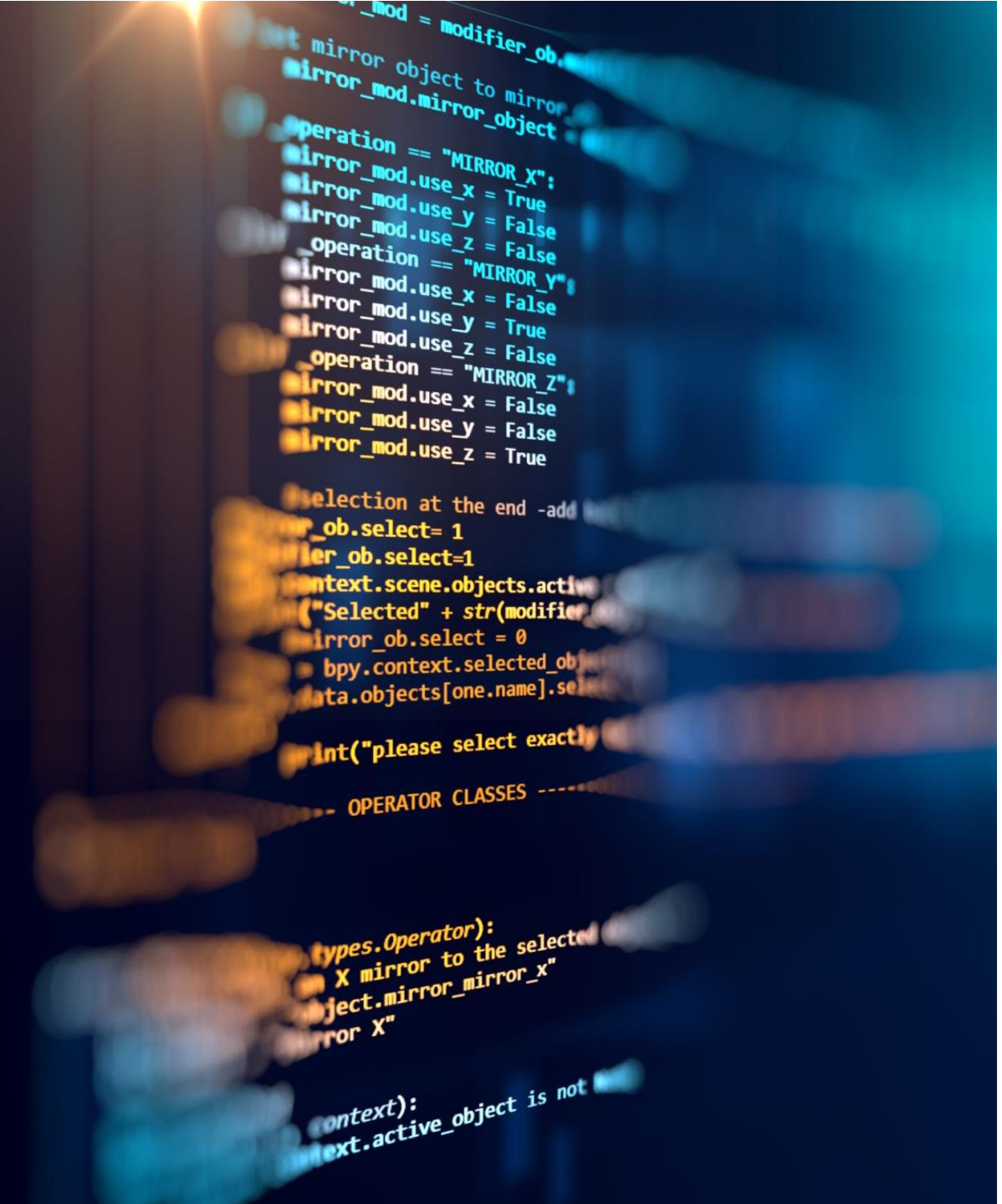


Nextflow 의 클라우드 컴퓨팅 활용법

1. Nextflow를 활용한 AWS-Batch활용법
2. Cloud computing 비용 절감 방법

발표자: 강승현

날짜: 26/AUG/2025

What is Nextflow

Nextflow is workflow manger

Wratten et al. (2021) *Nature Methods*

> *Workflow managers provide a framework for the creation, execution, and monitoring of pipeline.*
> <...>
>
> *They simplify pipeline development, optimize resource usage, handle software installation and versions, and run on different compute platforms, enabling workflow portability and sharing.*

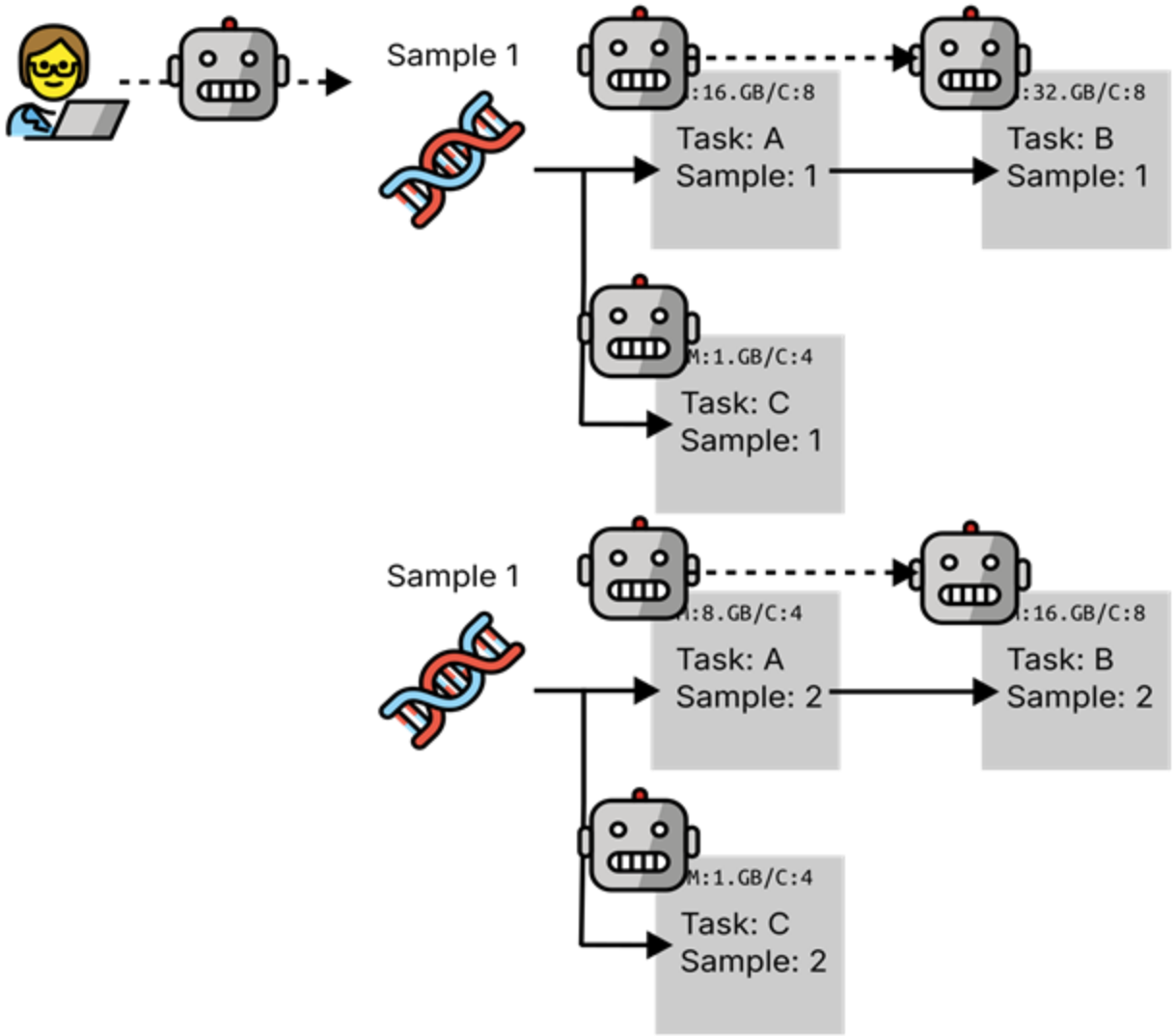
Pipelines using workflow managers benefit users by being

- More portable
  - Can be used on a very wide range of infrastructure
- More efficient
  - Less for users to install
  - More optimising resource usage
  - Extremely parallelised

Than manual analyses, or 'custom' pipeline frameworks

Nextflow available benefits to any user to reproduce the pipeline

Nextflow is workflow manger



nf-core/sarek

nf-core/scrnaseq

oncoanalyser

genomic-medicine-sweden/nallo is a bioinformatics analysis pipeline for long-reads

What if need to process pipeline with large data set

Nextflow is workflow manger

Many **schedulers** supported by Nextflow

slurm workload manager

IBM Spectrum LSF

grid

& more

```
$ nextflow run nf-core/rnaseq -profile sge <...>
$ nextflow run nf-core/rnaseq -profile slurm,raven <...>
$ nextflow run nf-core/rnaseq -profile binac <...>
$ nextflow run nf-core/rnaseq -profile awsbatch <...>
```

Many institutional clusters/HPC already supported by

nf-core/configs

as well as the **cloud**

aws

However, large consortium data analysis require cloud computing

PCAWG PanCancer Analysis OF WHOLE GENOMES

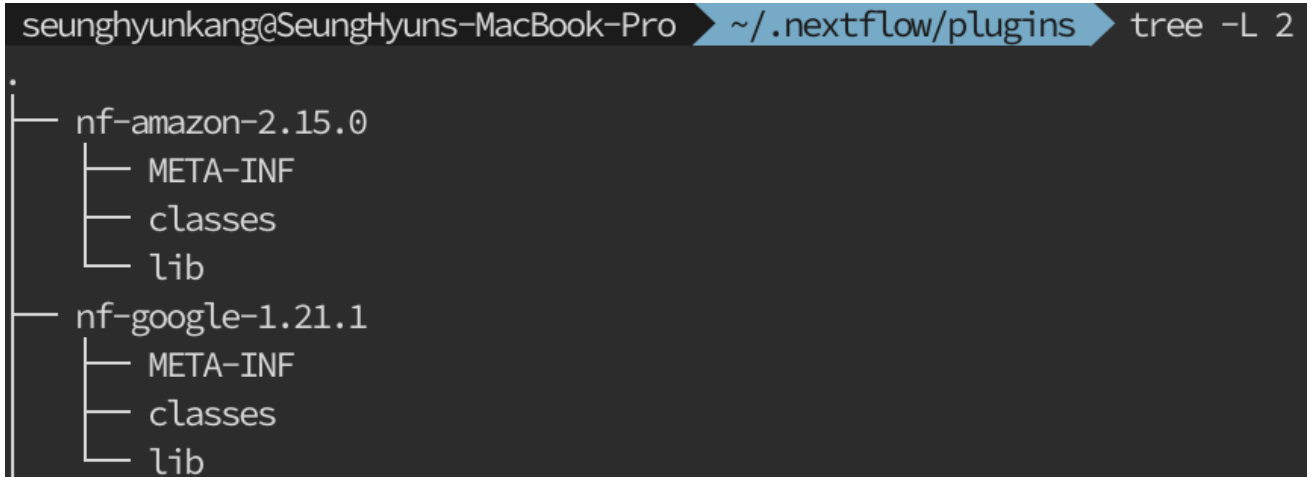
TCGA

Hartwig MEDICAL FOUNDATION

Nextflow have lots of plugin options

Nextflow plugins

Plugin ID	Functions
nf-amazon	**AWS Batch Executor 및 S3 파일 시스템 지원**
nf-azure	Microsoft Azure Batch 및 Azure Storage 지원
nf-google	Google Cloud Life Sciences 및 Google Cloud Storage 지원
nf-tower	**Seqera Platform (구 Tower) 연동 및 모니터링**
nf-wave	Wave Containers 서비스 연동
nf-schema	파이프라인 파라미터의 유효성을 검사하고, --help 출력을 자동 생성
nf-prov	파이프라인 실행 과정에 대한 상세한 출처(provenance) 정보를 캡처
nf-sqldb	파이프라인 내에서 SQL 데이터베이스에 읽고 쓰는 기능 추가



Now, Lets set up aws-batch setting with nextflow

Nextflow plugins

```
// nextflow.config

process {
    // 1. 실행 엔진을 'awsbatch'로 지정
    executor = 'awsbatch'

    // 2. 각 프로세스에 할당할 기본 CPU 및 메모리
    cpus = 2
    memory = '8.GB'

    // 3. 프로세스가 실행될 AWS Batch Job Queue 이름
    queue = 'your-batch-queue-name'
}

aws {
    // 4. AWS 리전 설정
    region = 'ap-northeast-2' // 예: 서울 리전
    batch {
        // 5. 컨테이너 내에서 사용할 AWS CLI 경로
        cliPath = '/usr/local/bin/aws'
    }
}

// 6. 작업 디렉토리를 S3 버킷으로 지정
workDir = 's3://your-nextflow-work-bucket/work'

// 7. Docker 컨테이너 사용 활성화
docker.enabled = true
```

```
nextflow run nf-core/rnaseq -profile test,awsbatch -resume
```

Nextflow plugin automatically installed when process executor require plugins

However, you can install plugin by below commands

Prerequisites: what you need to start

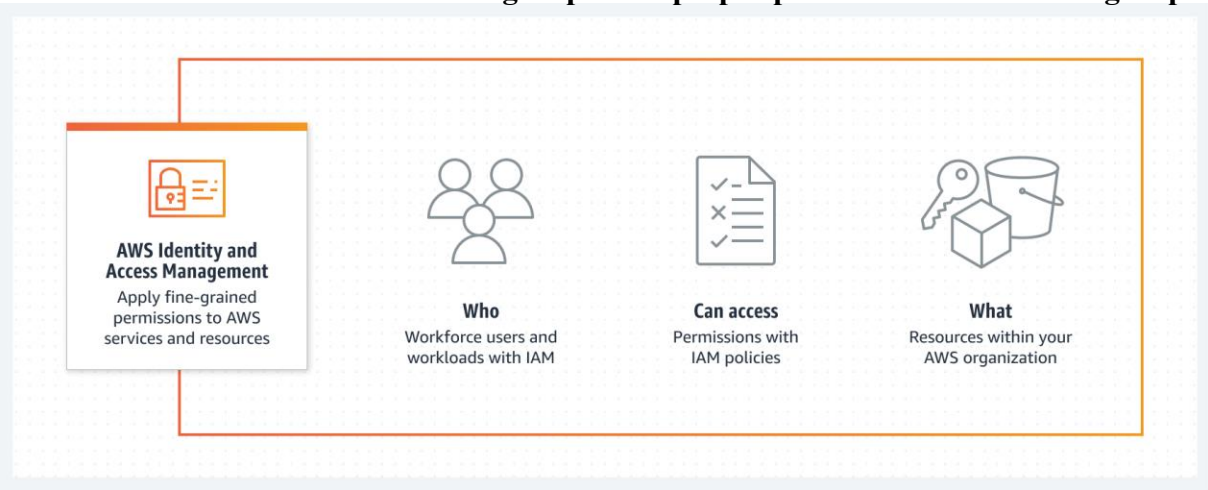
- Nextflow
- Docker account
- AWS
  - Account
  - IAM setting
  - VPC setting
  - Compute Environment setting

How to Set Up and Run Nextflow with AWS Batch

AWS setting: IAM

* For AWS root manger:

We recommand make user and user-group to set proper permision for each user group



- To pull access key from other platform such as CGC

```
                "ssm:PutParameter",
                "ssm:LabelParameterVersion",
                "ssm:DeleteParameter",
                "ssm:UnlabelParameterVersion",
                "ssm:GetParameterHistory",
                "ssm:GetParameters",
                "ssm:GetParameter",
                "ssm:DeleteParameters"
            ],
            "Resource": "arn:aws:ssm:*:*:parameter/*"
        },
        {
            "Sid": "VisualEditor1",
            "Effect": "Allow",
            "Action": "ssm:DescribeParameters",
            "Resource": "*"
        }
    ]
```

- To use AWS Batch:

```
"batch:CancelJob"
"batch:DescribeComputeEnvironments"
"batch:DescribeJobDefinitions"
"batch:DescribeJobQueues"
"batch:DescribeJobs"
"batch:ListJobs"
"batch:RegisterJobDefinition"
"batch:SubmitJob"
"batch:TagResource"
"batch:TerminateJob"
```

- To view EC2 instances:

```
"ec2:DescribeInstanceAttribute"
"ec2:DescribeInstances"
"ec2:DescribeInstanceStatus"
"ec2:DescribeInstanceTypes"
"ecs:DescribeContainerInstances"
"ecs:DescribeTasks"
```

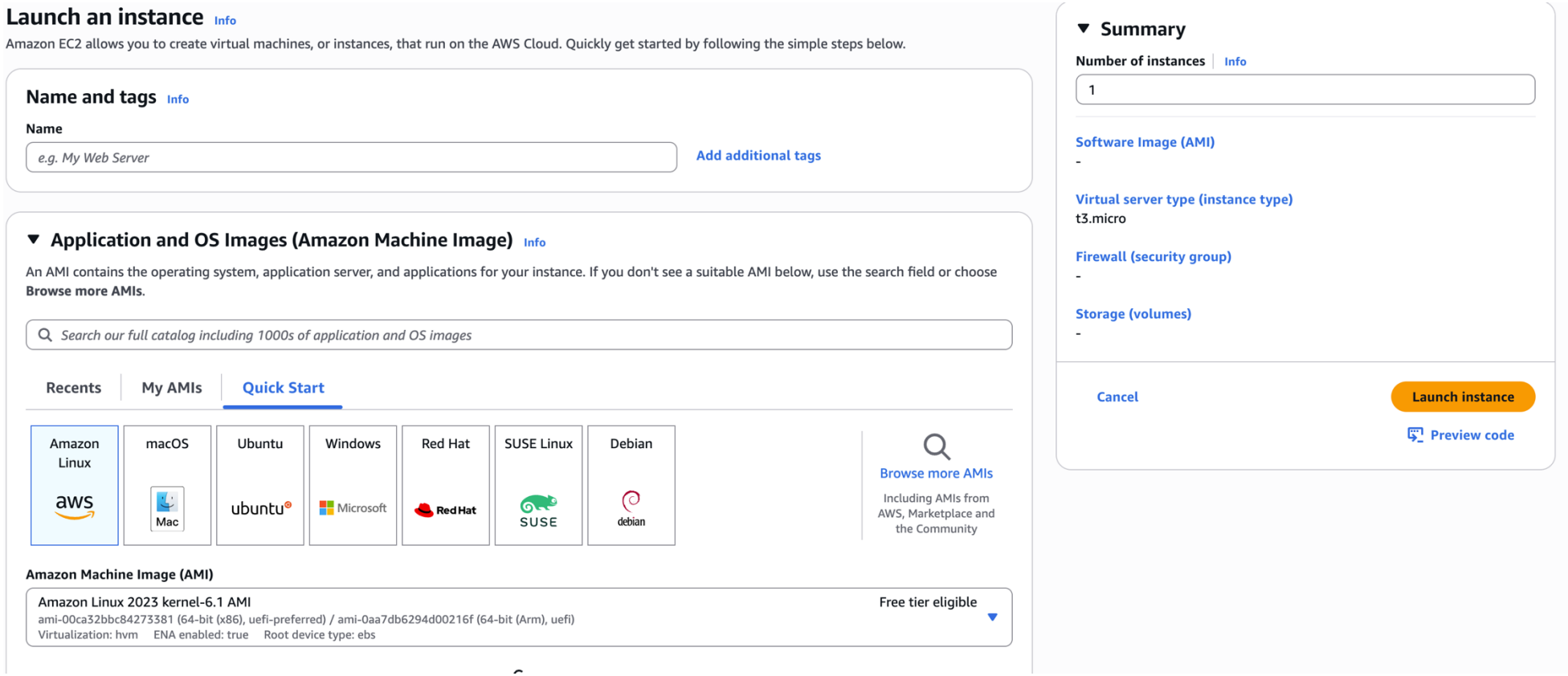
- To pull container images from ECR repositories:

```
"ecr:BatchCheckLayerAvailability"
"ecr:BatchGetImage"
"ecr:DescribeImages"
"ecr:DescribeImageScanFindings"
"ecr:DescribeRepositories"
"ecr:GetAuthorizationToken"
"ecr:GetDownloadUrlForLayer"
"ecr:GetLifecyclePolicy"
"ecr:GetLifecyclePolicyPreview"
"ecr:GetRepositoryPolicy"
"ecr:ListImages"
"ecr:ListTagsForResource"
```

AWS setting: VPC

- Set VPC according to your server and other users network environment

AWS setting: Create snapshot image



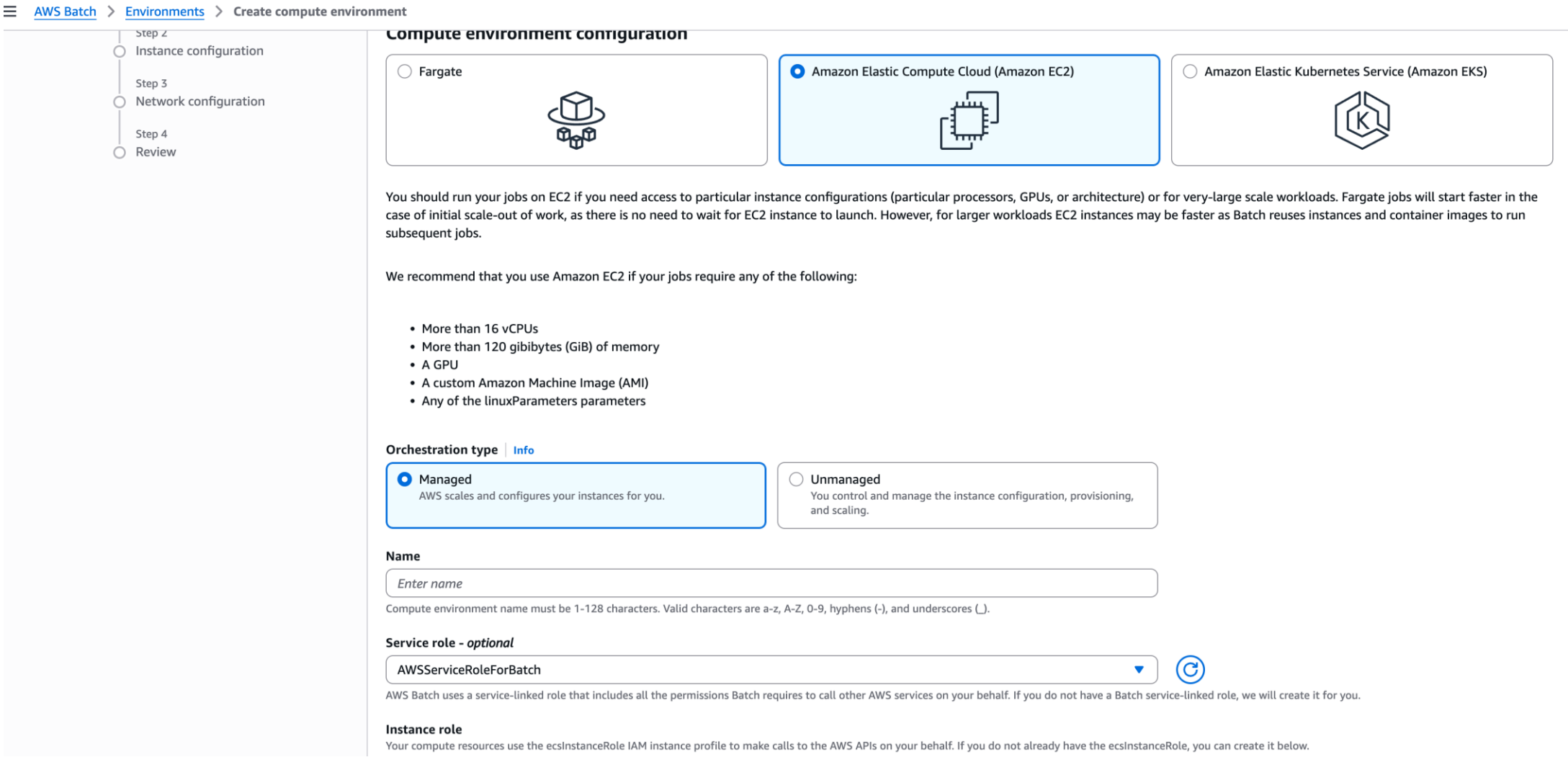
Amazon ECS-Optimized Amazon Linux 2 (AL2) x86_64 AMI:

- Docker already installed, but you need to download aws

Amazon ECS-optimized Amazon Linux 2023 AMI

- You need to install docker, but aws cli exsit

AWS setting: Create compute environment



AWS setting: Create queue

Orchestration type

- ○ Fargate
- ◉ Amazon Elastic Compute Cloud (Amazon EC2)
- ○ Amazon Elastic Kubernetes Service (Amazon EKS)
- ○ SageMaker Training

You should run your jobs on EC2 if you need access to particular instance configurations (particular processors, GPUs, or architecture) or for very-large scale workloads. Fargate jobs will start faster in the case of initial scale-out of work, as there is no need to wait for EC2 instance to launch. However, for larger workloads EC2 instances may be faster as Batch reuses instances and container images to run subsequent jobs.

We recommend that you use Amazon EC2 if your jobs require any of the following:

- More than 16 vCPUs
- More than 120 gibibytes (GiB) of memory
- A GPU
- A custom Amazon Machine Image (AMI)
- Any of the linuxParameters parameters

Job queue configuration

Name

Enter job queue name

Job queue name must be 1-128 characters. Valid characters are a-z, A-Z, 0-9, hyphens (-), and underscores (_).

Priority | Info

How to Set Up and Run Nextflow with AWS Batch

Nextflow setting

For Today, we will use oncoanalyser to practice

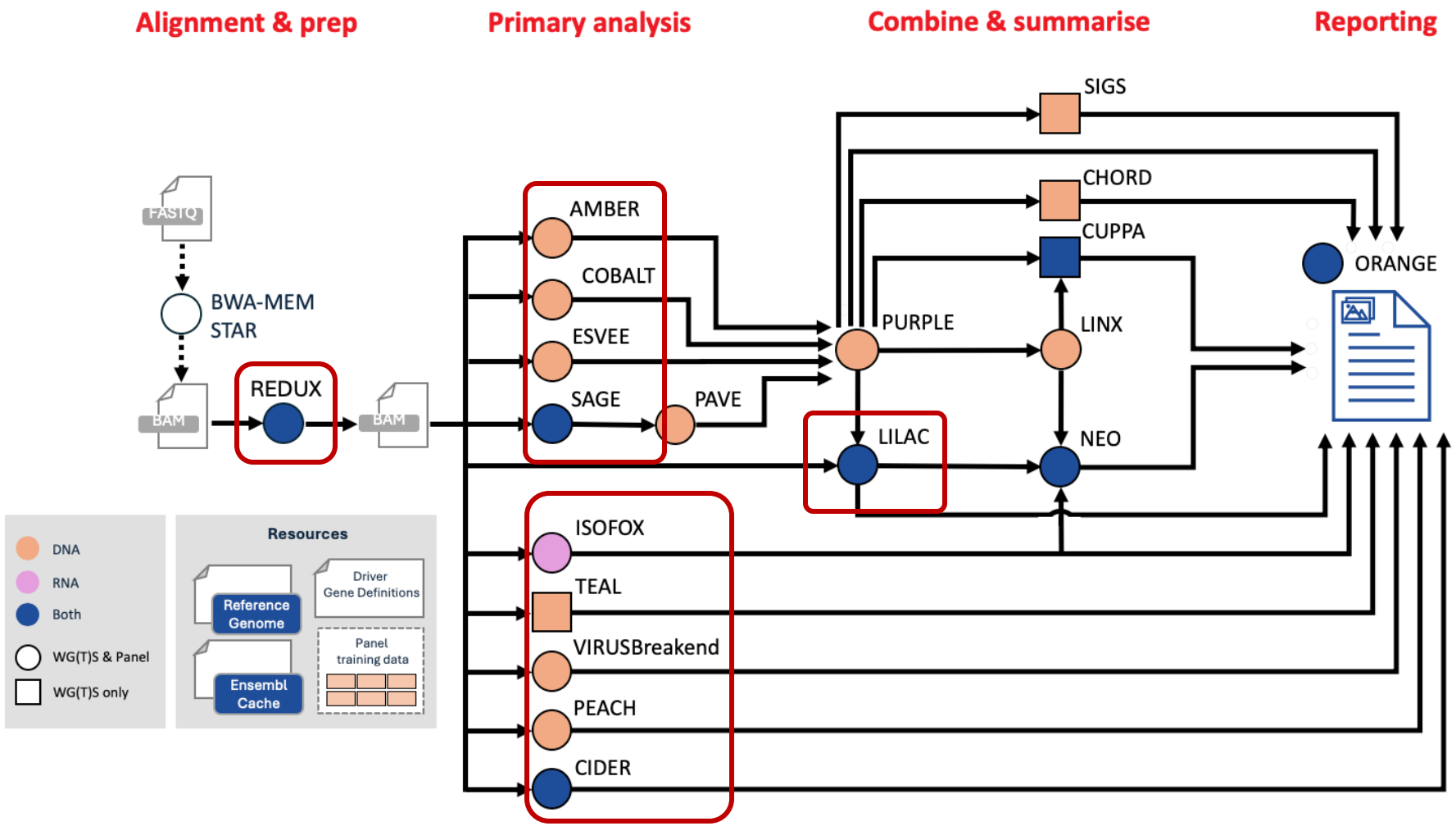
```
//Nextflow config file for running on AWS batch
params {
    config_profile_description = 'AWSBATCH Cloud Profile'
    config_profile_contact     = 'Alexander Peltzer (@apeltzer)'
    config_profile_url         = 'https://aws.amazon.com/batch/'

    awsqueue                   = false
    awsregion                  = 'eu-west-1'
    awscli                     = '/home/ec2-user/miniconda/bin/aws'
}

timeline {
    overwrite = true
}
report {
    overwrite = true
}
trace {
    overwrite = true
}
dag {
    overwrite = true
}

process.executor = 'awsbatch'
process.queue = params.awsqueue
aws.region = params.awsregion
aws.batch.cliPath = params.awscli
```

```
seunghyunkang@SeungHyuns-MacBook-Pro ~/Dropbox/oncoanalyser master nextflow run main.nf -c conf/awsbatch.config -bucket-dir s3://steve-oncoanalyser-20250825 -profile test,my_aws,docker -resume
```

How to reduce cloud cost ?

Where the large cloud computing cost comes from ?

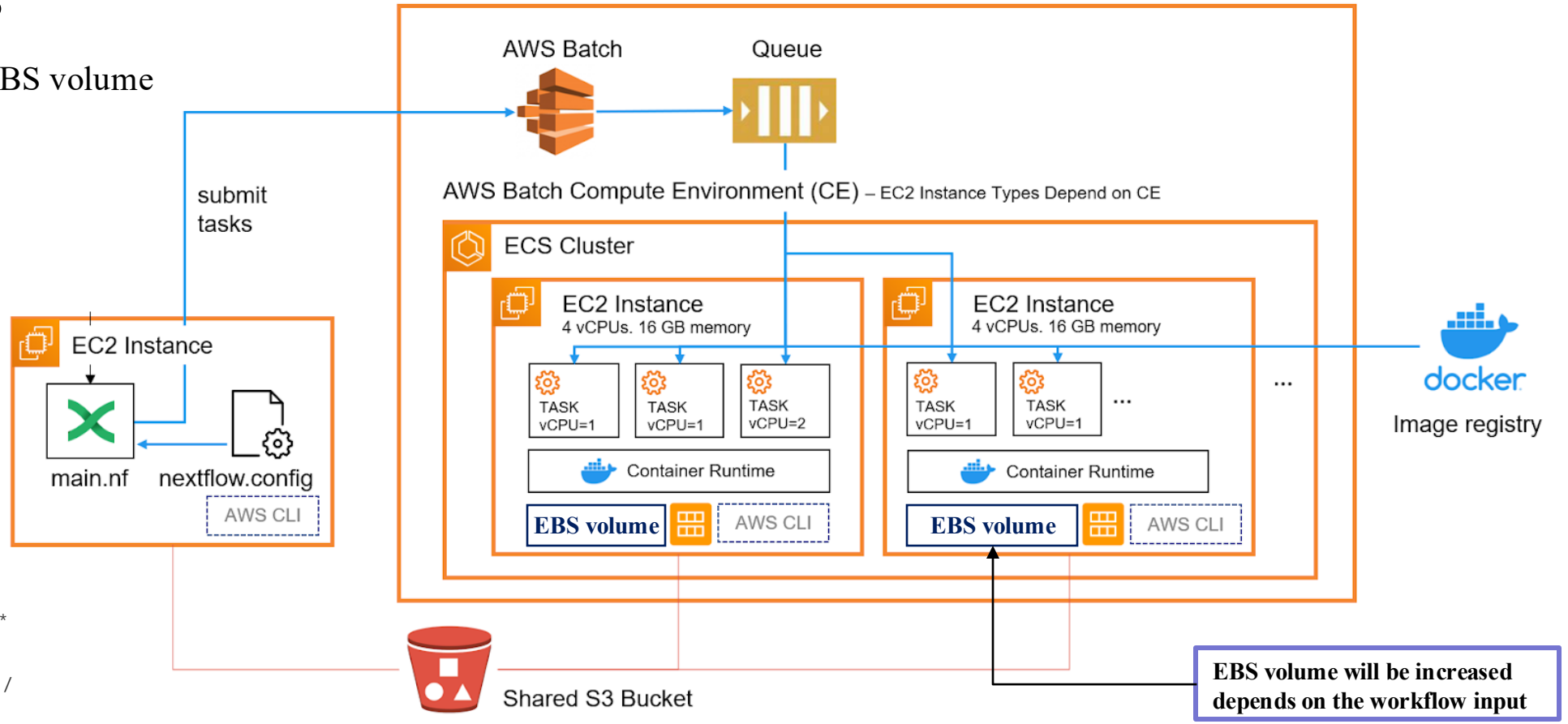
Here is oncoanalyser pipeline, there are 11 modules (there are more because some modules have multiple process) that require bam files



How to reduce cloud cost ?

Why need larger EBS?

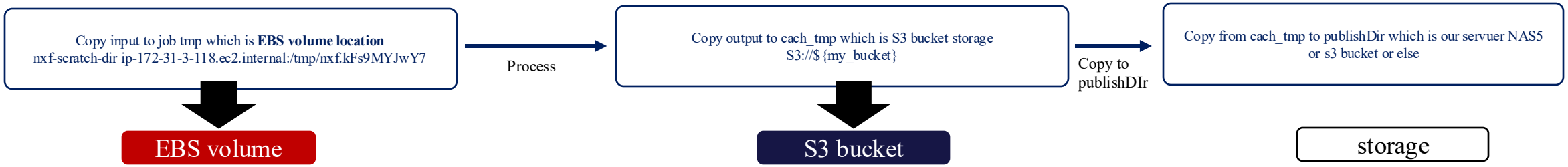
Large size inputs need larger EBS volume



0.08 * provision storage * times (sec) / (86,400 sec * 30)) = price

(월별 GB당 0.08 USD * 650 GB * 10,800 sec (= 3hr) / (86,400 * 30) = 0.216 USD per instances (process)

+ compute ec2 price xxx USD per hour



How to reduce cloud cost ?

Then how to reduce those cost?

Large size inputs need larger EBS volume

- Try to USE NVMe instance which contained ready to use storage (do not require EBS) - it only costs cpu usage times per hr
- Mount system
  - ✓ Mount-S3, FUSE, … etc
- Elastic File System (EFS)