Content

1 Introduction

The focus of work package 7 (WP7) of the NFU Data4LifeSciences program is on 'facilities for high-throughput data processing'. WP7 started February, 2015.

The need for efficient data processing facilities increases with the growing number of use cases with high volume data sets and multi-center study collaboration, e.g. in the genetics and medical imaging research areas. Use cases range from small-scale studies up to a typical cohort study of 5,000 DNA samples requiring more than 1,500,000 core hours (171 years) and over 500 Terabytes for GWAS analyses. High-performance computing and networks enable high-throughput data processing in life science & health research, providing researchers a head start in competitive research areas.

Requirements vary per use case, bringing in a challenge when aiming to address all use cases with a single type of compute infrastructure. Different types of infrastructures are required in terms of data I/O performance, capacity, job scheduling and data transfers. This work package deals with identifying best practices, in particular for the large scale & multi-centre research projects such as BBMRI.

The WP7 team consists of research support staff from AMC, VUmc, LUMC, UMCG, UMCU, as well as ICT infrastructure advisors and community managers from SURF (SURFsara and SURFnet).

2 Goal

2.1 True north

The ultimate goal of WP7 is to create a secure, cost-efficient, easily accessible and scalable data processing infrastructure for end-users in life science & health research without borders between local and national data processing services. End users are able to use HPC capacity on demand, based on a clear governance and cost model, that gives access to shared data and synchronized analysis tools directly available at the processing facilities.

2.2 Project goal

Multi-site studies such as supported by BBMRI cope with shared data sets that need to be analyzed at multiple sites, with shared pipelines and software configurations.

Data4Lifesciences WP7 aims to harmonize the UMC and national compute infrastructures and thus to create a data analysis infrastructure that allows the efficient and secure use of local and national compute and data resources. This includes:

- a data analysis infrastructure that allows each researcher to use national and/or UMC compute clusters;
- easy access to each other's and / or national HPC clusters to meet peak load computing capacity needs;
- reduce overhead in user accounting when giving access to each other and national clusters: central AAI;
- sharing best practices on using processing platforms for a diversity of use cases;
- sharing of pipelines, software configuration and installation packages on processing platforms;
- standardization and/or joint purchasing of hardware and services, reduce costs where possible.

In this phase (2015-2016) WP7 aims to advise on technical solutions based on system and user requirements. We perform pilots to build up experience, show proof of concept, and evaluate available and innovative solutions in order to be able to advise on a workplan for the implementation phase: 2017-2018.

We review and test the following compute scale out models (Table 1a) and connectivity and access models (Table 1b). Testing these models can lead to potential components of a federated compute and data infrastructure. Except for the Grid infrastructure, that is currently in production, pilots are initiated to test these models (see phase Proof of Concept) and learn about its ad/disadvantages. Note that the Grid infrastructure provides the right technology but provides insufficient usability for the life science research community.

Model	Pilot description	Reference	Phase
(1) Federative cloud environment	Use of Virtual Machines on distributed cloud environment - EGI federated cloud - Community cloud burst - Research cloud	https://www.egi.eu/infrastructure/cloud/	Proof of Concept
(2) Grid infrastructure	Use of grid middleware (EGI) for job submission on distributed compute clusters	https://surfsara.nl/project/life-science-grid	Production
(3) Cluster infrastructure with federative ID management	Using each other's cluster without middleware but with federative identity management	-	Proof of Concept
(4) Hybrid cluster-cloud model	Extend a local cluster with virtual nodes in a cloud environment	-	Proof of Concept

Table 1a. Scale-out models for data processing

Model	Pilot description	Reference	Phase
(5) Federated ID management	Using SURFconext or other identity management systems for a single-sign on at Dutch compute facilities	-	Proof of Concept
(6) Research LAN	Secure environment to exchange data and VMs	-	Proof of Concept

Table 1b. Connectivity and access models

3 Activities

3.1 Pilots

Technical pilots are initiated to cover different aspects of the shared data and compute infrastructure:

- Scale out data processing (1, 3, 4)
- Access & identity management: (1, 3, 5)
- Fast, secure and efficient data sharing (6)

3.2 Governance and provisioning

Besides the technical solutions, operational models of federation are investigated in terms of

- governance between institutes: unbrokered, brokered, collaborative
- architecture level of federation: datacenter, IaaS, PaaS, SaaS, services, experts/people
- terms of collaboration: SLA – friends-service, joint procurement of products and services

as well as finance models. Currently, low available research budgets lead to personal or departmental hardware purchases rather than local or central facilities. Together with national e-infrastructure providers (SURF, DANS, 3TU) policies for funding are being discussed with funders like NWO and ZonMw.

Characteristics are very different for each use case, which is a challenge when wanting to address all use cases with a single type of compute infrastructure. Thus, different types of infrastructures are required in terms of data I/O performance, capacity and job scheduling. Cloud research workspaces are available at VUmc and UMCG. The focus of the WP7 workgroup is on large scale & multi-centre research projects such as BBMRI

Challenges:

- Financial models in research grants.
- Service delivery of UMC cluster to external partners to exchange resources
- Costs between national and local facilities are often not comparable as support staff and electricity is excluded in the price towards researchers
- HPC becomes increasingly commodity, and requires different service delivery models, e.g. outsourcing becomes financially interesting

3.3 Deliverables

ID	Deliverable	Task lead
WP7.1	***Project plan, requirements and reports***	
7.1.1	Inventory of current UMC HPC facilities incl. maintenance, architecture, access	Irene
7.1.2	Criteria for harmonized HPC infrastructure model (e.g. scalability, access, exchangeability, performance)	Irene
7.1.2a	Technical requirements	Irene
7.1.2b	User & system requirements	Irene
7.1.2c	Mapping of user & technical requirements	Irene
7.1.3	Use cases	Irene
7.1.3a	Inventory use cases: most used big data analysis protocols medical imaging and genetics	Irene
7.1.3b	Description of use cases	Irene
7.1.4	Inventory of access models, AAI technologies	Paul
7.1.4a	Description of law and regulations regarding security	Jan Jurjen
7.1.4b	Requirements for a private analysis workspace in multi-center studies	Arnoud, Harry

ID	Deliverable	Task lead
7.1.5	Report PoCs scale out & connectivity & access based on pilots	Irene
7.1.6	Choice for a scale-out HPC infrastructure model, possibly per use case scenario.	Irene
7.1.7	Inventory of relating (European) projects that work on federative infrastructures, and way of collaboration	Irene
7.1.8	Implementation plan for a scalable HPC infrastructure model	Irene
7.1.9	Service and financial analysis of current (local and central) HPC facilities	Irene
7.1.10	Policy on finance streams for compute facilities (in collaboration with LCRDM).	Irene
7.1.11	Federation models: governance, level (IaaS, SaaS, ..), service level	Irene
7.1.12	Interim report (sept 2015, December 2016)	Irene
7.1.12a	version September 2015	Irene
7.1.12b	version December 2016	Irene
WP7.2	***Evaluation of current scale-out models for data processing***	
7.2.1	Description of current models	Silvia
7.2.2	Evaluation of EGI Grid infrastructure: lessons learned, best practices	Silvia
WP7.3	***PoC Scale out: federated cloud***	
7.3.1	Inventory of lessons learned & best practices based on current (European) projects	Pieter
7.3.2	Description of concept federated cloud model	Pieter
7.3.3	Test VMs on cloud UMCG, SURFsara HPC Cloud	Pieter
7.3.3a	Standard Virtual machines (VM) available for WP7 use cases (7.1.3)	Pieter
7.3.3b	Performance tests	Pieter
7.3.3c	Report on performance on network, compute	Pieter
7.3.4	Report and evaluation of pilot based on criteria federative HPC infrastructure model	Pieter
7.3.5	Implementation plan and costs	Pieter
WP7.4	***PoC Scale out: community cloud burst***	
7.4.1	Description concept community cloud burst	Christiaan
7.4.2	Test VMs on research cloud VUmc and SURFsara HPC Cloud	Christiaan
7.4.2a	Standard Virtual machines (VM) available for selected applications	Christiaan
7.4.2b	Performance tests	Christiaan
7.4.3	Report and evaluation of pilot based on criteria federative HPC infrastructure model	Christiaan
7.4.4	Implementation plan and costs	Christiaan
WP7.5	***PoC Scale out: hybrid cloud-cluster model***	
7.5.1	Description concept hybrid cloud-cluster model	Jeroen
7.5.2	Test performance scale-out model	Jeroen
7.5.2a	Test pipelines on infrastructure LUMC with scale out to SURFsara HPC Cloud	Jeroen
7.5.2b	Link queueing engines via API (LUMC, SURFsara + light path)	Jeroen
7.5.2c	Test automatic scale-out via API @ LUMC	Jeroen
7.5.2d	Test pipelines op infrastructure UMCU with scale out SURFsara HPC Cloud	Patrick
7.5.2e	Link queueing engines via API (UMCU, SURFsara + light path)	Patrick

ID	Deliverable	Task lead
7.5.2f	Test automatic scale-out via API @ UMCU	Patrick
7.5.3	Report and evaluation of pilot based on criteria federative HPC infrastructure model	Jeroen
7.5.4	Implementation plan and costs	Jeroen
WP7.6	***PoC Authentication and Authorization: federated ID management***	
7.6.1	Inventory of use cases (LUMC, UMCG, UMCU, UU and SURFsara) and describe possible solutions for federative authentication via SURFconext	Patrick
7.6.2	Development PoC code to generate ssh keys based on federative UMC account (SAML)	Patrick
7.6.2	Start implementation PoC code at UMCU, LUMC and UMCG/RUG. Parallel trajectory for federation to SURFsara resources - ssh based	Patrick
7.6.3	Pilot PoC implementations	Patrick
7.6.4	Wrap up results Go/NoGo federative authentication research resources in production	Patrick
WP7.7	***PoC Connectivity: E-LAN***	
7.7.1	Describe use cases, policies and possible technical solutions	Paul
7.7.1a	Description use case and technical implementation	Paul
7.7.1b	Policy framework	Paul
7.7.2	Acquire UMCs for contribution in E-LAN trajectory	Paul
7.7.3	First draft on collaborative research network architecture: "the UMC E-LAN"	Paul
7.7.4	Implementation: link local research network UMCs to E-LAN and SURFsara HPC cloud. (LUMC-SURFsara)	Paul
7.7.5	Test PoC at more UMCs (UMCG, UMCU, others?)	Paul
7.7.6.	Implementation plan and costs	Paul
7.7.7	Requirements Research LAN UMCs	Paul
7.7.8	Wrap up results E-LAN PoC, Go/NoGo E-LAN in production	Paul

4 Project team

4.1 WP team

Reporting structure

- Jan Willem Boiten (program manager NFU Data4Lifesciences)
- Anwar Osseyran (mentor NFU D4LS WP7)

NB. Results of this workpackage are being discussed in the NFU Data4lifesciences program committee and operational board.

Projectleader

- This project is led by workpackage leader WP7, Irene Nooren (SURFsara)

Project team

The project team consists of members of the UMCs, SURFsara and SURFnet:

- UMCU: Patrick Kemmeren
- LUMC: Jeroen Laros, Michele Huijberts
- UMCG: Pieter Neerincx, Hans Gankema
- AMC: Silvia Delgado Olabarriaga
- VUmc: Christiaan Geertsma
- SURFsara: Irene Nooren, Natalie Danezi, Gerben Venekamp, Rens Groenewegen, Maarten Kooijman
- SURFnet: Paul van Dijk

The input from Paul van Dijk and Gerben Venekamp is matched with activities in the AARC project.

Within the subgroups the following people are involved:

Part WP	Project team	Other resources required (UMCs)
WP7.1 Harmonisation plan & requirements	All, Irene Nooren (lead)	Projectleader
WP7.2 Scale out: PoC federated cloud	Pieter Neerincx (lead) Irene Nooren (SURFsara) Hans Gankema, Fokke Dijkstra (RUG-CIT) Natalie Danezi (SURFsara) Maarten Kooijman (SURFsara)	Technical advisor IT manager
WP7.3 Scale out: PoC community cloud burst	Christiaan Geertsma (VUmc, lead) Ander Astudillo (SURFsara) Natalie Danezi (SURFsara)	IT manager Technical advisor Researcher
WP7.4 Scale out: PoC hybrid cloud-cluster model	Jeroen Laros (LUMC, lead) Patrick Kemmeren (UMCU) Maarten Kooijman (SURFsara)	IT manager
WP7.5 Connectivity & access: federated ID management	Patrick Kemmeren (UMCU, lead) Jeroen Laros (LUMC) Gerben Venekamp (SURFsara) Paul van Dijk (SURFnet)	Community manager Technical advisor
WP7.6 Connectivity & access: research LAN	Paul van Dijk (SURFnet, lead) Jeroen Laros (LUMC) Patrick Kammeren (UMCU) Pieter Neerincx (UMCG)	Community manager Technical advisor

4.2 Stakeholders in other WPs

Input from other workpackages is needed as follows:

- WP2, WP4: input for deliverable 7.1.2. system requirements, UMC reference architecture that needs to match the data processing architecture.
- WP4: input for deliverable 7.1.4. requirements security and privacy for data processing infrastructure.
  See also the paragraph ‘Relation to other projects’.

5 Risks

Risks include:

Risk	Measure	Action
Insufficient funds available to move ahead with the plan in 2017	High	Each UMC has their own budget cycle. Align these cycles and/or get a clear view on possibilities for proposals
Required manpower is not (sufficiently) available	Medium	Plan resources. Get a clear view on the resource management processes at UMCs
Required resources (hardware, software) for testing are not (sufficiently) available	Medium	Plan in advacne
Legal obstacles for data sharing	High	Include WP6
User groups not sufficiently involved – no adoption by users	Medium	Get end-users involved as well

6 Relation to other projects

This project has a relation with the following projects and uses the following communication channels to keep each other informed and initiate collaborative actions when needed:

- NFU D4LS WP1 & WP6 (good research practice)
  - Solutions as developed by WP7 should be complaint to regulations as collected by WP1 & WP6
- NFU D4LS WP2 (architecture)
  - the solutions within WP7 need to fit into the reference architecture as provided by WP2
  - Irene is part of the WP2 project team, as well as results are shared during the NFU D4LS Operational Board meeting
  - Key contact for communication from WP2: Jeroen Belien
- NFU D4LS WP4 (biomedical data sharing & analysis):
  - Use the case cases from WP4 also for WP7
  - the solutions within WP7 need to able to easily connect to the workspace that is being developed within WP4
  - As a representative of WP7, Irene is part of the WP4 project team. Also, results are shared during the NFU D4LS Operational Board meetings
  - Key contacts for communication from WP4: Harry Pijl, Arnoud vd Maas
- SURF SIG Compute resources for life science research
  - Share HPC expertise and lessons learned
  - This SIG is on the agenda of the WP7 project team
- ELIXIR EXCELERATE:
  - develops European infrastructure for life science research
  - Irene is part of the ELIXIR EXCELERATE NL team, and takes part in WP4 ‘data and compute’
- EUDAT, PRACE
  - European research infrastructures for data and compute
  - Irene gets updates from this project via SURFsara and will bring in updates in NFU D4LS when appropriate.
- BBMRI2.0: infrastructure for biobanking
  - Jeroen Laros (LUMC) and Irene Nooren (SURFsara) are involved in BBMRI-NL WP5: ‘data and compute’.
- SURF Research Cloud: In this project a federated cloud IaaS/PaaS/SaaS is under investigation for the research in the Netherlands
  - Irene gets updates from this project via SURFsara and will bring in updates in NFU D4LS when appropriate.
- Landelijk coördinatiepunt research data management, workgroup infrastructure facilities.