MS1.1.2: INTERMEDIATE USE CASES FOR THE COMMUNITY CONNECT (COCO) SERVICE

AUTHORS: BART GIJSEN (TNO)
REVIEWERS: SYLVIA KUIJPERS, MARIJKE KAAT, RONALD VAN DER POL (SURFNET)
SVEN WARRIS (WUR)
DATE: SEPTEMBER 30, 2014
VERSION: 1.0

Reading Guide

This document is a refinement of the MS1.1.1 report "Initial use cases for the Community Connect (CoCo) service." In particular, the report is extended with section 4 that contains a description of the business perspective on the "DNA Sequencer as a Service" use case, and how the CoCo service supports this multi-domain use case.

1. Introduction

The advent of new networking technologies and Software Defined Networking in particular, are creating innovation opportunities for a wide range of use cases. In the GN3Plus CoCo project SURFnet and TNO focus on the opportunities for the eScience community. The objective of the CoCo project is to develop, demonstrate and validate a proof of concept for an "On-Demand Community Connection Service for eScience Collaboration".

The CoCo prototype service will enable scientists from multiple organizations to dynamically create virtual, private networks for sharing services and facilities as if they were collaborating within a single network environment. The CoCo service will be exposed through a simple and easy to use interface, without the need to involve network administrators for each virtual network instance. Federated authentication and authorization will be integrated to ensure authentication and confidentiality requirements of the CoCo service. Figure 1 illustrates a CoCo instance that connects endpoints, for example laptop's or other devices used by scientist or scientific instruments, from multiple domains.

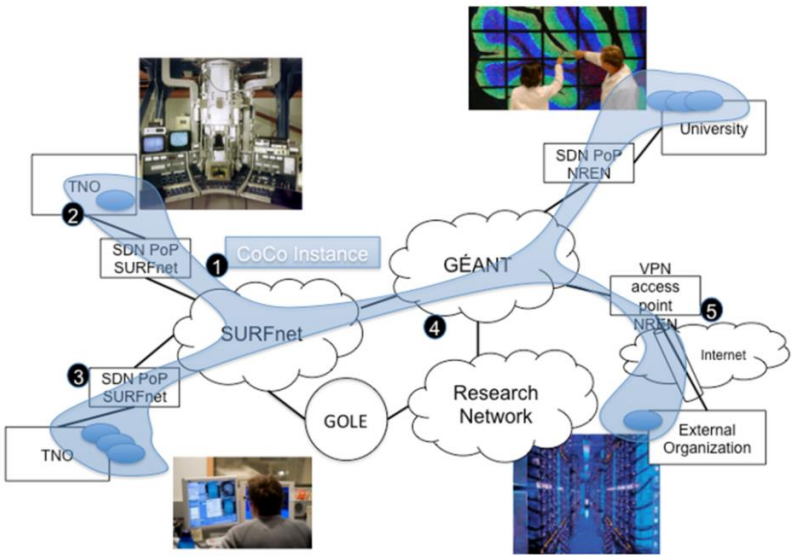


Figure 1: Topological overview of a CoCo instance connecting endpoints in multiple domains

The envisioned CoCo service is a general connectivity service that can be used to support multiple types of research collaboration simultaneously. This is illustrated in Figure 2 for two of the use cases that were identified in the initial use cases workshop (see also the descriptions of those use cases in section 3). For example, the "DNA sequencer as a Service" consists of a service specific portal in the service plane that end-users interact with, and that communicates with the CoCo agents in the control plane. The portal also coordinates service level and user group information for the specific service (e.g. a "DNA sequencer as a Service"). In addition, each portal administrates an abstract view of the resources that are used for deploying the service and enables on-demand resource management for end-users (or devices).

Management of the connectivity resources will be done via the CoCo agents in the control plane. The CoCo agents manage the multi-domain aspects for each of the services that they support. Not shown in this figure are the control planes for other resources such as virtual machines ("Compute") and storage ("Store"), because the control of those resources is out of the scope of the CoCo project.

Each domain has an OpenFlow based data plane infrastructure that is controlled by its own CoCo agent. In the data plane the CoCo instances (from multiple services) are deployed as separate, virtual private flows. This way the CoCo control and data plane components contribute to the creation of flexible and efficient multi-domain services that will innovate research collaboration.

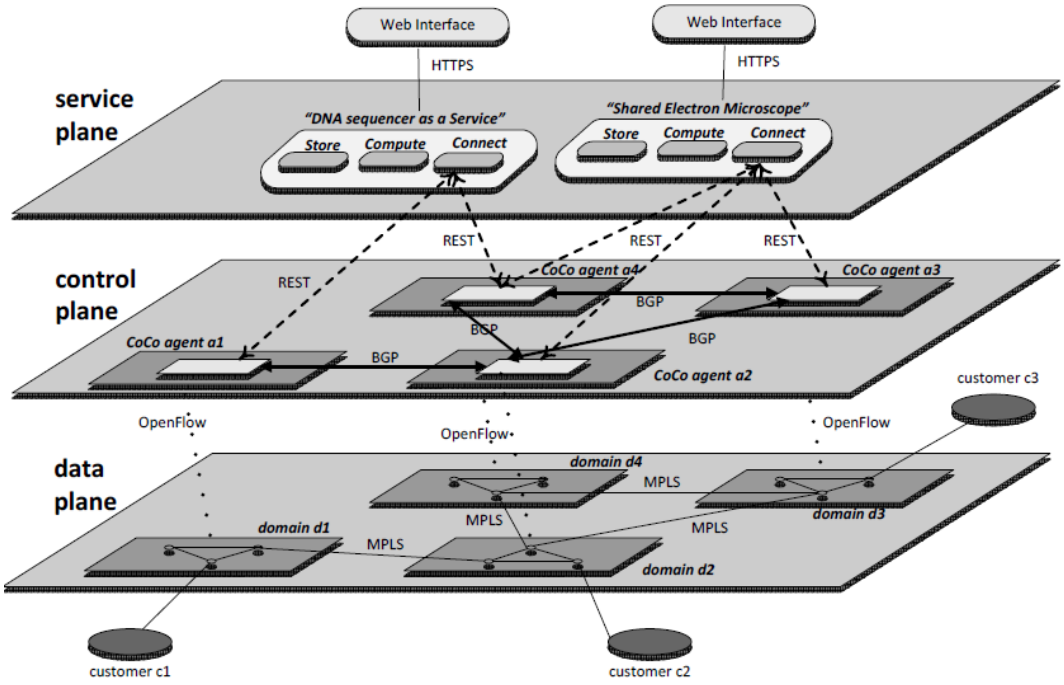


Figure 2: Schematic overview of use case deployment on CoCo service instances

2. CoCo use case workshop

In order to explore the demand in the eScience community related to the CoCo service a workshop was organized, with eight participants representing eight academic institutes. The workshop was held on January 21st 2014 at SURFnet's premises. After an introduction to the CoCo project the Genomics research project was presented by one of the participants. The presentation about this eScience project opened the discussion about the shortcomings of current networking services, as well as hopes and expectations from the future CoCo services. During this discussion and the subsequent interactive session with the workshop participants a number of suggestions, questions, constraints and expectations were exchanged in an open atmosphere.

Current eScience use cases and opportunity for improvement

- The "Large-Scale Population Imaging Studies" project presented by SURFsara outlines how brain scan slices are gathered from scanners at the Leiden University and Erasmus Medical Centers and the output data is forwarded for temporary storage and processing at the HPC cluster at SURFsara. After processing, the results from the data analysis can be inspected by experts from other research institutes. The amount of data produced by the MRI scanner is significant. For the transport of data from the campus where the brain scanner is located in the direction of SURFsara static light paths are used. In addition, Bandwidth on Demand (BoD) connections are used for visualization of the processed data.
- Another presented example was the "Genome of the Netherlands" project. In this project BoD connections are used to send and share DNA sequencing data between the participating institutes. The establishment of the twelve connected locations that constitute the communication infrastructure for this eScience project involved actions from many professionals and took a year and a half to complete. An important, time-consuming factor was the building of trust between collaborating parties. The established communication infrastructure is not used for other projects. Once the communication infrastructure will be phased out, it is expected that rebuilding such an infrastructure again for other projects will still take very considerable effort.
- For the Genomics project and other mentioned cases VPN and Bandwidth on Demand services are used. It is acknowledged that in most cases their application is limited to point-to-point connections. In addition the establishment of connections still requires considerable effort from network administrators, who need to be involved from the beginning. In this workshop specific network administration constraints and efforts were not addressed. In the follow-up of the use case activity, this issue needs further attention.
- In some cases the storage systems containing the scientific data are not connected to the broadband connectivity point from which data can be sent to collaboration partners. So in those cases even bringing data from storage to the connectivity point within a single domain can be problematic.
- In many cases classification of the confidentiality of data is unclear. For example, in the context of research involving medical data it is unclear if any, or what, legal framework is applicable. In practice one can assume that the persistent data will have to remain in one place, while copies or visual representations can be disclosed and transported to other sites.
- In the interactive session participants indicated that sharing of large scientific data sets is the primary challenge. Currently, in almost half of the cases large data sets are (almost) impossible to share with scientists in other domains. In addition, communication services supporting virtual meetings are used as collaboration tools. In most cases combinations of such data transport and virtual meeting services are used. At present day controlling scientific instruments remotely is rarely done. An exception is a jointly purchased electron microscope located at the campus of Leiden University and the LUMC. Even this instrument is most often used by scientists from the institutes where the microscope is located, because of the complexity of handling such an instrument and thus the required expertise. Although remote management of the microscope is possible, this option is only used to monitor long-running experiments from distant locations. For future eScience applications the CoCo

service is expected to be primarily used for scientific data, sensor and instrument sharing. In most cases additional communication services will be used as well.
- Realizing inter-domain connectivity for collaboration with international partners is mostly done ad-hoc. Workshop attendees indicated that for roughly half of the research projects the collaboration level is international.
- In Rotterdam the network domains of the Erasmus Medical Center (MC) and the Erasmus University are distinct. This has led to a different set of connectivity services; for example, Eduroam is provided for the Erasmus University while it is not supported for the Erasmus MC. Effort to align the supported services is planned as for future steps.
- The workshop participants indicate that the number of scientific institutions that jointly use a communication service for a particular eScience collaboration project is only a hand full in most cases. In roughly one out of ten cases scientists from more than ten institutions are involved in jointly utilizing a CoCo-like service for scientific collaboration.
- In practice, the demand for connectivity services from eScience researchers appears to be unpredictable. This is inherently due to the unpredictable nature of eScience. Although experiments can be planned, this does not mean that planning of successful results that are worth sharing with collaborating peers and the amount of produced data can be done. Once the scientific output has been generated successfully it should be possible to share the results as soon as possible. In some cases enormous amounts of data are generated and compute resources are required for the duration of a week and then remain unused for the following month. As a consequence, reserving high performance compute, high-volume storage and high-speed connectivity resources for longer periods of time is not efficient. This raises the need for on-demand and user initiated resource instantiation of virtual overlay networks on top of shared infrastructures.
- Using secure, public cloud services for large data processing is expensive in terms of bandwidth and often undesirable for other reasons such as data confidentiality and technical or commercial constraints regarding the data analysis software.

<u>Points of attention and constraints mentioned in the context of the CoCo service</u>

- The building of trust between collaborating parties is an important success factor.
- Due to data confidentiality one can assume that the persistent data will have to remain in one place, while copies or visual representations can be disclosed and transported to other sites.
- In the context of user initiated CoCo instances it is anticipated by the workshop participants that more than one, but a select group of users should be able to add / remove members (endpoint) to a CoCo instance.
- The opinions about required authorization methods for members, or groups of members, to access CoCo instances are more diverse. Almost all participants foresee that some kind of authorization control is required, but for example whether that should be group based or individual member based is unclear.
- For most applications of the CoCo service the number of interconnected domains in which endpoints reside will be less than ten. However, in some cases instances consisting of tens of domains are also expected.
- In terms of the number of endpoints of a CoCo instance will in most cases be around ten, although in one out of three cases instances with up to one hundred endpoints are also expected. Instances with more than a hundred endpoints are not foreseen by the workshop participants.
- There is no clear indication about the dynamic nature of the members (endpoints) of a CoCo instance. Some participants indicated that members would be added or removed from an instance at least once per two weeks, while others indicated that instances would typically be modified less than once per two months. In general, the dynamics of the instance members will strongly depend on the type of application for which the instance is used.
- In general, eScience researchers are interested in their scientific research and experiments, not in the compute, storage or connectivity resources that are used as tools to support their

research. Therefore, a CoCo service should be very easy and intuitive to use without requiring to study manuals and the service usage should be stable/reliable.
- Further, in eScience proposals the budget for IT and communication services is often underestimated.

Requested improvement by the CoCo service that were brought forth during the workshop

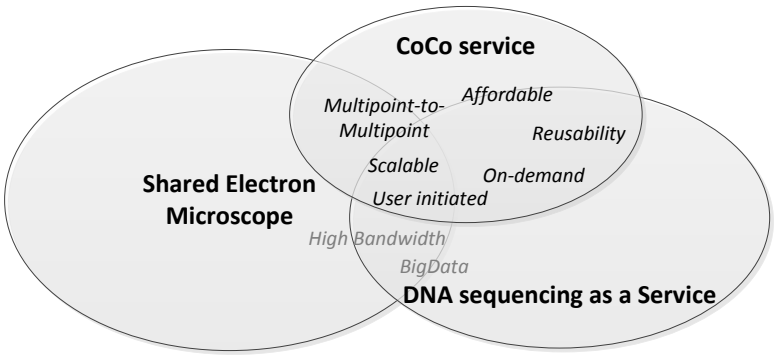
During the discussion about current communication and storage solutions for eScience projects a number of explicit or implicit requests for a future CoCo service were heard. Although not each of these requests can be satisfied by the CoCo service (for example because some of them are beyond the scope of CoCo), we list them below.

- A key challenge is the *on-demand* assembly of required data, compute and connectivity resources[1], with focus on scientific applications such as data, sensor or instrument sharing.
- *User initiated* connectivity service instances; proper installation of CoCo agents would still need to be supervised by network administrators.
- *Multipoint-to-multipoint* connectivity would in many cases add value to collaboration services.
- *Reusability* of connectivity service solutions, for example to transfer a solution from one network domain to another, or to reuse a solution for a subsequent, similar scientific project.
- For scientific research collaboration transfer of large data chunks (or more specifically, creation of temporary data copies: "cloud memory" as opposed to more persistent cloud-based storage) is becoming relevant more and more ⇔ CoCo service should support (and be *scalable* regarding) *broadband* communication interfaces, or even broadband link aggregation for large file transfer.
- The solution should be *affordable*, i.e. only a limited portion of the research budget will be available for the connectivity service which makes infrastructure sharing important.

[1] CoCo is primarily focused on on-demand connectivity, but this requirement illustrates that the CoCo instance should be deployable in platforms consisting of integrated, virtualized compute, storage and connectivity resources.

3. CoCo use case selection

The use cases that will be selected for further elaboration in the CoCo project should sketch a context of application for the CoCo service. The use cases should lend themselves for demonstrating and validating that the CoCo service improves sharing of existing eScience instruments, processing and communication infrastructure and services, on one or more of the aspects mentioned above. The on-demand, user initiated connectivity service that CoCo aims to provide covers (only) part of the use case solutions, as indicated in the diagram below.



Use case "DNA sequencing for third parties"[2]

Contributors: Wageningen University (WUR)/Plant Research International (Contact person:Sven Warris)

Description

The Wageningen University operates a DNA sequencing service for affiliated researchers from their own university and as a paid service for national and international research institutes. The amount of data that is generated in a run and the necessary compute capacity to analyse it, depends on the characteristics of the samples they receive.

Since the amount of storage and compute capacity that are available on-site is limited, the WUR needs flexible and trusted access to elastic compute facilities and/or a possibility to send output data to other locations. In some occasions confidentiality is important, depending on the nature of the samples and the requirements of the client/researcher. In the latter case restrictions will be imposed on the external storage and/or compute facilities.

Potential use case

In this use case there will be a DNA sequencing as a Service portal where scientists (users) can reserve a DNA sequencing service. The scientists can specify whether they want to receive the raw data from the sequence (and subsequently specify to which storage destination) or that they want to receive processed data. The users will be able to log in with the credentials of their home institute.

The operators of the service will be able to select compute facilities and/or storage destinations depending on the amount of data to be expected. The CoCo service is 'invoked' by the DNA sequencing service portal for establishing on-demand connectivity between the sequencer's storage device and the compute facilities and storage destinations specified by the customer. Once the data has been produced the CoCo instance is ready to be used for distribution of the data for processing and storage in (possibly) multiple domains.

[2] This use case is described in more detail in section 4.

An indication of endpoints and communication services involved in this use case are (this will be worked out in more detail in later use case actions):

- CoCo agent endpoints: WUR, SURFsara, (commercial) VM provider (e.g. Kentis?), other?
- Communication services / infrastructure: The existing Multi Service Port of the Wageningen University may be used for this purpose, with an additional OpenFlow enabled switch. SURFsara and additional VM providers may be connected via (on demand) lightpaths via NetherLight.
- Remarks:
    - Demonstrate the on-demand, multi-domain connectivity provided by the CoCo service, including flexible addition of destination endpoints.
    - Validate the stability of the CoCo service.
    - Functional and non-functional requirements need to be further discussed with the use case contributor.
    - Content level issues such as sharing of specific data selections, data specific confidentiality, etc. are out of scope for the CoCo project.

Use case "Shared electron microscope"

Contributors: LUMC (Contact person: Bram Koster)

Description:
The Netherlands Centre for Electron Nanoscopy (NeCEN) offers research institutes and companies access to advanced cryo-transmission electron microscopes. The two NeCEN microscopes are specifically designed to explore complex biological structures. The centre consists of a consortium of eleven institutes and the host location is at Leiden University.

The microscope generates a lot of data in the form of large images. The processing of these images is compute intensive. The software enables remote control for the microscopes, but this option is not necessary and not used often. In some occasions when experiments take several days, remote terminals are used for monitoring the experiment.

Potential Use Case
In this use case NeCEN offers an environment (portal) in which scientists can schedule experiments, and search/retrieve data for collaborative analysis. There is a central storage facility supplemented with external (cloud) storage. In the latter case, there is a need for a sustained low latency connection between the storage locations. Since the data is produced in a relative slow rate (compared to the capacity of wide area network connections) this connection doesn't have to be a high speed link. However, in some occasions when a lot of data from the external location has to be retrieved there is a need for temporary increased bandwidth.

An indication of endpoints and communication services involved in this use case are (this will be worked out in more detail in later use case actions):

- CoCo agent endpoints: Leiden University, LUMC? (LU and LUMC have a dedicated fiber connection for this purpose), other consortium members, (external) cloud storage provider.
- Communication services / infrastructure: All consortium member are already connected to SURFnet and most of them have a 10G Multi Service Port on SURFnet that can be used.
- Remarks:
    - In terms of demonstration of the added value of the CoCo service there is an overlap with the DNA sequencer as a service use case. When detailing the use cases it will be decided whether both use cases will be worked out, or only one of them.

4. Detailed use case: DNA sequencer as a Service

In this section we explore the DNA sequencer as a Service use case in more detail. After a short description of the context of DNA sequencing we will highlight the business drivers and thresholds in this scientific field. From that business viewpoint we will argue the added value of a DNA sequencer as a Service, and indicate the features that such a service should support. Subsequently we will decompose the technical challenges for creating such a service and pinpoint the contribution that the CoCo service can provide.

DNA sequencing

DNA sequencing and bioinformatics are essential methods in analysing the functions and structure of genomes, i.e. the genomics science field. At the Wageningen University & Research centre (WUR) a combination of three DNA sequencers are used as instruments: Illumina's HiSeq and MiSeq, and Pacific Biosciences' PacBio. Each of these sequencers has its own pro's and con's, making them more or less applicable for specific DNA sequencing tasks. For example, the MiSeq can sequence parts of DNA material in smaller batches, which can be useful in the preparation phase of a DNA sample, whereas the PacBio is more useful for sequencing long DNA reads. In some cases the analysis of a particular DNA sample requires the use of a combination of sequencers.

For the storage and processing of output from DNA sequencers specific bioinformatics solutions are used. The amount of data that a sequencer generates is typically large, depending on the size of the sequenced genome material and the sequencer. The PacBio typically generates hundreds of megabytes per sequencing run, while the HiSeq can produce up to hundreds of gigabytes of data in a run lasting over a week[3]. For some sequencers output data storage, including some post-processing, is located directly next to the sequencer and transportation of that bulk data via portable hard disks is not an unusual practice. Even internally within the WUR premises.

After the generation of the sequencing output several post-processing tasks can be performed. Examples of post-processing include read mapping and genome assembly (aligning and merging fragments of sequenced DNA to a longer DNA reference sequence or complete genome in order to reconstruct the original sequence), variant calling (for identifying deviations in a particular sequence region) and de novo assembly (genome assembly in case no reliable reference genome is available). All of these post-processing tasks are compute intensive and sometimes linear in the data size that needs to be processed. Read mapping can be parallelized to a high extent and is done on one of the WUR's HPC clusters or Keygene's life science grid, a commercial partner of the WUR. Variant calling is similarly compute intensive, but also more research intensive. It requires more iterative steps performed by the genomics scientist. De novo assembly is the most IT resource demanding processing technique. In addition to intensive computing it requires terabytes of memory (RAM) and is therefore deployed on special machines. In this context SURFsara is currently testing software developed for processing on a Hadoop cluster.

Inevitably DNA sequencing research involves manual tasks performed by genomics scientists. Automation of (relatively) standard tasks is part of current developments in bioinformatics. Progress is being made to introduce "pipeline" technology that is essentially a form of workflow management designed specifically to compose and execute a series of computational or data manipulation steps. Examples include the Galaxy[4] solution and the cloud based Tavaxy system. At the WUR bioinformatics experts use the Galaxy platform for workflow automation and Snakemake for command-line based pipelines.

[3] See for example [Comparison of Next-Generation Sequencing Systems, Lin Liu et.al., Journal of Biomedicine and Biotechnology, Volume 2012, Article ID 251364].

[4] See http://galaxyproject.org

Business perspective

The equipment used for DNA sequencing and bioinformatics is expensive and gets outdated relatively quickly, due to current rapid developments. For example, the investment cost for upgrading the WUR's HiSeq sequencer is approximately 600,000 Euro and this investment should be exploited within a three year time frame. Investments in the required computing software and equipment are also high and their applicability is, to some extent, dedicated to the purpose of genomics research. For research organisations the overall investment costs are high and can only be justified if the (re-)utilization of the sequencers and bioinformatics solutions is sufficiently high. The opportunity to offer DNA sequencing as a Service that can be consumed by scientists from multiple institutes can strongly improve this return on investment.

The DNA sequencer as a Service case indicated by the WUR clearly shows business potential for the WUR. However, this use case is also applicable for other institutes in the genomics science field. The use of DNA sequencers in The Netherlands is spread among more than a hundred expert groups from institutes that are gathered in the Dutch Techcentre for Life sciences (DTL)[5]. In general, technological improvements that continue to lower the cost of sequencing have strongly boosted the science of genomics[6].

DNA sequencer as a Service

The term "sequencing as a service" was already introduced in the previous decade, when the company Complete Genomics was founded[7]. Their initial business plan included a $500 million investment for building ten sequencing centres. In this use case we take a different business model for the DNA sequencer as a Service. In fact, to avoid the tremendous upfront investment in Complete Genomics' single provider case, we focus on a federative, multi-domain DNA sequencer as a Service model that re-uses sequencing capabilities.

An overview of the usage of the envisioned DNA sequencer as a Service is presented in Figure 3. It illustrates users of the service, typically DNA scientists, on the left hand side that interact with the service portal. On the right hand side the "back-end" DNA sequencing providers[8] are shown. Indicated are a number of user iterations with the portal that enable the user to configure a DNA sequencing experiment.

After authentication the user can specify details about the requested sequencing experiment. In the next step the service portal will assist the user in selecting appropriate DNA sequencing instruments. This is an important preparation step in DNA sequencing which is often ignored, when a project team only has access to a particular sequencer. Since a key objective of the DNA sequencer as a Service is to increase accessibility of multiple sequencers at different locations the DNA sequencer as a Service portal should support an appropriate sequencer selection interface. The selectable DNA sequencing instruments can be physical DNA sequencers, but the portal could also support interfaces to other online DNA analysis services[9].

The portal can also assist the researchers with preparing biological samples, find support for processing samples and inform the scientist on how to get the biological samples at the sequencing facility.

[5] www.dtls.nl

[6] en.wikipedia.org/wiki/Genomics

[7] www.nytimes.com/2008/10/06/business/06gene.html

[8] Note that this illustration presumes a multi-client domain and multi-provider context. However, the same service concept would apply for a single domain setting, where the indicated "providers" could just as well be individual pieces of DNA sequencing, processing and storage equipment. In fact, a single domain use case could be a shorter term application, whereas a multi-domain application could be the longer term objective.

[9] For example: via galaxyproject.org, or by off-loading processing to a commercial life science grid.

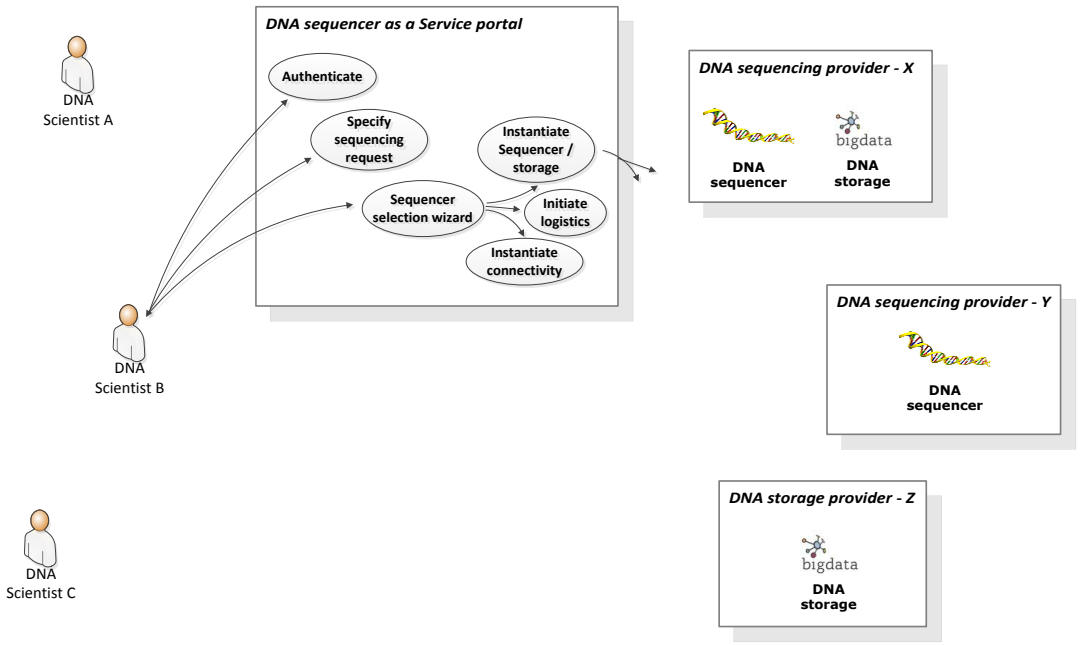


Figure 3: Using the DNA sequencer as a Service – service configuration

After initial configuration of the service the user is provided with an interface to activate and control DNA sequencing runs. Figure 4 presents an example of a series of interactions between the user and the sequencing instruments, via the portal.

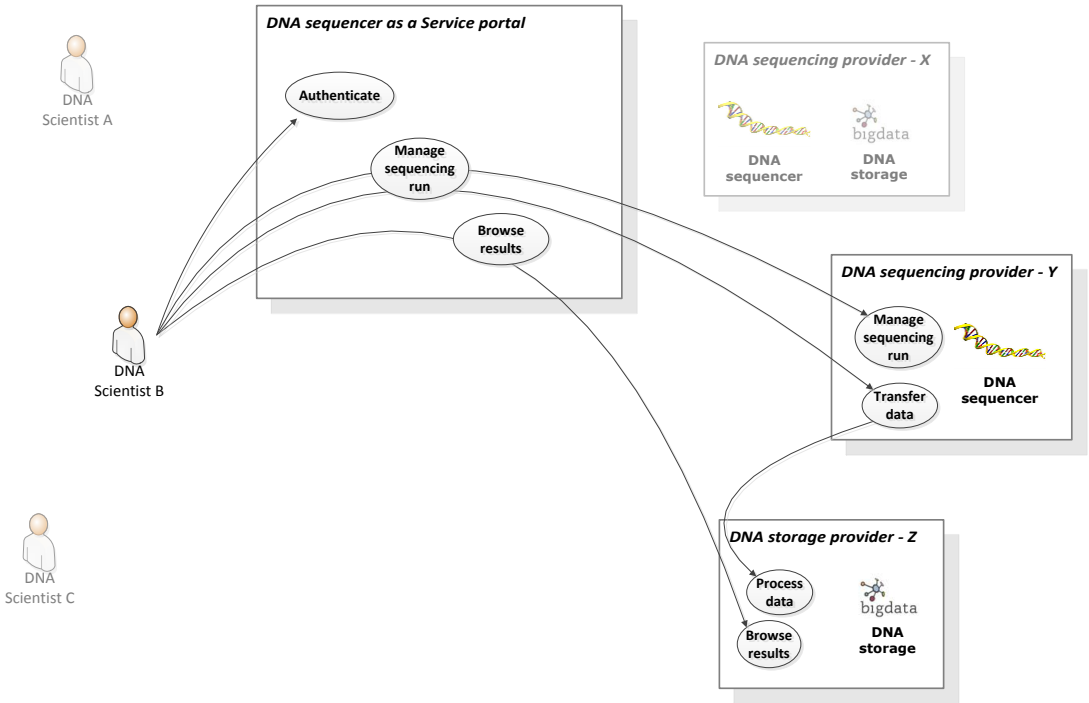


Figure 4: Using the DNA sequencer as a Service – remotely operating the DNA sequencer

One of the improvements of the portal is the ability for single sign-on authentication. Current practice requires scientists to login on multiple DNA sequencer, storage and processing systems. That can now be centralized in the portal. Further, the portal will support the user in managing the equipment, such as starting a run, viewing the run status and/or initiating the transfer of data from the DNA sequencer. In addition, the portal provides a useful platform for integrating pipeline technology to

enable genomics scientists to record and store their repeated manual tasks in workflow templates. This will contribute to the demand by DNA scientists for increased automation and ease-of-use. Finally, the DNA sequencer as a service concept offers the opportunity to provide a more unified look & feel for the user interface. For example, by embedding the visualisations provided by genome browsers in a standard format via the portal.

The success factor of these potential service portal improvements for the sequencing process depends on specific circumstances in the genomics research field that are out of scope for the CoCo project. Instead of elaborating on the precise benefits here, we focus on the (technical) feasibility and challenges of these service features in the following subsection.

(Technical) feasibility and challenges

Emerging workflow management solutions, such as the Galaxy platform, are focussed on automating DNA sequencing and processing processes. At their back-end such platforms will manage and interface to the resources that execute these processes. These resources include storage, processing and connectivity resources. Although the (integrated) management of the full range of these resources is beyond the scope of this project, the CoCo service is a good candidate for providing the on-demand, multi-domain connectivity between the resources. The foreseen identity management feature of the CoCo service will contribute to the ease-of-use by relieving end-users from the need to login on separate systems. Moreover, the CoCo service is being designed to be incorporated in future integrated resource management solutions.

Figure 5 and Figure 6 highlight how a DNA sequencing as a Service portal (either implemented with technology such as Galaxy or otherwise) can use the CoCo service. Figure 5 presents domains involved in the DNA Sequencer as a Service: the DNA scientist domain, the domain of the DNA sequencing as a Service portal and the premises of the DNA sequencing providers. The DNA sequencing providers can be multiple; in this case we illustrated a domain from which DNA sequencing services are provided and another domain providing storage and processing capabilities.

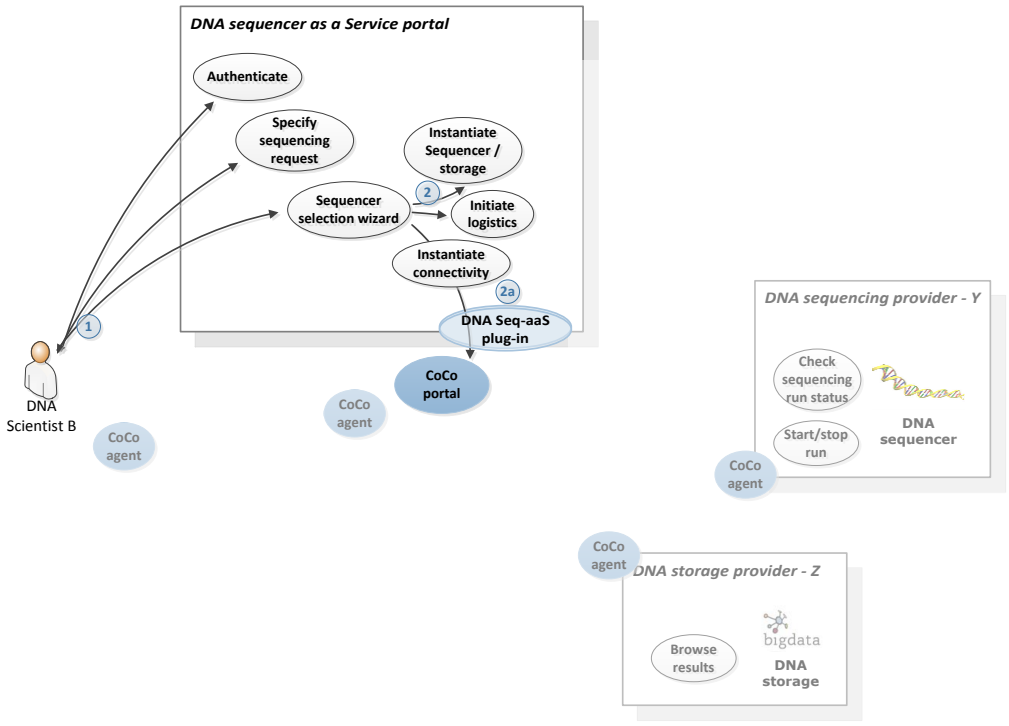


Figure 5: Connectivity initiation via CoCo

In the first step of the service initiation phase (see Figure 3 and Figure 5) the DNA scientist authenticates him-/herself via the DNA sequencing as a Service portal, specifies the sequencing request and selects the appropriate instruments. In this selection step (step 2 in Figure 5) the user also specifies the required connectivity. In the DNA sequencing as a Service portal a "plug-in" is used for that, as indicated by step 2a in Figure 5. This plug-in is a small software program in the DNA Sequencer as a Service portal that is the interface between the DNA sequencing as a Service portal and the CoCo portal in the service layer. In the connectivity initiation phase the plug-in stores application specific (meta) data containing, for example the user and DNA instrument names that can be selected in the portal by a DNA scientist[10]. Connectivity data that is used in the control and forwarding plane, e.g. IP address prefixes and VLAN IDs that are provided by the user, are inserted via the plug-in and stored in the CoCo portal.

In each of the domains the DNA Sequencing as a Service would require the installation and/or configuration by the network administrator of control and data plane components, including:

- OpenFlow switch(es),
- a VLAN per CoCo instance,
- a CoCo agent and
- some additional network configurations (e.g. VLAN tag insertion).

After installation of the CoCo connectivity for the DNA sequencer as a Service, there are no more actions required for the network administrators. In fact, (de)activation of CoCo instances can be done by users (or scheduled via the plug-in), such as illustrated for the DNA scientist in Figure 6. Via the portal the user can for example log-in to the portal and initiate a DNA sequencing run (step 1 in Figure 6). To this end, the DNA sequencing as a Service "plug-in" implements a script that can instruct the CoCo portal to (de)activate and modify CoCo instances via the REST interface to the CoCo agents (step 2).

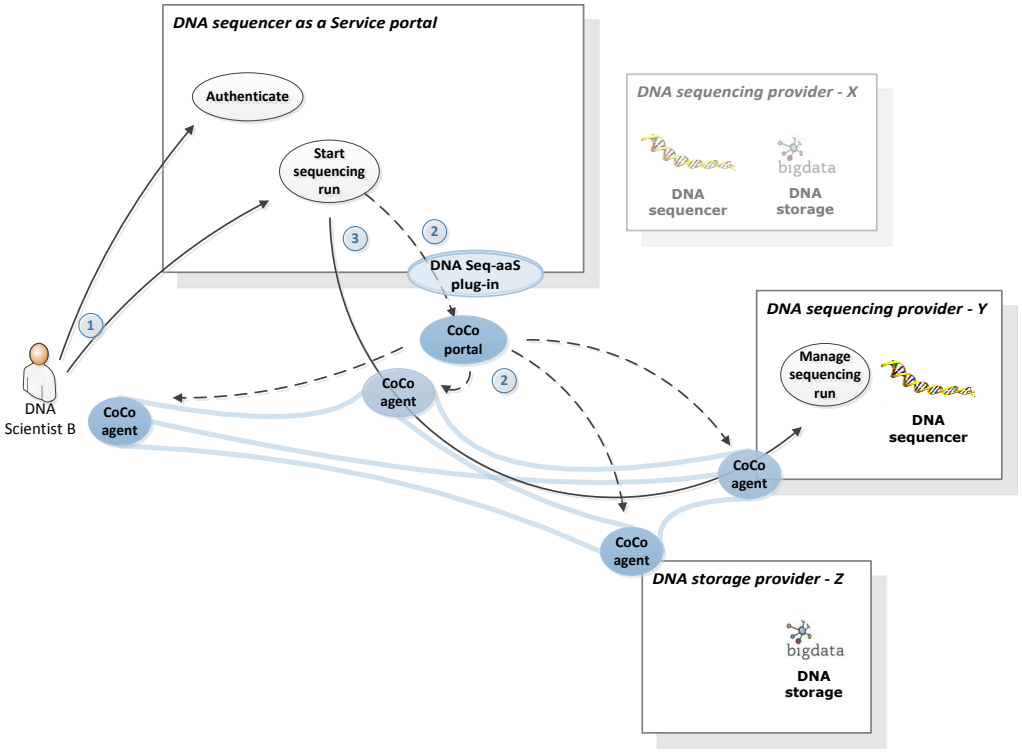


Figure 6: CoCo support for remote interaction

[10] Also other types of application specific information can be included. For example, a time schedule to (de)activate CoCo instances at specific time-slots.

The CoCo agent in each domain translates these instructions into forwarding rules on the local OpenFlow switches. Exchanging CoCo specific data between the agents is not a function of the portal plug-in. That exchange of information is done directly between neighbor CoCo agents.

Once the CoCo instance is instantiated this on-demand, multi-domain virtual private LAN can be used to forward the "Start sequencing run" command to the DNA sequencer (step 3 in Figure 6). The CoCo instance can be used for subsequent user interactions, until the user deactivates the CoCo instance.

We already mentioned that DNA sequencers can produce very large data volumes that need to be transported to the processors and the scientific end users. Optimal solutions are still lacking for the transfer of these large data sets. Although high bandwidth and BigData technology for transporting high-volume data is out of the scope of the CoCo project, the CoCo service can be deployed to establish user (or platform) initiated and multi-domain connectivity over broadband links.

Apart from the technical feasibility and challenges there are other obstacles that can prevent successful innovation via a DNA sequencer as a Service. One of them could be reluctance of DNA scientists to the phenomenon of outsourcing of DNA sequencing itself. However, this is unlikely, because outsourcing of DNA sequencing is common practice. For example, the WUR's PacBio sequencer is currently located at Keygene's premises. In addition, it is common practice to outsource sequencing runs of a more repetitive nature to Chinese institutes[11]. This increasing experience with the logistic and legal aspects of outsourcing is encouraging that the concept of a DNA sequencer as a Service will not be hindered from this perspective.

A similar, non-technical aspect of DNA sequencing is the willingness of scientists to adopt the concept of a DNA sequencer as a Service, instead of their current hands-on way of working. This should be investigated further (for example via a proof of concept), but the ongoing trend of workflow automation appears to bring the end users sufficient benefits to adopt such technology.

Further, in some cases DNA sequencing produces confidentiality sensitive data, in particular when the sequencing samples concern human DNA or when the DNA sample is analysed for commercial purposes. The (federated) identity management and virtual private networking principles included in the design of the CoCo service provide the means to realise the required confidential connectivity.

[11] Note, that Complete Genomics was bought by the Chinese genomics organization BGI-Shenzhen in March 2013 [www.completegenomics.com/news-events/press-releases/BGI-Shenzhen-Completes-Acquisition-of-Complete-Genomics-198854331.html].

5. Conclusion and next steps

During the workshop the participants from the eScience community provided a number of suggestions, constraints and requests for use cases. Not all of the requests are in the scope of the CoCo service. For example, specific BigData solutions and the demand for combining data, compute and connectivity resources is beyond the scope of the CoCo project.

Added value of the CoCo prototype service is foreseen in terms of improvement of current multi-domain connectivity services regarding:

- *on-demand* connectivity service and enabling composition with on-demand data and compute services;
- *User initiated* connectivity service instantiation;
- *Reusability* of connectivity service solutions;
- Supporting *scalable, broadband* (and *aggregation* of) communication services;
- The solution should be *affordable* (exploiting infrastructure sharing / overlay).

The elaborated DNA Sequencer as a Service use case illustrates the business value of the CoCo service in the field of expensive and rapidly out-dated eScience instruments. Sharing of DNA sequencers by eScientists from multiple (inter)national institutes is a key to the progress in that research field. The CoCo service enables easy-to-use connectivity functions that are required for creating a DNA Sequencer as a Service. The elaborated use case clarifies how the CoCo service can be used as the connectivity pillar in such a collaborative eScience service.

The next steps regarding the use case activity are:

- In the following use case actions specific attention will be paid to the feasibility and effort needed by network administrators to install a CoCo agent on OpenFlow switches and apply appropriate network configurations.
- Also, authentication and confidentiality requirements will need to be addressed in the final use case activity.
- Once the CoCo prototype has been finalized some demonstration version of the DNA sequencer as a Service portal can be realised. Also a discussion can be planned with the broader DTL community about a more elaborate business case and the idea to launch a national service portal.
- Finally, using the intermediate use case milestone representatives from the IUAC will be requested for their feedback on the added value of the CoCo portal in their eScience fields.

Appendix: results from interactive session during the CoCo use case workshop

The following sheets present the outcome of responses from the workshop participants during the interactive, electronic voting session (in Dutch). For each use case question a number of options were given and each participant selected one of the options. In addition to the discussions these responses provide insight in the potential use of the CoCo services.

Met wie wordt in uw instelling / project samengewerkt?

1. alleen mensen binnen mijn eigen instelling	10%
2. ook met anderen in Nederland	40%
3. ook met anderen buiten Nederland	50%
4. weet niet	0%

Uit hoeveel instellingen komen de leden waarmee via eInfrastructuur wordt samengewerkt?

1. Meestal 2	20%
2. Vaak een hand vol	60%
3. Zo rond 10	10%
4. Soms ook wel meer dan 10	10%

Wat voor type eInfrastructuur toepassingen worden daarbij veel gebruikt?

1. Data op afstand	30%
2. Samenwerking op afstand	20%
3. Instrument / sensor op afstand	0%
4. Meer van bovenstaande	40%
5. Anders, namelijk ...	10%

Hoe beheert u grote datasets?

1. Die data sets zijn niet of moeilijk op afstand benaderbaar	44%
2. Die staan op een centrale plek en kunnen makkelijk op afstand benaderd worden	22%
3. Die staan op meerdere plekken en kunnen makkelijk op afstand benaderd worden	22%
4. Weet niet	11%

Gebruikt u virtuele machines / cloud?

1. Ja, en die staan allemaal op 1 plek	12%
2. Ja, wij gebruiken VM's op meerdere locaties	50%
3. Nee, wij gebruiken geen VM's	38%
4. Weet niet	0%

Hoe vaak denk u leden aan uw community toe te voegen of te verwijderen?

1. Tenminste 1x per 2 weken	25%
2. Tenminste 1x per 2 maanden	12%
3. Minder vaak	25%
4. Weet niet	38%

Hoeveel leden zouden (meestal) toegang hebben tot een CoCo instantiatie?

1. Hoogstens 5	22%
2. Ongeveer een 10-tal	44%
3. Zo rond de 100	33%
4. Soms ook ruim meer dan 100	0%

Wie zouden leden moeten mogen toevoegen of verwijderen?

1. alleen één projectleider	0%
2. een beperkte groep mensen	90%
3. iedereen in het project	10%
4. weet niet	0%

Hoe wenst u toegang binnen uw groep te regelen?

1. alle leden mogen overal bij	11%
2. ik wil leden kunnen toevoegen aan groepen en die groepen al dan niet toegang verlenen	33%
3. ik wil per lid kunnen bepalen waar die wel en niet bij mag	44%
4. anders, namelijk …	11%

Voor welk type toepassing zou CoCo het meest nuttig zijn?

1. Eduroam++	0%
2. Multi-domein communicator	0%
3. Instrument / sensor / data op afstand	38%
4. Meerdere van bovenstaande	50%
5. Anders, namelijk …	12%

….