



## Complete Dynamic Multi-cloud Application Management

Project no. 644925

Innovation Action

Co-funded by the Horizon 2020 Framework Programme of the European Union



Call identifier: H2020-ICT-2014-1

Topic: ICT-07-2014 – Advanced Cloud Infrastructures and Services

Start date of project: January 1st, 2015 (36 months duration)

## Deliverable D3.5

### Overview of Implemented Use Cases

**Due date:** 31/12/2017  
**Submission date:** 30/01/2018  
**Deliverable leader:** CNRS  
**Editors list:** C. Blanchet (CNRS)

#### Dissemination Level

- |                                     |   |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | PU: Public  |
| <input type="checkbox"/>            | PP: Restricted to other programme participants (including the Commission Services)        |
| <input type="checkbox"/>            | RE: Restricted to a group specified by the consortium (including the Commission Services) |
| <input type="checkbox"/>            | CO: Confidential, only for members of the consortium (including the Commission Services)  |

## List of Contributors

Participant	Short Name	Contributor
Interoute S.P.A.	IRT	Domenico Gallico, Matteo Biancani
SixSq Sàrl	SIXSQ	Charles Loomis
QSC AG	QSC	Ralf Fischer, Adelheid Weinert, Maria Kourkouli, Ulrich Hacker, Doris Hacker
Technische Universitaet Berlin	TUB	Dirk Thatmann
Fundacio Privada I2CAT, Internet I Innovacio Digital A Catalunya	I2CAT	Eduard Escalona
Centre National De La Recherche Scientifique	CNRS	Christophe Blanchet
Universiteit van Amsterdam	UvA	Fatih Turkmen, Yuri Demchenko

## Change history

Version	Date	Partners	Description/Comments
0.1	21/12/2017	CNRS, QSC, i2CAT, SiXSQ, TUB, UvA	Initial version.
0.2	29/01/2018	CNRS, QSC, i2CAT, SiXSQ, UVA	Revised version, ready for internal review

# Table of Contents

<b>List of Contributors .....</b>	<b>2</b>
<b>Change history .....</b>	<b>3</b>
<b>List of Figures .....</b>	<b>5</b>
<b>List of Tables .....</b>	<b>6</b>
<b>Executive Summary .....</b>	<b>7</b>
<b>1. Introduction .....</b>	<b>8</b>
<b>2. CYCLONE implemented use cases .....</b>	<b>9</b>
2.1.1. UC1 - Securing human biomedical data .....	10
2.1.2. UC2 - Cloud virtual pipeline for microbial genomes analysis .....	11
2.1.3. UC3 - Live remote cloud processing of sequencing data .....	11
2.1.4. UC4 - Virtual Power Plant .....	12
2.1.5. UC5 - Internet of Services Lab (IoSL) .....	12
2.1.6. UC6 - Multi-cloud deployment and federated identity integration of ENtRANCe components .....	13
2.1.7. UC7 - Open Scientific Data .....	13
2.1.8. UC8 - Benchmark Driven Placement .....	14
2.1.9. UC10 - Smart Utility 4.0 .....	14
2.1.10. UC11 - Assembling genomes from sequencing reads .....	15
2.1.11. UC12 - Metagenomics .....	15
2.1.12. UC13 - Shared environment between cloud Galaxy portals .....	16
2.1.13. UC16 - Attribute-based Authorizations with XACML .....	16
<b>3. Use cases related software developments .....</b>	<b>19</b>
3.1. <i>CYCLONE Key technical areas .....</i>	<i>19</i>
3.2. <i>Software implications according to the use cases requirements .....</i>	<i>19</i>
<b>4. Conclusions .....</b>	<b>21</b>
<b>References .....</b>	<b>21</b>
<b>5. Acronyms .....</b>	<b>23</b>

---

# List of Figures

Figure 1: DACI Single Deployment Parameters ..... 17

---

# List of Tables

Table 1: List of CYCLONE use cases.....9

Table 2: Summary of use cases integration with the different CYCLONE components ..... 10

Table 3: CYCLONE key technical areas and associated software developments ..... 19

Table 4: Details of achieved use cases requirements and related CYCLONE software implications ..... 20

---

## Executive Summary

The goal of this document is to provide an overview of all CYCLONE use cases and their achieved integration at the end of the project. The use cases are available from the CYCLONE main site and use case devoted portal, from the CYCLONE SlipStream/Nuvla instance, and, in some cases, from the corresponding community cloud portal. Potential users will find there a detailed description as well as demonstrators ready-to-run on the project testbed and usual cloud environments (public and academic) like Exoscale or AWS.

CYCLONE developments strongly contribute to the adaptation of use cases to multi-cloud environment:

- SlipStream/Nuvla for easiness of integration and one-click multi-cloud deployment of complex applications
- Federation Provider for authentication and authorization management based on eduGAIN federation identity
- DACI for the authorizations management and the enforcement of policies
- CNSMO for secure (VPN, FW) multi-cloud deployment with high-level features (DNS, LB)

All of the use cases identified during the project have been integrated with CYCLONE software to achieve multi-cloud deployments, and most of them are available for usage.

---

# 1. Introduction

CYCLONE integrates and extends open source software to create a unified cloud application management solution for application service providers, DevOps, and researchers. One of the main added values of the CYCLONE developments is the ability to deploy legacy and new applications on multi-cloud infrastructures and to meet related needs such as deployments of several services of a complex application in many virtual machines for scalability or availability, identity and authorizations management, or network filtering and isolation.**Errore. Il nome file non è valido.**

During the project, the partners have elaborated the use cases: initial ones presented in the DOA about bioinformatics and energy applications, and new ones selected to complete this portfolio and to help to identify new requirements. Based on these requirements, CYCLONE partners have defined the software implications related to the CYCLONE key technical areas, focusing on the multi-cloud perspective, and developed new software components to meet the needs until the final evaluation of all CYCLONE use cases. This work was described in deliverables D3.1 [1], D3.2 [2] and D3.3 [3] and D3.4 [4].

This deliverable is organized as follows:

In chapter 2, we describe the use cases implemented during the project. For each use case, we provide first an overview and the goals having motivated the adaptation to the CYCLONE multi-cloud environment, and then their achieved integration at the end of the project.

In Chapter 3, we describe the link between these use cases and the related software development done by CYCLONE.

In Chapter 4, we provides the conclusion for this deliverable and for the activities carried on across WP3.



## 2. CYCLONE implemented use cases

CYCLONE use cases come from several domains, *e.g.* bioinformatics, energy, teaching, security. Table 1 gives an overview of the CYCLONE use cases, their domains and the partner who is the principal investigator. We also assigned previously a priority of treatment. For detailed descriptions, the reader should refer to the CYCLONE use case portal [6]. It is important to notice that some of the use cases rely on legacy applications, especially in the bioinformatics fields, where the proposed applications were previously deployed on local computing clusters or web portal.

**Table 1: List of CYCLONE use cases**

ID	Title	Domain	Resp.	Priority
UC1	Securing human biomedical data	Bioinformatics	CNRS	high
UC2	Cloud virtual pipeline for microbial genomes analysis	Bioinformatics	CNRS	high
UC3	Live remote cloud processing of sequencing data	Bioinformatics	CNRS	high
UC4	Virtual Power Plant	Energy	QSC	high
UC5	Internet of Services Lab (IoSL)	Teaching	TUB	high
UC6	ENTRANCE	App mgmt.	TUB	medium
UC7	Open Scientific Data	Data management (Earth Observation)	SixSq	medium
UC8	Benchmark Driven Placement	Bioinformatics	SixSq	high
UC10	Smart Utility 4.0	Energy	QSC	high
UC11	Assembling genomes from sequencing reads	Bioinformatics	CNRS	high
UC12	Metagenomics	Bioinformatics	CNRS	medium
UC13	Shared environment between cloud Galaxy portals	Bioinformatics	CNRS	high
UC16	Attribute-based Authorizations with XACML	Security	UvA	high
UC17	Genomic Variant Analysis for Cancer and Rare Disease Diagnosis	Bioinformatics	CNRS	high

\* UC9, UC14 and UC15 were interrupted or closed due to minor re-scoping of some activities of the project related to the components developments and the target audience.

During the development of the project, the use cases helped to defined 48 common requirements. They were attributed a level of relevance according to the following scale: MAY be satisfied (7 requirements), SHOULD (15) and MUST be (26). At the end of the project, these requirements are mostly fulfilled with the new multi-cloud components developed by CYCLONE partners: 38 are satisfied, 4 partially and 6 were finally not implemented. Table 2 provides a synthesis of the use case achievements (see Deliverable D3.3 [3] for details).

**Table 2: Summary of use cases integration with the different CYCLONE components**

ID	Title	Deployment	Security	Network
UC1	Securing human biomedical data	✓	✓	–
UC2	Cloud virtual pipeline for microbial genomes analysis	✓	✓	✓
UC3	Live remote cloud processing of sequencing data	✓	✓	✓
UC4	Virtual Power Plant	✓	-	✓
UC5	Internet of Services Lab (IoSL)	✓	✓	-
UC6	ENTRANCE	✓	✓	-
UC7	Open Scientific Data	✓	-	-
UC8	Benchmark Driven Placement	✓	-	-
UC10	Smart Utility 4.0	✓	✓	✓
UC11	Assembling genomes from sequencing reads	✓	✓	✓
UC12	Metagenomics	✓	✓	✓
UC13	Shared environment between cloud Galaxy portals	✓	✓	-
UC16	Attribute-based Authorizations with XACML	✓	✓	✓
UC17	Genomic Variant Analysis for Cancer and Rare Disease Diagnosis	✓	✓	✓

The following paragraphs provide a short description of the use cases and their achievements. When useful, related links to ready-to-go demonstrators, in the CYCLONE SlipStream/Nuvla instance [13] and the CNRS bioinformatics appliance catalogue [12], are provided below in the use case related paragraphs. The reader can refer to the CYCLONE use case portal [6] for the detailed use cases descriptions.

#### 2.1.1. UC1 - Securing human biomedical data

Thanks to the steady drop of genome sequencing technology costs (NGS), an increasing number of clinicians are including biological results obtained with these technologies in their day-to-day diagnosis practice. Today, much genomics analyses are realized on the exome, which is the expressed part (5%) of the genome. However, the full genome sequencing is being envisaged and will be soon included in daily medical practices.

In the near future, some of the genomic data processed on the IFB cloud platform will concern human biomedical data related to patients and thus, will be subject to strict privacy restrictions. To ensure the data security while carrying out the analysis in a federated cloud environment, it is necessary to ensure the security in all involved sites belonging to the federation and ensure their integration (especially if the cloud federation involves both public and private cloud infrastructures).

The cloud appliance NGS-Unicancer was developed by the platform INCa-SLC and provides a simple web interface to launch the biomedical genomic analysis pipeline. The appliance was enhanced by the Federation Provider developed by CYCLONE and is ready for on-demand deployment on the CNRS IFB's cloud infrastructure. The user deploys the appliance NGS-Unicancer through the CNRS IFB's web interface in "1-click" and uses the CYCLONE Federation Proxy to get access to the VM web interface based on its identity in the federation. Then the user can easily upload its data, run the analysis and get the results.

[https://nuv.la/module/cyclone/UC1\\_Securing\\_human\\_biomedical\\_data](https://nuv.la/module/cyclone/UC1_Securing_human_biomedical_data)

### 2.1.2. UC2 - Cloud virtual pipeline for microbial genomes analysis

In the post-NGS area, sequencing bacterial genomes is very cheap (few hundreds €). Most of the time, users are no longer content to analyse a single genome; they want to compare large collections of related genomes (strains). This entails that biologists have to pay too much attention and dedicate their time to sequence the genomes, instead of thoroughly analysing the genomic data. Thus, this brings light to the increasing need for automating the annotation of bacterial genomes.

The Insyght environment performs the annotation of microbial genomes and the visualization of the synteny (local conservation of the gene order along the genomes). The platform requires set of bioinformatics tools to analyse the data and stores the results of the tools in a relational database, several public reference data collections and a web interface to allow the user to consult the results and perform the manual annotation. Installing the platform requires solid skills in system administration. Thus, performing the analysis of collections of genomes requires large computing resources that can be found with the distribution of the jobs over several computers, generally the computing nodes of a cluster.

[https://nuv.la/module/cyclone/Bacterial\\_Genomics](https://nuv.la/module/cyclone/Bacterial_Genomics)

<https://biosphere.france-bioinformatique.fr/catalogue/appliance/19>

This use case was adapted to multi-cloud deployments. A successful validation was done on 2 sites of CYCLONE testbed (CNRS LAL and Exoscale). With the CNSMO-VPN tool, the whole virtual infrastructure is isolated for security and operation purposes (for example to ease the exchanges of the biological data between the nodes) from the others cloud users and VMs. The CNSMO-DNS helps the deployment of the batch cluster mode. The multi-cloud deployment of the UC2 then benefits from the following CYCLONE components:

- SlipStream for the generic deployment recipe compliant with several flavors of clouds
- SlipStream for the one-click deployment of a complex application
- Federation proxy to manage authentication both for Web and CLI accesses
- CNSMO to isolate the whole deployment from other cloud users and VMs, and help the deployment of the batch computing cluster.

### 2.1.3. UC3 - Live remote cloud processing of sequencing data

Bioinformatics deals with the collection and efficient analysis of biological data, particularly genomic information from DNA sequencers. The terabytes of raw data, produced by the sequencers for each run, require significant computing resources for analysis that may not be available locally. Especially for small sequencing facilities, the analysis of the data is indeed the bottleneck. These sequencers are located at a dozen places in France, while the users are distributed throughout the country and possibly further afield via international collaborations. The researchers may then not have sufficient on premise computing resources to analyse the raw data and will transfer them again to a remote computing center.

The utilization of federated cloud computing resources helps online data analysis: reducing the data transfers, reducing the need for long-term storage of the raw data in the sequencing center, reducing the time to analyse the raw data, and obviating the need for managing an extensive, local IT infrastructure (from the researcher's point of view).

New features were added, the application recipe includes:

- A scalable virtual cluster with Docker-Swarm
- An isolated network with CNSMO-VPN
- A web interface to monitor Docker containers with ready-to-use bioinformatics tools
- And authentication mechanisms using the academic federation eduGAIN for web, SSH and virtual desktop access.
- [https://nuv.la/module/cyclone/UC3\\_NGS\\_mapping](https://nuv.la/module/cyclone/UC3_NGS_mapping)
- <https://biosphere.france-bioinformatique.fr/catalogue/appliance/13>

#### 2.1.4. UC4 - Virtual Power Plant

The main idea of a Virtual Power Plant (VPP) is to integrate distributed energy resources (DER) for the generation of renewable energy and to combine them to one reliable power plant. The Virtual Power Plant connects several small and medium sized power plants, DERs, such that they are manageable as a single larger one. This comprises resources like wind, solar and bio-mass which are distributed geographically. The Virtual Power Plant provides a solution through joint control of these small and decentralized renewable energy resources, to provide reliable electricity in accordance with demand. The Virtual Power Plant optimally combines the advantages of various renewable energy sources. Wind turbines and solar modules generate electrical energy in accordance to how much wind and sun is available and when it is available. Bio-mass is used to make up the difference: it is converted into electricity as needed in order to balance out short-term fluctuations, or is temporarily stored. Combining the different kind of DERs gives the benefit of increasing the usage of renewable energy.

To control the VPP as a tailored energy supply, a new approach of ICT for managing distributed energy resources is necessary. Each Virtual Power Plant provides a single operating profile to the energy system and can react in a flexible way. The combination of the energy generation resources to a VPP enables to access a wider range of markets for selling energy. One single DER can only provide the small amount of energy it is generating. In combining the advantages of the various energy resources the Virtual Power Plants makes it possible to access energy markets usually only available for power plants which are generating a higher amount of energy and even to sell their flexibility to providers of system services.

The components and applications of the use case have been defined and deployed on Nuvla using the capabilities of SlipStream to create a base component on which several of the more specialized components are based. This way the features needed in all components only need to be defined once are included using the base component. From the CYCLONE tools the CNSMO component is integrated into the use case to form the isolated network for secure data transfer between the VPP management and the DER (Distributed Energy Resource) cloud.

#### 2.1.5. UC5 - Internet of Services Lab (IoSL)

"Internet of Services Lab"<sup>1</sup> (IoSL) is a SNET/TUB teaching project where students work a semester in groups of three to six. They most often implement software related to numerous research projects and topics: Cloud, Mobile, Social, and Ubiquitous Computing. This type of teaching project is quite popular in other computer science related study paths at TUB as well as other universities.

There are different cloud-related innovation areas where CYCLONE can provide benefits within the IoSL:

1. Rapid provisioning of resources for student projects.
2. Utilization of SlipStream modules for reproducible application deployments.
3. Federation Provider.

Student projects usually consist of three tier web applications with complementary apps which use the same backend. Some of these applications are data-intensive and strongly benefit from the one-click deployment of complex application with clusters and auto scaling mechanisms of Nuv.la. The students are encouraged to use containerization i.e. Docker to simplify the shipment of their software.

The details of the deployment depend on respective projects. However, there are two basic setups. For the simple three tier web applications, first step is to deploy web, data and business logic as containers on one VM with one-click. Second step is to use one-click deployment of complex application by distributing one container on one VM and linking them together using Nuv.la. For data-intensive applications, deployment of complex application is more suitable as the analysis is run distributed on a cluster of data nodes.

#### 2.1.6. UC6 - Multi-cloud deployment and federated identity integration of ENtRANCE components

Entrance is a research project funded by the German Ministry of Education and was carried out between March 2014 and July 2016. Entrance enables users to share files in a self-sovereign manner using an encryption mechanism called Attribute-based Encryption (ABE). ABE is similar to classic public key systems but installs in addition an attribute layer. In CP-ABE private keys are equipped with attributes and an access structure (a Boolean formula) is cryptographically bound to a ciphertext. The attributes of the private key have to solve the access structure before the decryption succeeds. Compared to classic public key crypto systems the owner of an ABE system is in charge for both, secret key generation and key distribution to recipients, and the data encryption. For both, encryption and key generation the system owner has to manage attributes. During the key generation phase attributes are attached to secret keys and during file encryption an access structure must be designed and attached to the ciphertext. Different ABE flavors, such as KP-ABE, CP-ABE and hybrid ABE exist. Since ABE can support functions, such as attribute revocation, malicious key tracing or access rights delegation, it is generally more feature rich than classic public key systems.

The Entrance System has been extended to support **Cyclone's Federation Provider** for user authentication. This addresses the fundamental question of how to provision and authenticate users in a secure way. Entrance took advantage of the new German ID card (nPA) but this approach limits the number of possible data receivers to owners of the German ID-card. With the CYCLONE Federation Provider integration new groups of data receivers can be easily provisioned by an ABE system owner.

#### 2.1.7. UC7 - Open Scientific Data

Publicly funded research creates an immense amount of data that has academic and commercial value both inside and outside of the original research domain. The taxpayers through funding agencies are increasing the pressure to make these datasets generally available through "open data" programs. A major obstacle has been finding viable business models that keep the maintenance costs for the original funding agency reasonable by monetizing the datasets for commercial use. SixSq is exploring solutions with ESA in which public data is hosted on European cloud infrastructures.

To make such a solution viable on the technical side would require: 1) detailed knowledge of the storage locations of dataset components, 2) means of placing analysis applications near the data of interest, and 3) ranking of multiple providers based on price or other characteristics. All potential features of the CYCLONE brokering and matchmaking components. There may also be a need for remote access to the public datasets to, for example, perform a combined analysis with private datasets that could not be moved to the cloud. In this case, the CYCLONE networking features may be useful.

A SixSq intern created proof of concept service for generating and updating an Earth Observation data product. It performs satellite image processing according to the client's requests, while meeting execution time and cost constraints. The architecture of the developed service is comprised of multiple entities including a benchmark system to enable cost predictions of the image processing unit. The benchmarks, cloud service offers, and resulting data products were all stored within the SlipStream Service Catalog. This proof of concept validated the general architecture for generating and maintaining an open scientific data set on the cloud using the SlipStream Service Catalog and cloud application automation.

#### 2.1.8. UC8 - Benchmark Driven Placement

A significant part of the recent design discussion for the CYCLONE brokering and matchmaking components dealt with benchmarks, both general benchmarks and application-specific benchmarks. At the recent UCC 2015 conference [7], there was a presentation from a bioinformatics group in Cardiff that used their own benchmarks for common bioinformatics tools on a number of community and public clouds to drive placement of the virtual machines [8]. The benchmarks are collected continuously from applications being run as well as directed probes of particular infrastructures.

Their work would be an interesting direct validation of the CYCLONE brokering and matchmaking design. CYCLONE would benefit by reusing their benchmarks for this validation. They would possibly benefit by replacing their ad hoc deployment engine with SlipStream, freeing them from the maintenance of their engine and allowing them to expand their benchmarks from single virtual machines to full application pipelines.

SixSq created a specific CIMI resource within SlipStream to allow users to register benchmarks associated with specific cloud services and cloud resources. This resource integrates well with the overall SlipStream Service Catalog, allowing users to take into account their benchmarks (or benchmarks published by others) when selecting cloud resources. Currently, the selection must be managed directly by the application, using the information in the Service Catalog. Eventually, the SlipStream deployment engine will directly support benchmark-driven placement.

#### 2.1.9. UC10 - Smart Utility 4.0

The first energy use cases focused on the energy generation by the distributed renewable energy resources. The aggregation in a Virtual Power Plant in order to contribute to efficient energy management was implemented. Given that the future energy management concerns the sectors of energy production, energy consumption and the transportation, it was essential to extend the energy use cases in the field of consumption and ICT platform services. The inclusion provides an energy management that enables the partners based on the aggregated energy data collected within the platform to balance the generation and consumption of energy - thus forming the Smart Utility.

The use case combining renewable energy generation with industrial consumers to achieve the optimal usage of the generated energy, has been defined on Nuvla integrating the DACI and CNSMO tools developed by CYCLONE. The DACI authorization is used for the flexible specification of policies for provider and tenants for data access. The tenants are the operators of the energy sources, the VPP and the consumers. The advantage provided by the package is that it allows separating the policy specification from the application code. The aspect oriented method of integration was used and consequently no modification of the application code was necessary. Integrating the DACI tool into the application provides a flexible method to specify access policies and quickly react to changes like e.g. adding or blocking a tenant, thus adding a layer of data access security.

Furthermore, the CNSMO network tools were included in the application to provide dynamic VPN and FW services for the data transfer between the VPP partners and clouds, i.e. within the Smart Utility. It provides a centralized network provisioning, presenting an abstracted view of the cloud for the users of the platform



and services. The VPN service allows the user to dynamically add and remove partners from the private network.

The use case includes renewable energy resources and industrial consumers in the region near Cologne.

The deployment via the CYCLONE platform provides a reliable and time effective method to set up and manage the Smart Utility application. The added value provided by DACI and CNSMO for security features and network management within the application enables the application provider to save the effort for development and maintenance of these services. The provider of the Smart Utility application can focus on the core problem having not to provide the know-how for the added services.

#### 2.1.10. UC11 - Assembling genomes from sequencing reads

To decipher the genome of new organisms, biologists split it into a very large number of redundant fragments that are subsequently sequenced (shotgun sequencing). For large eukaryotic genomes, this generates billions of reads (partial sequence of the fragments). The genome is reconstructed by piecing up overlapping reads. Algorithmically, this is done by searching for an Eulerian path in a de Bruijn graph built from the reads. For large eukaryotic genomes, this graph is particularly large and requires computers with a sizeable amount of memory (sometimes more than 1TB of RAM).

The goal in use case 11 was to provide a pipeline for bacterial genome assembly that could be deployed and used easily from a simple web interface. Our implementation is based on *bistro*<sup>1</sup>, an OCaml library for specifying and running scientific workflows. *bistro* features cached, parallel and resumable (on failure) execution, and is able to use Docker containers to run external programs. In addition, it provides an extension implementing a lightweight web server to run a pipeline, follow its execution and browse through its results. Notably, this web server is able to automatically derive an HTML form to input the data and parameter values.

For deployment, we wrote a Slipstream recipe that specifies the configuration of a virtual infrastructure running the pipeline. This infrastructure is composed of a light front-end VM that runs the web server and a virtual cluster operated via Docker Swarm. The workload required by the bioinformatics computing analysis can be then easily distributed in case of heavy requirement on a scalable bunch of VM thank to the CYCLONE developments. The cloud deployment relies on CYCLONE federation provider and CNSMO for the access management.

[https://nuv.la/module/cyclone/UC11\\_Assembling\\_genomes](https://nuv.la/module/cyclone/UC11_Assembling_genomes)

<https://biosphere.france-bioinformatique.fr/catalogue/appliance/111>

#### 2.1.11. UC12 - Metagenomics

Metagenomics is the study of the genomes present in communities of organisms found in environmental samples. Metagenomics is a generic term describing a range of studies that use high-throughput DNA sequencing (reads) to characterize microbial systems, including whole-genome shotgun (WGS) sequenced metagenomic and metatranscriptomic studies, as well as amplicon-based approaches that target specific marker genes (often rRNA genes). The analysis pipeline of WGS metagenomics consists of 4 broad steps: i) quality control of the reads, ii) sequence reconstruction and grouping, iii) gene and regulatory elements predictions, iv) function prediction for the protein-coding genes. The amplicon metagenomics pipeline comprises mostly 3 steps: i) quality control of the reads, ii) OTU clustering and picking followed by their taxonomic assignment, iii) statistical analysis. Both approaches, WGS and amplicon, involve the comparison of reads (or assembled reads) with databases of protein sequences for the former and rRNA genes for the latter. Metagenomics studies generate large amount of reads and these comparisons can be extremely computer-intensive since they involve matching this massive number of reads with large protein sequence

---

<sup>1</sup> <https://github.com/pveber/bistro>

or rRNA gene databases. Hence the need to distribute these computations over several Cloud facilities. The PathoTRACK application targets the design of software that will screen reads (*i.e.* DNA fragments) obtained by massive sequencing against specific or generalist databases for the identification of threat agents or pathogens produced by synthetic biology.

The application is now available as a complex application recipe (with several components) in NuvLa and registered in the bioinformatics Biosphere portal. The user data are uploaded through the web interface, and compared to several reference databases shared among the components. The current stable version consists of different components that are deployed in one-click as separate VM: a web user interface, a SQL service relying on the mariadb database, a job manager that uses the Snakemake workflow system to distribute the computing workload over a devoted cluster, several computing nodes and a REST API linking the different components. The cloud deployment relies on the CYCLONE federation provider access management, and on CNSMO for the security (VPN) and easiness (DNS) to deploy securely the whole environment.

The PathoTrack application need to be deployed as a multi-cloud environment on several providers to ensure:

- To keep the private reference datasets on premises
- To distribute the required computing workload on available large-scale public cloud resources

The application was demonstrated at a life science summer school about metagenomics (June 2017) and validated by attending bioinformaticians.

[https://nuv.la/module/cyclone/UC12\\_metagenomics\\_pathotrack](https://nuv.la/module/cyclone/UC12_metagenomics_pathotrack)

<https://biosphere.france-bioinformatique.fr/catalogue/appliance/100>

#### 2.1.12. UC13 - Shared environment between cloud Galaxy portals

Galaxy is an open and web-based platform for bioinformatics research that gathers multiple bioinformatics tools and allows chaining their execution in a workflow. Users of the Galaxy environment need to share histories, workflows and data although they run different instances of the same VM (most bioinformatics programs have been designed to work with ‘regular’ file systems). Once shared with other users or groups, they can use the workflow or the history as is or edit it and thus import it in its own environment. A Galaxy server is generally dimensioned to be able to handle large computation, and thus pre-empting a large amount of resources in the form of a computing cluster as the backend. The goals of having Galaxy in the CYCLONE cloud federation is to allocate on-demand adapted amount of resources through appliance, while keeping the user appreciated feature that is workflow sharing and re-using inside the Galaxy environment.

In the UC13 recipe, the galaxy portal and the bioinformatics tools are deployed automatically with Docker containers. The application now uses the CYCLONE Federation Proxy to authenticate users in Galaxy portal. Biologists can now access to their own portal according to their eduGAIN identity. The Federation proxy also provides the owner of the cloud instance a simple way to allow shared access with other biologists. The load balancing between the different user instances relies on the CNSMO-LB component. It was presented (as a poster) at the international Galaxy 2017 conference.

<https://nuv.la/module/cyclone/UC13-Galaxy>

<https://biosphere.france-bioinformatique.fr/catalogue/appliance/68>

<https://biosphere.france-bioinformatique.fr/catalogue/appliance/79>

#### 2.1.13. UC16 - Attribute-based Authorizations with XACML


Attribute-based access control allows for flexible management of permissions by enabling the use of user/resource attributes when specifying security policies.



It can be used to capture the existing authorization models such as role-based access control (RBAC) where core concepts are simply treated as attributes (e.g. role as a user attribute). In this use-case, we will demonstrate the use of XACML policies for managing permissions of CYCLONE users over available resources. The attributes provided by eduGAIN will be used to generate the policies. For instance, the following attributes: name, surname, organization type, affiliation, e-mail and username can be used to specify authorization policies. Certain types of policies can be automatically generated (i.e. from ad hoc group-based policies to XACML) from existing authorization management methods/systems of use-case owners.

The use case UC16 has been implemented (as part of the component DACI) and integrated with some of the energy related use cases (e.g. UC4. The architectural details of the components employed in the use case have been extensively discussed in deliverables D4.2, D4.4 and D4.5. Example access control policies for the use case have been specified according to requirements of various use cases including UC10.

All the code related to the use case is stored on the project repository [10] and the relevant cloud deployments are available at [11]. Figure 1 shows the deployment parameters of an application that employs the components developed in the use case. Since all subservices of DACI work on the same VM node in this application, the services are assigned different port numbers.

 <b>DACISingleDeployment</b>	Component	cyclone/CycloneDACI/DACI_Security_Services-SingleDeployment
	Default multiplicity	1
	Default cloud	default
	Parameter mappings	
	Input <code>authz_srv_port</code>	defaults to 8089
	Input <code>context_srv_port</code>	defaults to 8090
	Input <code>domain</code>	defaults to demo-uva
	Input <code>redis_address</code>	defaults to localhost
	Input <code>tenant_srv_port</code>	defaults to 8092
	Input <code>token_srv_port</code>	defaults to 8091

**Figure 1: DACI Single Deployment Parameters**

One interesting requirement for the integration of the use case with UC10 was related to the enforcement of policies. The enforcement has been implemented in such a way that the UC10 code base has been left intact thanks to the use of the aspect oriented programming techniques. The cloud application that integrates this use case with UC10 is available at [9].

### UC17 Genomic Variant Analysis for Cancer and Rare Disease Diagnosis

In the context of personalized medicine, an increasing number of patients suffering from a cancer or a rare disease will have their genome sequenced. The objective of this sequencing is to determine which variants (point mutations, gene copy number variation, insertions/deletions, DNA rearrangements) are observed in the genome. These variants are then annotated, which means that an expert tries to infer which regulatory motif on DNA or gene or metabolic/regulatory pathway is affected by the observed variants and determine whether the observed modifications in this motif, gene or pathway could be involved in the disease. Based on the expert's findings, a specific (personalized) treatment, if one exists, is prescribed to the patient.

From a bioinformatics point of view, this analysis starts with the mapping of the sequencing reads on a reference human genome and the determination of variants. This first stage, named variant calling, is carried out by a pipeline that has the reads as input, chains together several analysis programs and outputs a VCF file (i.e., a file, in a standardized format, that contains the location of the different types of variants having been found in the genome). The next stage consists in annotating the variants, as explained above. This stage is less automatic and involves running several tools locally, using web services, and accessing several databases. These two stages can be run on a cloud infrastructure that is not certified for biomedical data as far as the studied data have been anonymized before. A third step is particularly important with the ability to combine the results of this annotation stage with the

---

patient clinical data to filter out non-relevant variants. But patient clinical records are kept in hospital and cannot leave the premises of the hospital.

The application relies on the Snakemake workflow engine and several bioinformatics tools (bwa, samtools, picard, GATK, VEP). The bioinformatics components are deployed with the Conda framework, on a base cloud image of CentOS 7. At the first step, the workflow (See Figure 7) ensures that the reference data set is available (VEP/homo\_sapiens\_merged\_vep\_86\_GRCh37, 8+ GB) or downloads it. The workflow can be coupled to a Grid Engine cluster distributed over an isolated network (CNSMO-VPN, coupled with CNSMO-DNS to help the deployment of the batch cluster mode) for large workloads.

[https://nuv.la/module/cyclone/UC17\\_VCF\\_for\\_Diagnosis/uc17-wf-nte](https://nuv.la/module/cyclone/UC17_VCF_for_Diagnosis/uc17-wf-nte)

<https://biosphere.france-bioinformatique.fr/catalogue/appliance/109>

### 3. Use cases related software developments

This section provides indication of the four relevant technical where the different software components developed by CYCLONE are grouped. A specific mapping across software components and use cases is provided.

#### 3.1. CYCLONE Key technical areas

At the initial steps and during the project, the CYCLONE consortium has defined four key technical areas:

- 1) Cloud Access Management through cloud proxies,
- 2) Matchmaking, Brokering, and Mediation of Cloud Resources,
- 3) End-to-end Security for HTTP-based Applications,
- 4) Dynamic network configuration and management.

**Table 3: CYCLONE key technical areas and associated software developments**

Key technical areas	Software developments
<b>Deployment:</b>	Cloud deployment
<b>1) Cloud Access Management</b>	Complex App deployment
<b>2) Matchmaking, Brokering, and Mediation of Cloud Resources</b>	Multi-cloud deployment
<b>Security:</b>	Federation Proxy
<b>3) End-to-end Security for HTTP-based Applications</b>	Web authentication
	SSH authentication
	DACI
<b>Network:</b>	CNSMO-VPN
<b>4) Dynamic network configuration and management</b>	CNSMO-FW
	CNSMO-LB
	CNSMO-DNS

#### 3.2. Software implications according to the use cases requirements

For all use cases, the CYCLONE consortium identified common requirements (see the CYCLONE use case portal [6] for a detailed a description of them). We defined precisely the links between the requirements, the key technical areas, and the implications related to the software components developed by CYCLONE (see the CYCLONE main web site for details [6]). Table 4 shows a detailed view of these links between the requirements of CYCLONE use cases and software tools developed in the context of the CYCLONE project.

**Table 4: Details of achieved use cases requirements and related CYCLONE software implications**

	Key Tech. Areas	Cloud deployment	Complex App	Multi-cloud	Federation Proxy	Web authn	SSH authn	DACI	CNSMO-VPN/SDN	CNSMO-FW	CNSMO-LB	CNSMO-DNS
UC1	1, 2, 3	✓	-	✓	✓	✓	-	-	-	-	-	-
UC2	1, 2, 4	✓	✓	✓	✓	✓	✓	-	✓	-	-	✓
UC3	2, 4	✓	✓	✓	✓	-	✓	-	✓	-	-	-
UC4	2, 4	✓	✓	✓	-	-	-	-	✓	✓	-	-
UC5	1, 3	✓	✓	-	✓	✓	✓	-	-	-	-	-
UC6	1,3	✓	-	-	✓	✓	-	-	-	-	-	-
UC7	2, 1	✓	✓	✓	✓	✓	-	-	-	-	-	-
UC8	2	✓	✓	✓	-	-	-	-	-	-	-	-
UC10	1,3,4	✓	✓	✓	-	-	-	✓	✓	✓	-	-
UC11	2	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-
UC12	2, 3, 4	✓	✓	✓	✓	✓	✓	-	✓	✓	-	✓
UC13	1, 2	✓	-	✓	✓	✓	-	-	-	-	✓	-
UC16	1,3	✓	✓	✓	✓	✓	✓	✓	-	-	-	-
UC17	1,2,3,4	✓	✓	✓	✓	✓	-	-	✓	✓	-	✓

---

## 4. Conclusions

In this document, we provided an overview of all CYCLONE use cases. The main motivations of these use cases were to help to identified relevant requirements, and to evaluate their satisfaction with newly developed components, for the multi-cloud deployment of legacy and new applications.

The overall analysis of all CYCLONE use cases including the identified common requirements, and their level of achievements at the end of the project, demonstrates that these goals have been reached. Among the 48 identified common requirements, 38 were satisfied, with almost all requirements qualified as “MUST be satisfied”.

CYCLONE developments strongly contribute to the adaptation of use cases to multi-cloud environment, especially on the following point:

- SlipStream/Nuvla for easiness of integration and one-click multi-cloud deployment of complex applications
- Federation Provider for authentication and authorization management based on eduGAIN federation identity
- DACI for the authorizations management and the enforcement of policies
- CNSMO for secure (VPN, FW) multi-cloud deployment with high-level features (DNS, LB)

The initial project use cases as described in the DOA, has been enlarged with a set of use-cases and deployment scenarios which clearly provide indication of the tremendous impact that the project artefacts (in terms of software modules) are likely to have in real world deployment. This impact, also evaluated in D3.4 from a social and business point of view, is a tangible proof of the high exploitation potential of CYCLONE prototype.

---

## References

- [1] CYCLONE Deliverable D3.1, <http://www.cyclone-project.eu/assets/images/deliverables/Evaluation%20of%20Use%20Cases.pdf>
- [2] CYCLONE Deliverable D3.2, <http://www.cyclone-project.eu/assets/images/deliverables/Consolidated%20Evaluation%20of%20Use%20Cases.pdf>
- [3] CYCLONE Deliverable D3.3
- [4] CYCLONE Deliverable D3.4
- [5] CYCLONE web site, <http://www.cyclone-project.eu>
- [6] CYCLONE use case portal, <https://cyclone.france-bioinformatique.fr/usecases>
- [7] <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7430473>
- [8] [http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7431434&searchWithin%3Dbenchmark%26filter%3DAND%28p\\_IS\\_Number%3A7431374%29](http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7431434&searchWithin%3Dbenchmark%26filter%3DAND%28p_IS_Number%3A7431374%29)
- [9] UC16 Integration with the Energy usecase (UC4), [https://nuv.la/module/cyclone/CycloneDACI/Demo\\_DACI\\_Finesce\\_Application](https://nuv.la/module/cyclone/CycloneDACI/Demo_DACI_Finesce_Application)
- [10] DACI source code, <https://github.com/cyclone-project/cyclone-DACI>
- [11] UC16 Slipstream Deployments, <https://nuv.la/module/cyclone/CycloneDACI>
- [12] CNRS Catalogue of Bioinformatics Cloud Appliances, <https://biosphere.france-bioinformatique.fr/catalogue>
- [13] CYCLONE SlipStream/Nuvla instance, <https://nuv.la/module/cyclone/>

## 5. Acronyms

B2B	Business to Business
CSP	Cloud Service Providers
DC	Data Center
DNA	DeoxyriboNucleic Acid
E2E	End to End
FLA	Federation Level Agreement
IaaS	Infrastructure-as-a-Service
IPR	Intellectual Property Rights
IT	Information Technology
MaaS	Metal as a Service
MCSP	Multi-Cloud Service Provider
NaaS	Network-as-a-Service
Net-HAL	Network Hardware Abstraction Layer
NFV	Network Function Virtualization
NGS	Next-Generation Sequencing
OLA	Operational Level Agreement
PaaS	Platform-as-a-Service
PC	Project Coordinator
PMB	Project Management Board
PoP	Point of Presence
SaaS	Software-as-a-Service
SCI	Smart Core Interworks
SDN	Software Defined Networks
SLA	Service level agreement
SP	Service Provider
TC	Technical Coordinator
TCTP	Trusted Cloud Transfer Protocol
TMB	Technical Management Board
VCF	Variant Call Format
VM	Virtual Machine
WP	Work Package
WPL	Work Package Leader

**<END OF DOCUMENT>**