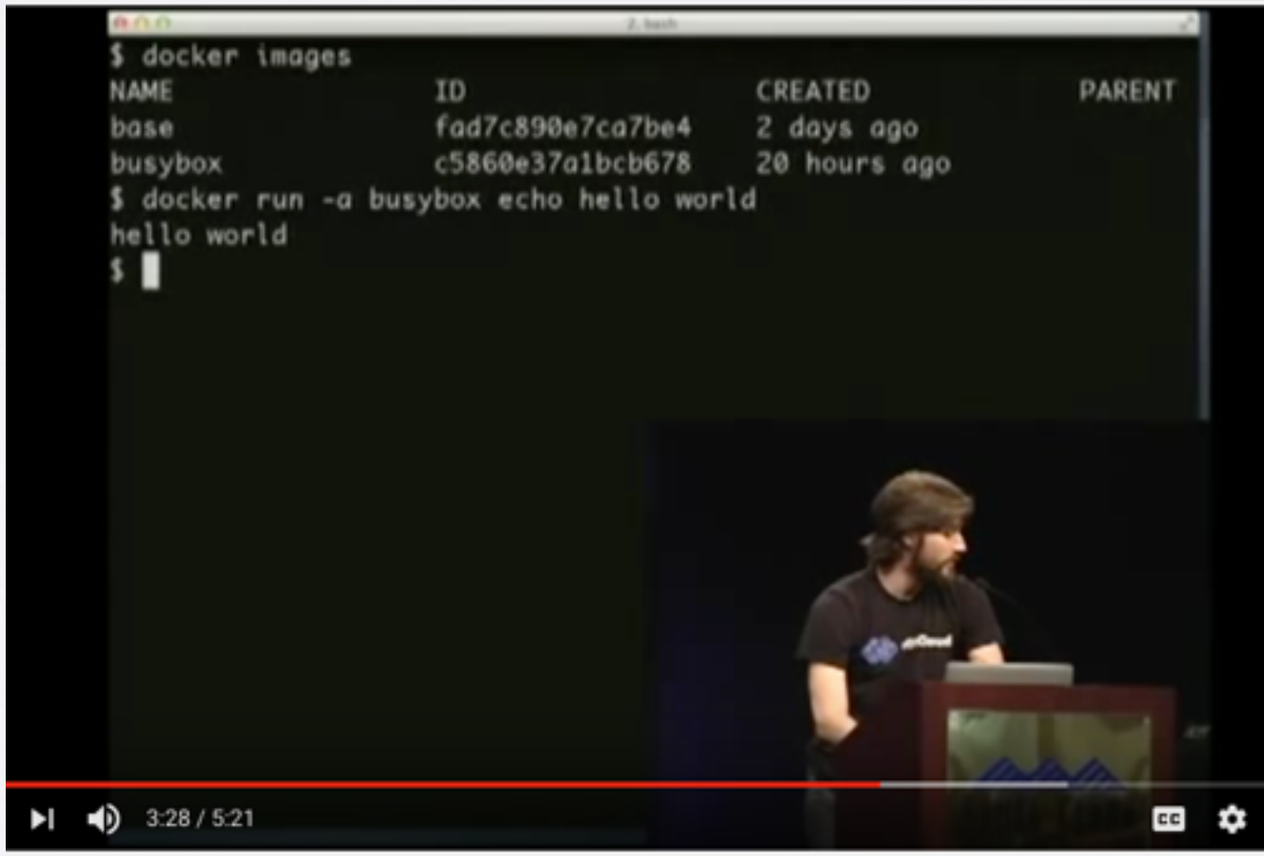
To reproduce the result of a typical computational biology paper requires 280 hours.

≈ 1.7 months!

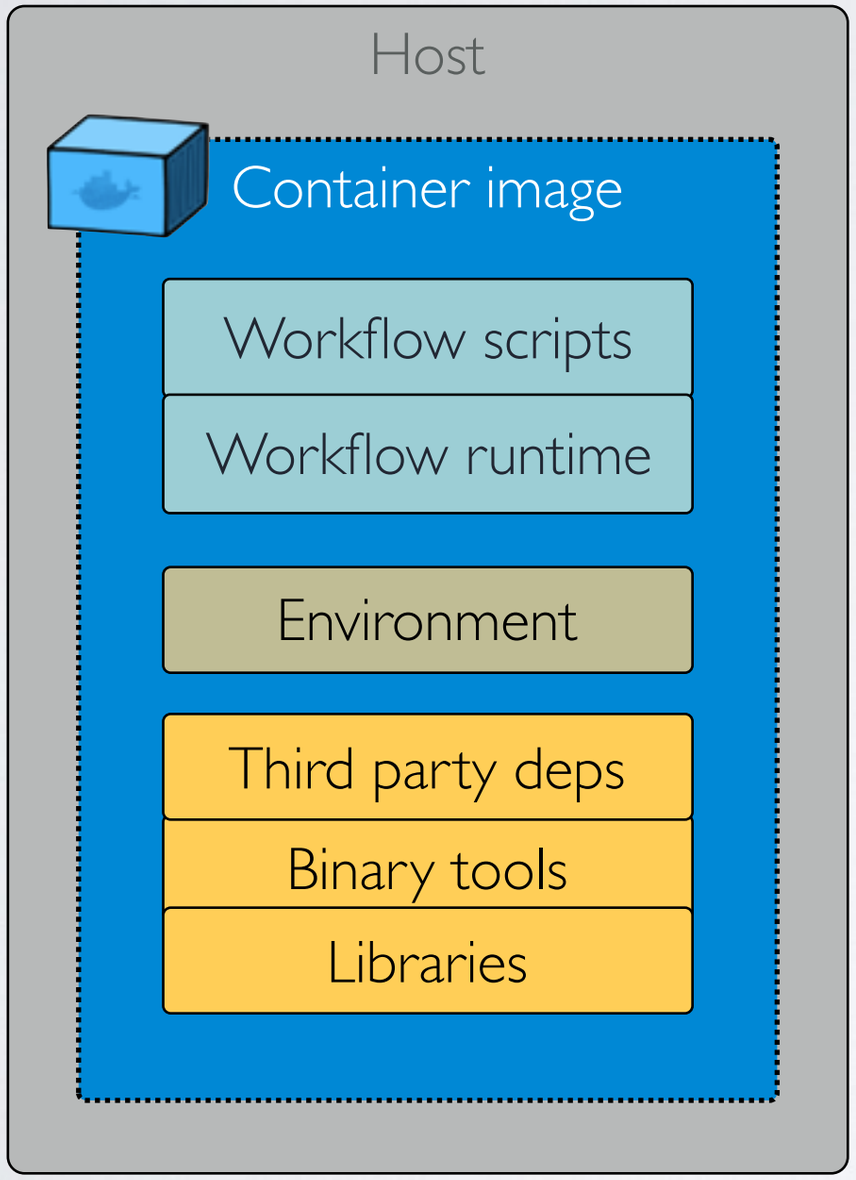
CONTAINERS

Containers are emerging as a solution to the problem of reproducibility of scientific workflows

- 100~ scientific publications mentioning Docker
- Large adoption in bioinformatics:
  - DockStore
  - BioContainers
  - BioShadock
  - Bioboxes



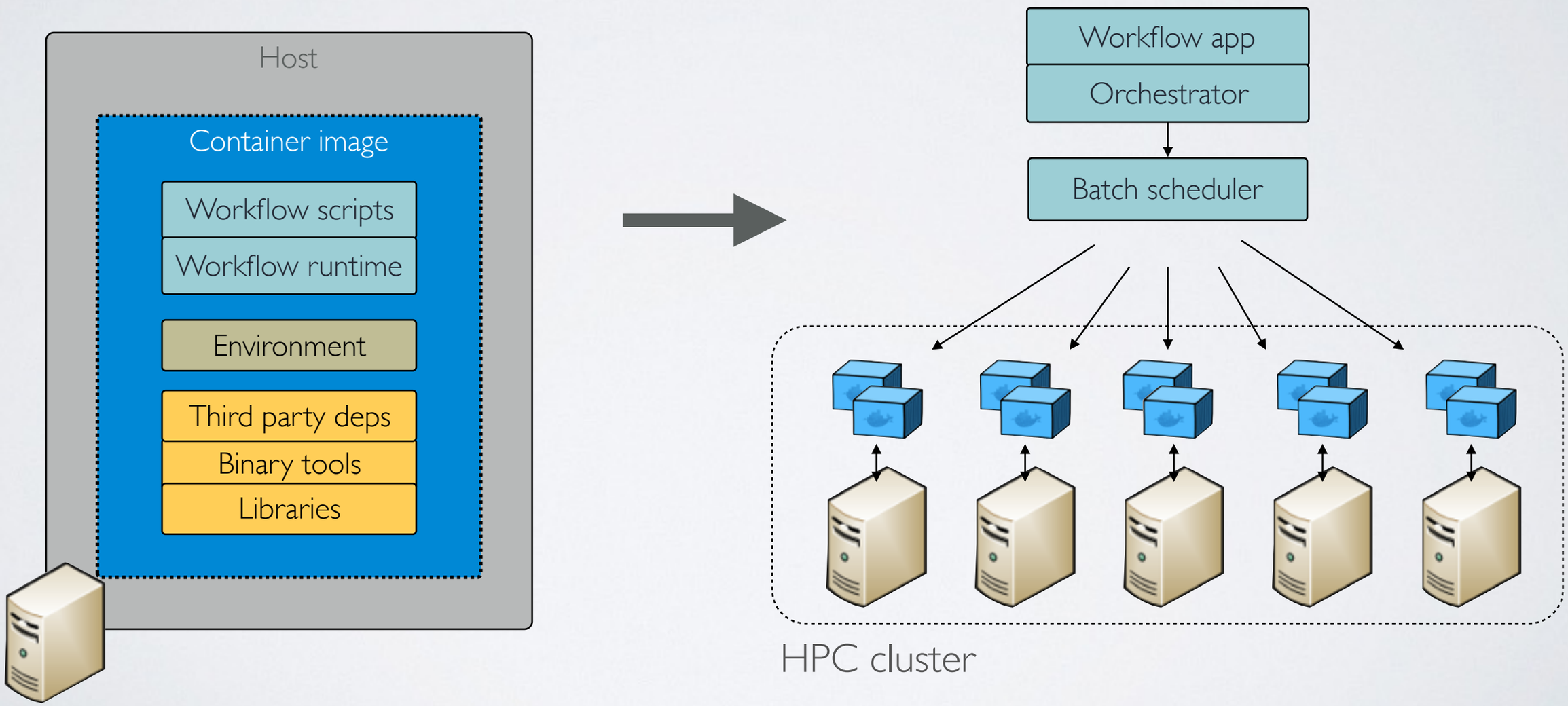
CONTAINER ISOLATION



- Allows you to create a ready-to-run package with all software dependencies
- Just one dependency instead of dozens
- Consistent results over time

HOW TO MANAGE A CONTAINERISED WORKLOAD AT SCALE?

CONTAINERISED WORKLOADS



ARE THE RIGHT TOOL FOR SCIENTIFIC WORKLOADS?

SERVICES ORCHESTRATION

$\neq$

JOBS SCHEDULING

CHALLENGES

- Isolate each task execution in its own container
- Manage jobs scheduling and dependencies
- Allow user to use any existing tools and scripts
- Automatic errors failover & execution checkpoints
- Enable portability across platforms (HPC and cloud)

Orchestration
& Parallelisation

nextflow

containers

Scalability
& Portability

Deployment &
Reproducibility

UNIVA GRID ENGINE slurm amazon web services

Platform Computing an IBM Company apache Ignite kubernetes BSC

Git GitHub GitLab Bitbucket

TASK EXAMPLE

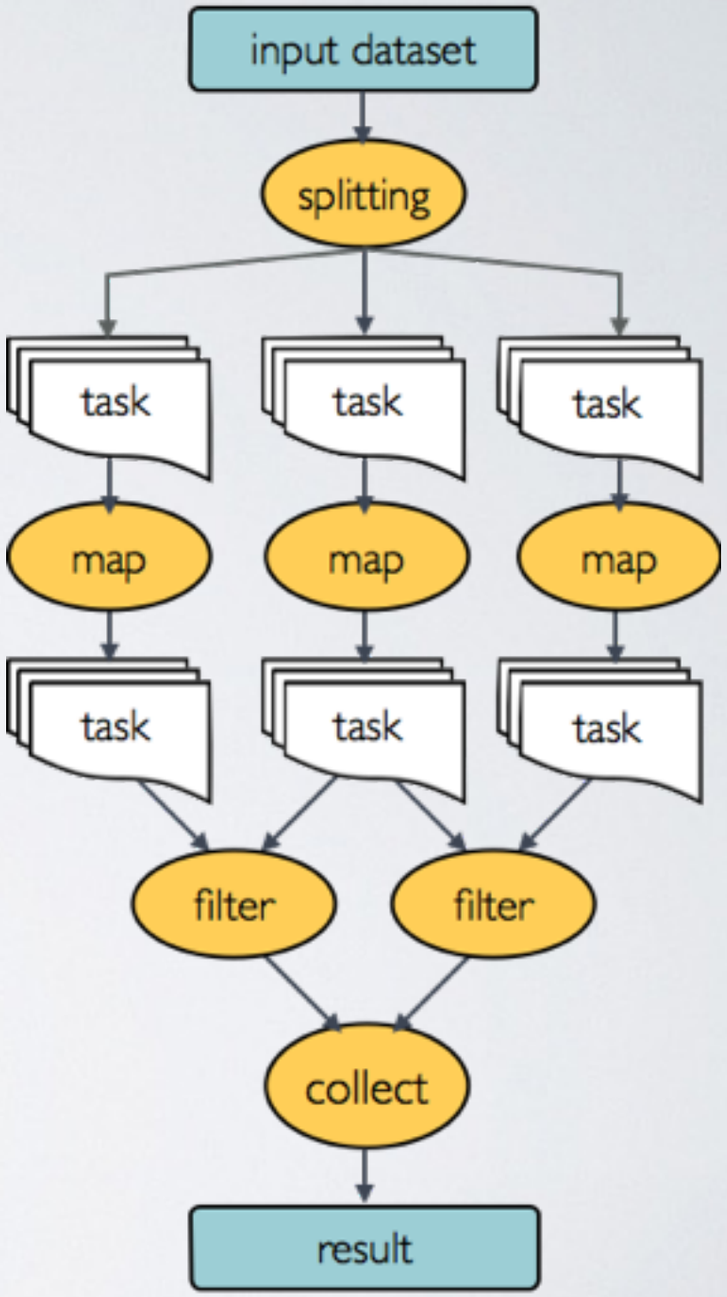
```
#!/bin/bash
blastp -query sampla.fasta -outfmt 6 \
    | head -n 10 \
    | cut -f 2 \
    | blastdbcmd -entry - > sequences.txt
```

TASK EXAMPLE

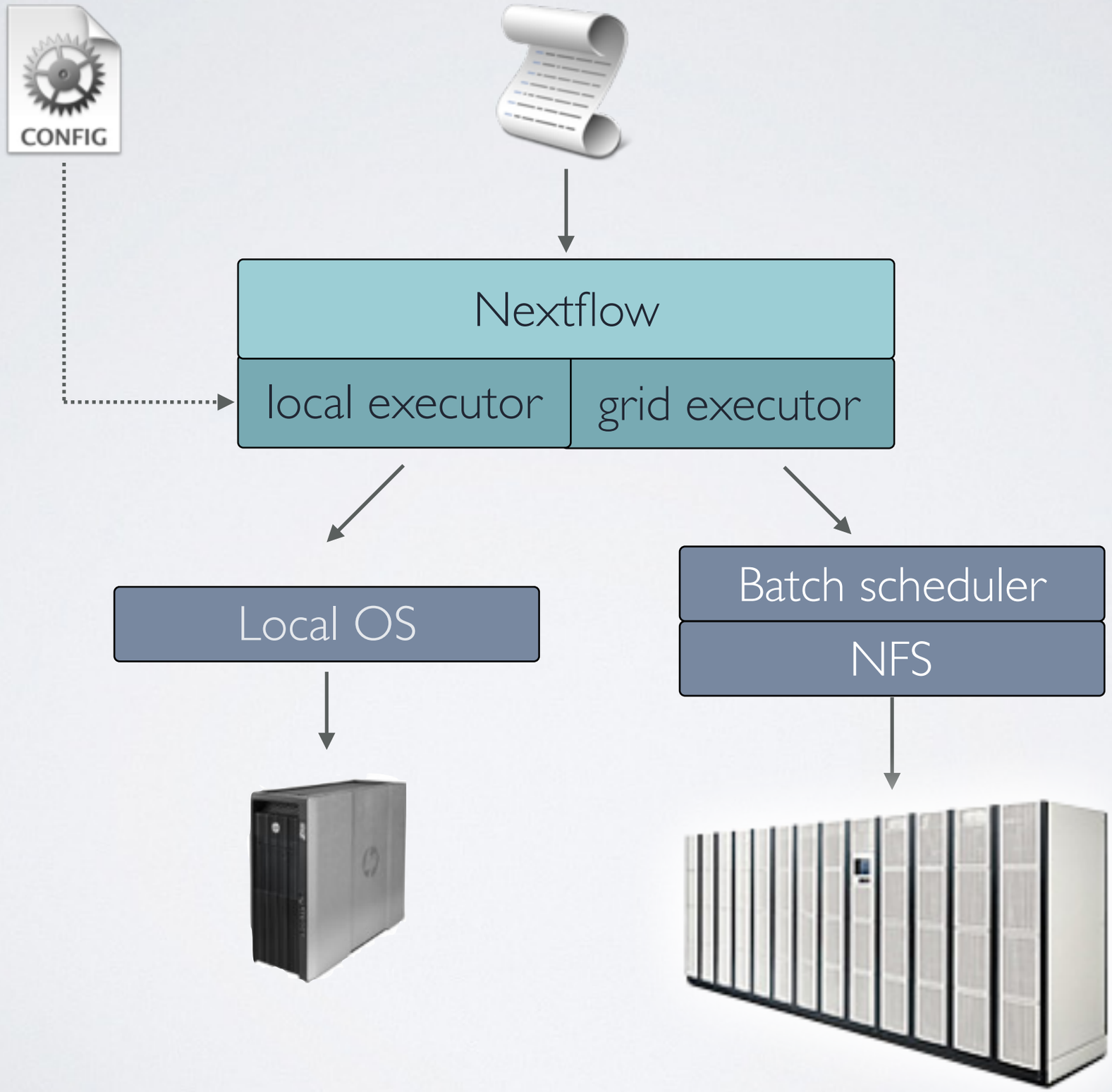
```
process foo {

    input:
    file 'sample.fasta' from fasta_files

    output:
    file 'sequences.txt' into result_files

    script:
    """
    #!/bin/bash

    blastp -query sampla.fasta -outfmt 6 \
        | head -n 10 \
        | cut -f 2 \
        | blastdbcmd -entry - > sequences.txt
    """
}
```

REACTIVE NETWORK

- Declarative computational model for parallel process executions
- Processes wait for data, when an input set is ready the process is executed
- They communicate by using dataflow variables i.e. async FIFO queues called channels
- Parallelisation and tasks dependencies are implicitly defined by process in/out declarations



PORTABILITY



UNIVA

slurm
workload manager

Platform Computing
an IBM Company

PBS Works

HTCondor
High Throughput Computing

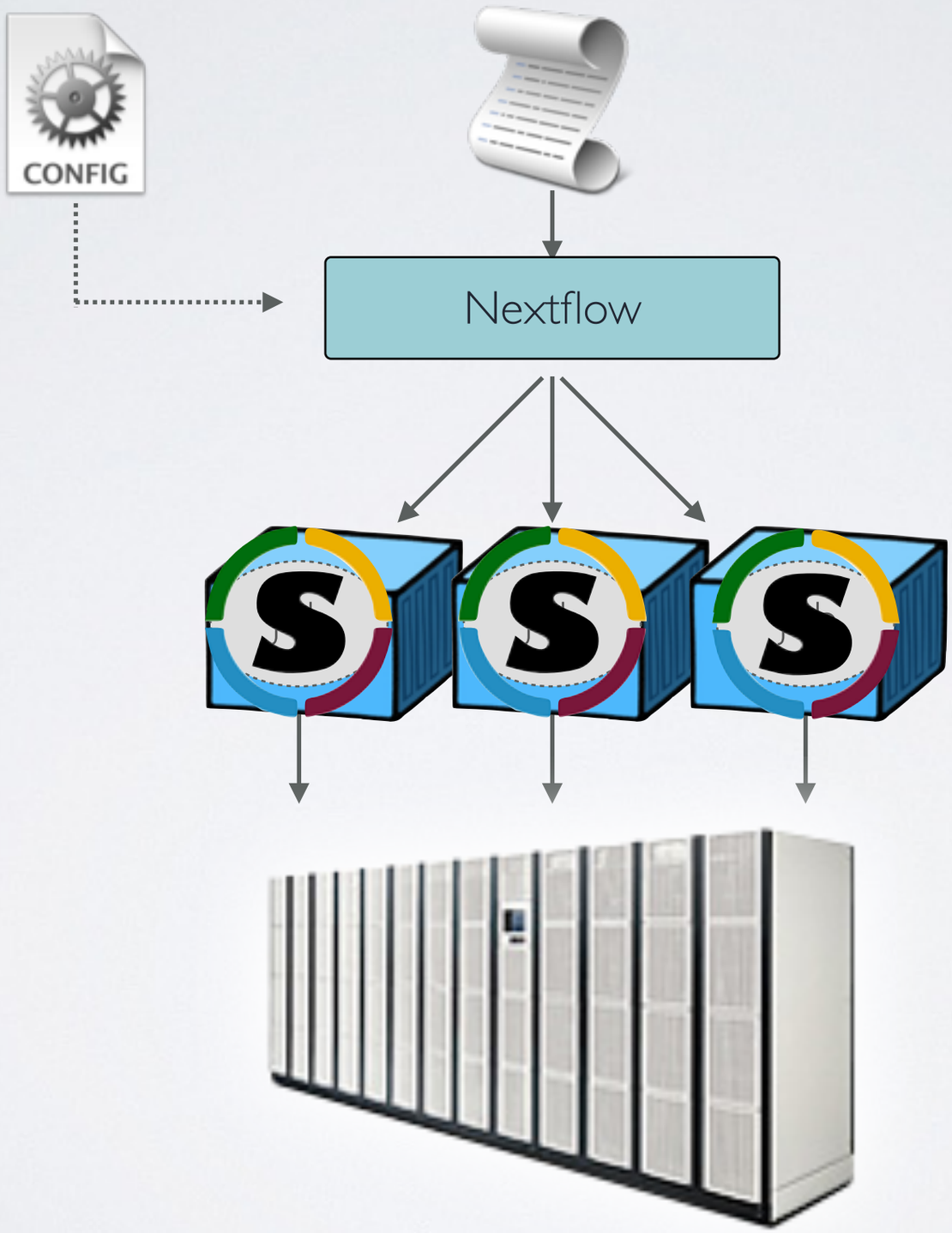
kubernetes

amazon web services

CONFIGURATION FILE

```
process {
  executor = 'sge'
  queue = 'cn-el6'
  memory = '10GB'
  cpus = 8
  time = '2h'
  container = 'ncbi/blast:3.2'
}
```

CONTAINERISATION



BENEFITS

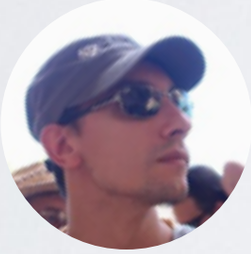
- Dead easy deployment
- Precise control on the execution runtime
- Portable across different execution platforms
- Decouple application logic from infra/configuration
- Enable reproducibility across systems and over time

WHO IS USING NEXTFLOW?

ACKNOWLEDGMENT

Notredame Lab, CRG

Evan Floden



Emilio Palumbo

Cedric Notredame

CRG Centre for Genomic Regulation

nextflow

http://nextflow.io