CYCLONE: A Platform for Data Intensive Scientific Applications in Heterogeneous Multi-cloud/Multi-provider Environment

Yuri Demchenko, Miroslav Zivkovic, Cees de Laat
University of Amsterdam
{y.demchenko, M.Zivkovic, C.T.A.M.deLaat}@uva.nl

José Ignacio Aznar Baranda
I2CAT
jose.aznar@i2cat.net

Christophe Blanchet, Mohamed Bedri, Jean-François Gibrat
CNRS IFB
{christophe.blanchet, mohamed.bedri, jean-francois.gibrat}@france-bioinformatique.fr

Oleg Lodygensky
LAL
oleg.lodygensky@lal.in2p3.fr

Mathias Slawik, Ilke Zilci
TU Berlin
{mathias.slawik, ilke.zilci}@tu-berlin.de

Rob Branchat, Charles Loomis
SixSq Sàrl
{rob, cal}@sixsq.com

***Abstract*—This paper presents results of the ongoing development of the CYCLONE as a platform for scientific applications in heterogeneous multi-cloud/multi-provider environment. The paper explains the general use case that provides a general motivation for the CYCLONE architecture and provides detailed analysis of the bioinformatics use cases that define specific requirements to the CYCLONE infrastructure components. Special attention is given to the federated access control and security infrastructure that must provide consistent security and data protection for distributed bioinformatics data processing infrastructure and distributed cross-organisations collaborating teams of scientists. The paper provides information about selected use cases implementation using SlipStream cloud automation and management platform with application recipe example. The paper also addresses requirements for providing dedicated intercloud network infrastructure which is currently not addressed by cloud providers (both public and scientific/community).**

Keywords- CYCLONE cloud automation platform for scientific application, SlipStream Cloud Management platform, Data intensive scientific applications requirements, Intercloud Architecture Framework (ICAF), Federated Cloud Infrastructure

I. INTRODUCTION

Modern data intensive research require continuously increasing power of computing resources and storage volume that are in many cases required on-demand for specific operations in data lifecycle (e.g. for data collection, extraction, processing and reporting) and to be elastically scaled. Cloud Computing [1, 2] provides a platform and environment for data intensive scientific integration and enables effective use of Big Data technologies and distributed data resources [3, 4]. Large volumes of data used in modern research, their distributed nature, and need to support distributed collaborative researcher groups create new challenges for the scientific applications engineering.

The presented paper provides information about ongoing development and implementation of the CYCLONE platform for data intensive scientific applications development, integration and automated deployment in heterogeneous multi-cloud multi-provider environment [5]. The paper refers to the general use case that motivates the general CYCLONE platform architecture requirements and discusses in details specific bioinformatics use cases to define specific requirements to CYCLONE platform and infrastructure components.

The CYCLONE platform is based on existing individual components developed by the authors in previous projects: SlipStream [6], StratusLab [7], TCTP [8], OpenNaaS [9], and ensures their integration in a single platform supporting the whole lifecycle of the scientific applications development. The paper refers to the general Intercloud Architecture Framework (ICAF) [10, 11] and the Intercloud Federation Framework (ICFF) [12] proposed in the earlier authors' work.

The remainder of the paper is organized as follows. Section II describes the general use case and CYCLONE architecture. Section III discusses bioinformatics use cases and identifies specific requirements to bioinformatics applications. Section IV provides suggestions for initial use cases implementation, section V describes the SlipStream cloud applications management platform, its functionality. Section VI provides general information about CYCLONE federated security infrastructure. The paper concludes with the summary and remarks on future development in section VII.

II. GENERAL USE CASE FOR INTERCLOUD APPLICATIONS DEPLOYMENT PLATFORM AND CYCLONE ARCHITETCURE

Multiple individual use cases for multi-cloud applications that require cloud and non-cloud resources integration into one intercloud infrastructure that executes a single or multiple enterprise or scientific workflows can be abstracted into general scenario and use case illustrated in Figure 1. It includes two interacting applications, that in general can be multi-cloud, that contain both application related and management components. Application component interacts with end users, management component is controlled by application administrator and interacts with the (inter)cloud management software. The figure also shows Cloud Applications Deployment and Management Software and Tools as an important component to support cloud applications deployment and operation during their whole lifecycle. Intercloud infrastructure should also provide two other components or services: federated access control and security, and intercloud network infrastructure that needs to be provisioned as a part of overall application infrastructure.

Intercloud applications and infrastructure may include multiple existing cloud platforms. In the generally distributed

heterogeneous multi-cloud multi-provider environment the problem of applications integration and management is becoming critical and require smooth integration with the application workflow and automation of most of development and operation functions, ideally integration with the application specific development and operation (DevOps) tools [13]. Currently widely used cloud automation tools such as Chef [14], Puppet [15] allow single cloud provider application deployment. They don't solve problem of multi-cloud resources/services integration and provisioning of inter-cloud network infrastructure.

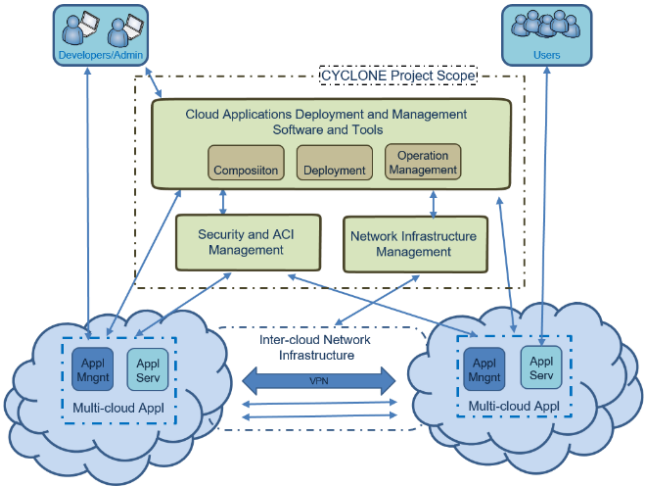

Figure 1. General use case for multi-cloud applications deployment.

The CYCLONE project attempts to solve this problem by leveraging original cloud management platform SlipStream and extending its with necessary functionality and components, in particular for inter-cloud resources deployment and network infrastructure provisioning, enabling federated access control for users and end-to-end security for data transfer, enabling dynamic trust establishment between cloud and application domains.

Intercloud platforms should deliver open integration environment and preferably standardized APIs, protocols, and data formats, allowing for cross-cloud resources interoperability. Practical Intercloud platform development should target two major stakeholders and user communities the Application Service Providers (ASPs) as well as their customers to address real life challenges and problems in a consistent and constructive way.

The effective cloud automation and management platform should allow dynamic cloud resources allocations depending on the workload and application workflow. This task can be solved for single cloud using its native elasticity and load balancing tools, however in intercloud environment such functionality will require involving real cloud platform load (including resources availability) and application monitoring [16].

III. BIOINFORMATICS USE CASES AND REQUIREMENTS

A. *Overall description*

Bioinformatics deals with the collection and efficient analysis of biological data, particularly genomic information from DNA sequencers. The capability of modern sequencers to produce terabytes of information coupled with low pricing (less than US$1000 for a human genome) that makes parallel use of many sequencers feasible causes a "data deluge" that is being experienced by researchers in this field [17, 18].

Bioinformatics software is characterized by a high degree of fragmentation: literally hundreds of different software packages are regularly used for scientific analyses with an incompatible variety of dependencies and a broad range of resource requirements. For this reason, the bioinformatics community has strongly embraced cloud computing with its ability to provide customized execution environments and dynamic resource allocation.

The French Institute of Bioinformatics (IFB) [19] consists of 32 regional bioinformatics platforms (PF) grouped into 6 regional centers spanning the entire country, and a national hub, the "UMS 3601–IFB-core". The IFB has deployed a cloud infrastructure on its own premises at IFB-core, and aims to deploy a federated cloud infrastructure over the regional PFs devoted to the French life science community, research and industry, with services for the management and analysis of life science data.

The CYCLONE project has identified two basic bioinformatics use cases that aim to address specific identified limitations while carrying out the analyses starting from general cloud federation requirements and specific use cases features to enhance current cloud infrastructure processes and services provisioning mechanisms.

B. *UC1 - Securing human biomedical data*

1) *Description*

Continuous decrease of the genome sequencing costs (NGS) allows increasing number of clinicians to include genome analysis and data into their day-to-day diagnostic practice. Today, most of 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.

It is expected that 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 the subject to strict personal data protection regulations. To ensure the data security while carrying out the analysis in a federated cloud environment, the security of all involved sites belonging to the federation must be ensured (especially when involving both public and private cloud infrastructures).

2) *Workflow*

The use case workflow to ensure data security includes the following steps (see Figure 2): (1) a biomedical user connects to the cloud through the IFB web authenticated dashboard; uses it to (2) run an instance of the appliance containing the relevant pre-configured analysis pipeline. At step (3) the VM containing genome applications is deployed on the cloud (testbed); then (4) the user signs into the web interface of the VM, (5) uploads the patient's biomedical data, and (6) runs the analysis in a secure environment. Finally, (7) the user gets the results.

The bioinformatics treatment generally relies on a comparison with the current release of the reference human genome hg19 (Human Genome version 19 or GRCh37). The hg19 is a database consisting of many files containing the public genomics data. It can be used remotely (with sufficient connectivity) or can be previously deployed by the cloud providers as a public data set available to all users.

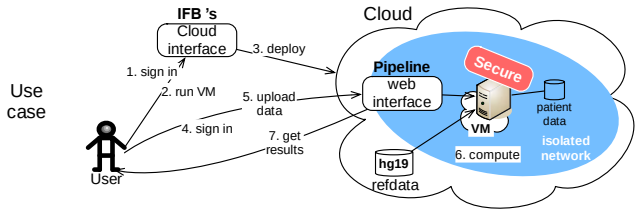

Figure 2: Functional schema of the use case "Securing human biomedical data". *The figure shows the application components and describes the different steps of the workflow.*

3) Implementation requirements

This use case demonstrates the capability of CYCLONE infrastructure to provide biomedical staff acting as cloud users with the deployment of their own cloud infrastructure for the biomedical analyses consisting of a single virtual machine (potentially elastically scaled) with a web interface in a secured environment. The access to this environment will be based on the identity and authorizations in the federation.

The tests within the project will be carried out only on non-personally identifiable data (either anonymized benchmark data or simulated data).

C. UC2 - Cloud pipeline for microbial genomes analysis

1) Description

In the post-NGS research, sequencing bacterial genomes is very cheap (few hundreds €) what allows researchers to compare large collections of related genomes (strains). Thus, this brings requirements to increasing need for automating the annotation of bacterial genomes.

The IFB-MIGALE platform (one of the bioinformatics platforms of the IFB) has developed an environment for the annotation of microbial genomes and a tool for the visualization of the synteny (local conservation of the gene order along the genomes). The platform automatically launches a set of bioinformatics tools (e.g. BLAST, INTERPROScan) to analyse the data and stores the results of the tools in a relational database (PostgreSQL). These tools use several public reference data collections. A web interface allows the user to consult the results and perform the manual annotation (manual annotation means adding manually metadata and biological knowledge to the genome sequence). Installing the platform requires advanced skills in system administration and application management. Performing the analysis of collections of genomes requires large computing resources that can distributed over several computers, generally the computing nodes of a cluster.

The proposed CYCLONE cloud federation will allow the life science researchers to deploy their own comprehensive annotation platform over one or more cloud infrastructures. Such deployments can be done with the dynamic allocation of network resources for the isolation of the VMs inside a dedicated private cloud including virtual network and replicated user data.

2) Workflow

As illustrated in Figure 3, a bioinformatician (1) connects to the cloud web dashboard, uses it to (2) run and (3) deploy with one click a genomes annotation platform consisting of many VMs, comprising of a master node of the virtual cluster that provides also the visualization web-interface, associated with several computing nodes. Then the user (4) uses secure communication over SSH to connect to the master and (5) uploads the raw microbial genomic data (MB) to the cloud storage. SCP/SFTP protocols are used from a command line tool or a GUI, to ensure AuthN/Z for the data transfer, and to overcome the performance issues of HTTP for large datasets. Still in command line interface, the user (6) runs the computation to annotate the new microbial genomes. The first step consists of many data-intensive jobs performing the comparisons between the new genome and the reference data

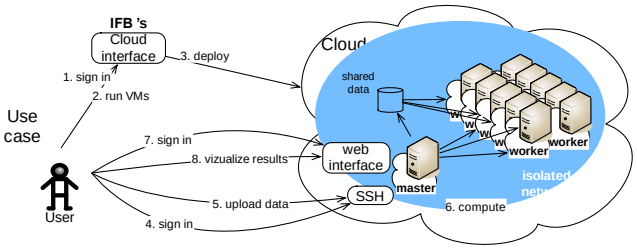

Figure 3: Functional schema of the use case "Cloud virtual pipeline for microbial genomes analysis". *The figure shows the application components and describes the different steps of the workflow.*

The results are stored in a relational database (provided by a cloud service or a VM deployed within the platform). Then the scientist (7) signs in the annotated data visualization environment provided by the Insyght web-interface to 8) navigate between the abundant homologues, syntenies and gene functional annotations in bacteria genomes.

3) Implementation requirements

This use case will evaluate the capability of CYCLONE infrastructure to provide a bioinformatician with a one-click deployment of a complex application. This deployment needs to be done in an isolated network for security and confidentiality reasons. The access to this environment will be based on the identity and authorizations in the federation. The application may be deployed over several clouds because it could require large computing resources not available in one place or due to functional reasons some data or tools are only available in certain environments or locations.

IV. Implementation of Use Cases and CYCLONE Infrastructure Components

A. Deployment of the use cases

The deployment of the bioinformatics use cases was done in a progressive manner, starting with one VM integrating

security features (UC1 Securing human biomedical data). Afterwards, we then deployed a complex application requiring the coordinated deployment of several VMs (UC2 Cloud virtual pipeline for microbial genomes analysis).

1) Deployment UC1 Securing human biomedical data

The first deployed bioinformatics use case "Securing human biomedical data" is a single-VM application requiring enhanced security features such as a trusted federated authentication mode and a deployment done only on certified (by the French Health Ministry) cloud infrastructure. The cloud appliance NGS-Unicancer is developed by the bioinformatics platform of the Centre Léon Bérard (Lyon, France, www.synergielyoncancer.fr) in the context of the project NGS-Clinique (INCA - Institut National du Cancer). It provides a simple web interface to launch the biomedical genomic analysis pipeline. The appliance was enhanced by the CYCLONE Federation Provider and is ready for on-demand deployment on the IFB-core cloud infrastructure. The user deploys the appliance NGS-Unicancer through the IFB web interface in "1-click" and uses the CYCLONE federation provider to get access to the VM web interface based on its identity in the federation. The user can then easily upload its data, run the analysis and get the results. In Figure 4, the upper part describes the use case workflow, middle layer represents the workflow steps that are linked to the related CYCLONE software components and services. The bottom part shows the testbed infrastructure components.

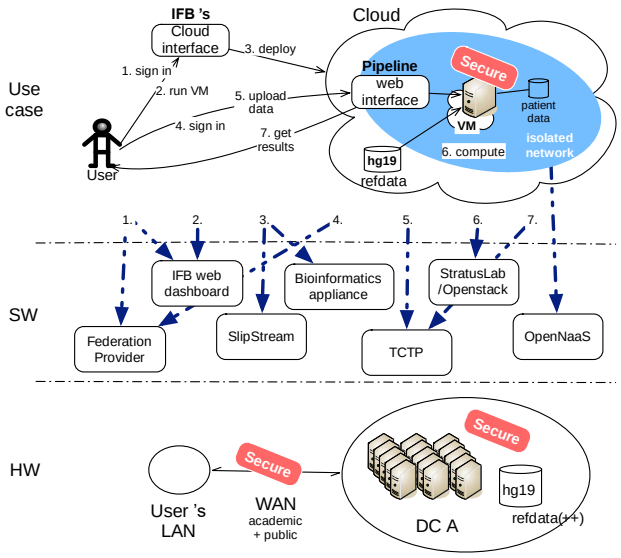

Figure 4: Functional relations between the use case "Securing human biomedical data" and the Cyclone components.

2) Deployment UC2: Cloud virtual pipeline for microbial genomes analysis

The second bioinformatics use case "Cloud virtual pipeline for microbial genomes analysis" is developed by the platform IFB-MIGALE (Jouy-en-Josas, France, migale.jouy.inra.fr). This application requires several components: a user web interface, a relational postgreSQL database, and a complete computing cluster with a master and several nodes to perform the data-intensive analyses. This infrastructure already running in a classical static way on bare-metal servers in IFB-MIGALE premises was ported to the cloud and extended with a « 1-click » deployment features by using SlipStream recipes. The image was exported from the IFB's cloud and registered in the StratusLab Marketplace. Afterwards, IFB-core wrote a deployment recipe based on SlipStream that instantiates the complete application with all the required VMs on the CYCLONE infrastructure.

B. Secure Network Infrastructure Provisioning

Due to the generically distributed Bioinformatics resources and applications required to be integrated in typical use cases, there is a strong requirements to provision dedicated secure virtualised network infrastructure. The following are specific requirements:

- Dynamically provisioned high-performance network infrastructure to interconnect multiple data sources locations and enable high-performance computations over large volume of data. This functionality is being developed based on OpenNaaS network provisioning and monitoring platform [9].
- Guaranteed network performance (QoS) is specifically important for remote data visualisation such as IGV - Integrative Genomics Viewer. The QoS (mainly bandwidth and latency) will require the traffic engineering mechanisms at the network level.
- Multi-domain Virtual Private Network (VPN) is required to interconnect distributed data resource and application components deployed in different clouds. Firewall configuration is identified as an important part of intercloud applications integration.

V. SLIPSTREAM: CLOUD APPLICATION MANAGEMENT PLATFORM

Within CYCLONE, software developers and service operators manage the complete lifecycle of their cloud applications with SlipStream, an open source cloud application management platform. Through its plugin architecture, SlipStream supports most major cloud service providers and the primary open source cloud distributions. By exposing a uniform interface that hides differences between cloud providers, SlipStream facilitates application portability across the supported cloud infrastructures.

To take advantage of cloud portability, developers define "recipes" that transform pre-existing, "base" virtual machines into the components that they need for their application. By reusing these base virtual machines, developers can ensure uniform behaviour of their application components across clouds without having to deal with the time-consuming and error-prone transformation of virtual machine images. Developers bundle the defined components into complete cloud applications using SlipStream facilities for passing

information between components and for coordinating the configuration of services.

Once a cloud application has been defined, the operator can deploy the application in "one click", providing values for any defined parameters and choosing the cloud infrastructure to use. With SlipStream, operators may choose to deploy the components of an application in multiple clouds, for example, to provide geographic redundancy or to minimize latencies for clients. To respond to changes in load, operators may adjust the resources allocated to a running application by scaling the application horizontally (changing the number of virtual machines) or vertically (changing the resources of a virtual machine).

SlipStream combines its deployment engine with an "App Store" for sharing application definitions with other users and a "Service Catalog" for finding appropriate cloud service offers, providing a complete engineering PaaS supporting DevOps processes. All of the features are available through its web interface or RESTful API.

A. *Functionality used for use cases deployment*

The bioinformatics use cases described above principally used SlipStream's facilities and tools to define applications and its deployment engine through the RESTful API.

The definition of an application component actually consists of a series of recipes that are executed at various stages in the lifecycle of the application. The main recipes, in order, are:

- **Pre-install**: Used principally to configure and initialize the operating system's package management.
- **Install packages**: A list of packages to be installed on the machine. SlipStream supports the package managers for the RedHat and Debian families of OS.
- **Post-install**: Can be used for any software installation that can not be handled through the package manager.
- **Deployment**: Used for service configuration and initialization. This script can take advantage of SlipStream's "parameter database" to pass information between components and to synchronize the configuration of the components.
- **Reporting**: Collects files (typically log files) that should be collected at the end of the deployment and made available through SlipStream.

There are also a number of recipes that can be defined to support horizontal and vertical scaling that are not used in the defined here use cases.

The applications are defined using SlipStream's web interface, the bioinformatics portal then triggers the deployment of these applications using the SlipStream RESTful API.

B. *Example recipes*

The application for the bacterial genomics analysis consisted of a compute cluster based on Sun Grid Engine with an NFS file system exported from the master node of the cluster to all of the slave nodes. The master node definition was combined into a single "deployment" script that performed the following actions:

1. Initialize the yum package manager.
2. Install bind utilities.
3. Allow SSH access to the master from the slaves.
4. Collect IP addresses for batch system.
5. Configure batch system admin user.
6. Export NFS file systems to slaves.
7. Configure batch system.
8. Indicate that cluster is ready for use.

The deployment script extensively uses the parameter database that SlipStream maintains for each application to correctly the configure the master and slaves within the cluster. A common pattern is the following:

```
ss-display "Exporting SGE_ROOT_DIR..."
echo -ne "$SGE_ROOT_DIR\t" > $EXPORTS_FILE
for ((i=1; i<=`ss-get
      Bacterial_Genomics_Slave:multiplicity`; i++ ));
do
  node_host=`ss-get
      Bacterial_Genomics_Slave.$i:hostname`
  echo -ne $node_host >> $EXPORTS_FILE
  echo -ne "(rw,sync,no_root_squash) ">> $EXPORTS_FILE
done
echo "\n" >> $EXPORTS_FILE # last for a newline
exportfs -av
```

which is used to export one NFS directory. The `ss-get` command retrieves a value from the parameter database. In this case, it determines the number of slaves and then loops over each one, retrieving each IP address (hostname) and adding it to the NFS exports file. This command will wait for a value to be set, allowing it to act as a semaphore and allowing coordination between the components of the application. An analogous `ss-set` command allows the values of parameters to be set.

A similar pattern is used for the SSH and batch system configurations in this script.

VI. CYCLONE FEDERATED SECURITY INFRASTRUCTURE

The CYCLONE security infrastructure aims to enable holistic security functionality in federated multi-cloud infrastructures by offering a set of ready-to-use components. The components are selected to best serve the needs of the distributed bioinformatics data processing infrastructure and the distributed cross organization collaborating research teams. The main identified security scenarios include: i) federated identity management, ii) federated authorization management, iii) end-to-end secure data management.

Federated identity management is enabled by the CYCLONE Federation Provider. The CYCLONE Federation Provider provides federated user identities via the SAML 2.0 web authentication workflow with the local identity providers. It supports all Identity Providers of eduGAIN [21]. Providing a federated identity includes user attribute mapping, which means transforming SAML user assertions into JSON Web Token claims. The current implementation with Keycloak [22] and SimpleSAMLphp allows browser based log-in to services which implement OpenID Connect

[23] client side which provides the library support in a variety of programming languages. The use of OpenID Connect and JSON Web Tokens add value to the CYCLONE system, since these are much easier to implement in comparison to SAML and its tokens for service providers. Moreover, with extension of the SlipStream authentication service to implement the OpenID Connect Authorization Code Flow, eduGAIN users will be able to log in to the SlipStream web dashboard to configure their deployments.

Federated authorization management addresses the collaboration between VM developers and bioinformaticians and the definition of access rules. It will be enabled by lists of users and groups defined at the pre-deployment stage which will be configured on the VMs. The user claims can be used to define a simple configuration with Require Statements in .htaccess on any Apache-hosted application with the OpenIdConnect Module enabled.

CYCLONE security infrastructure will address the end-to-end secure data management with the possible integration of TCTP [8]. Further security infrastructure will address dynamic trust infrastructure provisioning and trust bootstrapping protocol [24, 25].

VII. Conclusion and Future Development

This paper presents an on-going research and development of the advanced CYCLONE platform for data intensive applications development, deployment and management in heterogeneous multi-cloud multi-provider environment.

The current stage of the development concludes the development of a general platform components' architecture, functional design and pilot implementation of the core infrastructure provisioning functionality for selected bioinformatics usecases that represent one of the most demanding use cases that require high-performance dynamically configured multi-cloud infrastructure and complex data protection and access control mechanisms. The CYCLONE architecture is built around core functionality of the SlipStream Cloud Management platform and extends its with new functionalities to provision intercloud network infrastructure and dynamically created security infrastructure.

Acknowledgement

The research leading to these results has received funding from the Horizon2020 project CYCLONE and the French programs PIA INBS 2012 (CNRS IFB). We are thankful to our colleagues who collaborate with us to deploy selected bioinformatics applications in the CYCLONE testbed: Christian Baudet (Centre Léon Bérard, Lyon, France) and Thomas Lacroix (IFB-MIGALE, Jouy-en-Josas, France).

References

[1] NIST SP 800-145, "A NIST definition of cloud computing", [online] http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[2] NIST SP 500-292, Cloud Computing Reference Architecture, v1.0. [Online] http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505

[3] NIST Big Data Working Group Documents [online] http://bigdatawg.nist.gov/V1_output_docs.php

[4] Demchenko, Yuri, Peter Membrey, Cees de Laat, Defining Architecture Components of the Big Data Ecosystem. Proc. The 2014 Intern Conf. on Collaboration Technologies and Systems (CTS 2014), May 19-23, 2014, Minneapolis, USA

[5] Slawik, Mathias, Yuri Demchenko, José Ignacio Aznar Baranda, Robert Branchat, Charles Loomis, Oleg Lodygensky, Christophe Blanchet,CYCLONE Unified Deployment and Management of Federated, Multi-Cloud Applications, Proceedings The 8th IEEE/ACM International Conference on Utility and Cloud Computing (UCC2015), December 7-10, 2015, Limassol, Cyprus

[6] Slipstream Cloud Management Platform. [online] http://sixsq.com/products/slipstream/

[7] StratusLab Scientific Cloud [online] http://www.stratuslab.eu/

[8] M. Slawik, "The Trusted Cloud Transfer Protocol," in Proc. *2013* IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom), 2-5 December 2013, Bristol, UK

[9] OpenNaaS [online] https://github.com/dana-i2cat/opennaas

[10] Demchenko, Y., C.Ngo, M.Makkes, R.Strijkers, C. de Laat, Intercloud Architecture for Interoperability and Integration. Proc. The 4th IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2012), 3 - 6 December 2012, Taipei, Taiwan.

[11] Demchenko, Y., et al, Intercloud Architecture Framework for Heterogeneous Cloud based Infrastructure Services Provisioning On-Demand. The 27th IEEE International Conference on Advanced Information Networking and Applications (AINA2013). 25-28 March 2013. ISBN-13: 978-0-7695-4953-8.

[12] Y.Demchenko, C. Lee, C.Ngo, C. de Laat, Federated Access Control in Heterogeneous Intercloud Environment: Basic Models and Architecture Patterns. In Proc IEEE International Conference on Cloud Engineering (IC2E), March 11, 2014, Boston, USA

[13] Davis, Jennifer; Daniels, Katherine (2015). Effective DevOps. O'Reilly. ISBN 978-1-4919-2630-7.

[14] Chef: Cloud Automation deployment and DevOps platform [online] https://www.chef.io/chef/

[15] Puppet: Cloud Automated Provisioning and Management https://puppetlabs.com/

[16] Dumitru, C. , Oppresscu, AM., Živković, M. , R van der Mei, Grosso P. , de Laat, C. A queueing theory approach to Pareto optimal bags-of-tasks scheduling on clouds, Euro-Par 2014 Parallel Processing, pp. 162-173.

[17] Marx, V., Biology: The big challenges of big data.Nature, 2013 vol. 498 (7453) pp. 255-260

[18] Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., *et al*., Big Data: Astronomical or Genomical? PLoS Biol, 2015 vol. 13 (7): e1002195.

[19] French Institute of Bioinformatics – CNRS IFB UMS3601, http://www.france-bioinformatique.fr/

[20] Lacroix, T., Loux, V., Gendrault, A., Hoebeke, M. and Gibrat, J.F. Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol! Nucl. Acids Res., 2014 vol. 42 (21): e162.

[21] eduGAIN: Federated Identity Provider infrastructure interconnection service [online] http://services.geant.net/edugain/Pages/Home.aspx

[22] Keycloak: Integrated SSO and IDM for browser apps and RESTful web services [online] http://keycloak.jboss.org/

[23] OpenID Connect [online] http://openid.net/connect/

[24] Ngo, C., Y.Demchenko, C. de Laat, Toward a Dynamic Trust Establishment Approach for Multi-provider Intercloud EnvironmentThe 4th IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2012), 3 - 6 December 2012, Taipei, Taiwan

[25] Membrey, P., K.C.C.Chan, C.Ngo, Y.Demchenko, C. de Laat, Trusted Virtual Infrastructure Bootstrapping for On Demand Services. The 7th International Conference on Availability, Reliability and Security (AReS 2012), 20-24 August 2012, Prague.