Scaling bioinformatics analysis using Nextflow and AWS

Francesco Strozzi
Bioinformatics Manager
NEXTFLOW ECOSYSTEM

Orchestration & Parallelisation

Scalability & Portability

Deployment & Reproducibility

Executors

Containers

NEXTFLOW

Git

GitHub

GitLab

Bitbucket
NEXTFLOW EXECUTORS

- slurm
- Kubernetes
- Grid Engine
- UNIVA
- Platform Computing
- Apache Ignite
- Google Cloud Platform
- Azure
- PBS Works
- HTCondor
AWS BATCH

- Automates dynamic computing clusters creation
- Based on Docker
- Automatically optimise the EC2 instance types used depending on jobs requirements
- Using the spot instances and with the new per second billing, it has a dramatic impact on the costs for data analysis
process {
  executor = 'slurm'
  queue = 'my-queue'
  memory = '8 GB'
  cpus = 4
  container = 'user/image'
}
process {
  executor = 'awsbatch'
  queue = 'my-queue'
  memory = '8 GB'
  cpus = 4
  container = 'user/image'
}

Nextflow automatically:
- Converts the tasks into AWS Batch jobs
- Transparently manages all the input and output data transfer with S3
OUR EXPERIENCE WITH NEXTFLOW & AWS BATCH

- Nextflow fully supports AWS Batch starting from version 0.26

- There is a minimal setup to do in AWS (an AMI with enough space and a computing environment)

- You can get the best of both Nextflow and the cloud
  - Scalability is a non-issue (apart for the bill on your account)
  - Reproducibility and composability

- In Enterome we use Nextflow with AWS Batch routinely for both development and production workflows
  - Using AWS Batch alone without an advanced orchestrator can be fairly complicate
  - We started the AWS Batch integration together with the Nextflow team in 2017

- Optimizations:
  - S3 is the main filesharing service used in this model, does not scale very well in the current implementation
  - Submitting thousands of jobs can take time due to the API requests (jobs array / concurrent requests will help)
Meet up and Discuss: CoFest on Friday

https://www.nextflow.io/

https://github.com/nextflow-io/nextflow

https://twitter.com/nextflowio

https://gitter.im/nextflow-io/nextflow

Paolo Di Tommaso – Nextflow creator
THANK YOU