NGI H2020 Profile

**NGI CZ**

11-05-2014

# Target user communities

* *Provide information here about the* ***top three*** *target international user communities (Research Infrastructures of the ESFRI roadmap, other international research collaborations and projects) that are part of your NGI strategic user engagement roadmap*
* *Provide information about how resources (data, storage,…) will be made available in your NGI to the community and according to which policy*

|  |  |
| --- | --- |
|  | Research Community/Project description (list in order of descending priority) |
| Community 1 Biological and medical sciences | NGI\_CZ is an official member of ELIXIR CZ Node, which is a joint project of thirteen institutions in the Czech Republic.  Within the ELIXIR CZ, we have a long history of close cooperation with CEITEC (Central European Institute of Technology, a large national facility; www.ceitec.cz). We focus on its 3 research projects:   * *structural biology* – extends beyond ELIXIR, encompasses Czech part of INSTRUCT and EuroBioImaging * *genomics and proteomics* * *brain and mind research* – connected to The Virtual Brain project ([www.thevirtualbrain.org](http://www.thevirtualbrain.org))   Recently, we signed Memorandum of Understanding with a national node (BBMRI\_CZ) of the large research biobanking infrastructure– national node of BBMRI ESFRI project. |
| Community 2  Physical sciences and engineering | We closely cooperate with Institute of Physics of the AS CR:   * *laser physics* - Extreme Light Infrastructure (ELI) http://www.eli-beams.eu/), the unique ESFRI laser centre located in the Czech Republic. * *elementary particle physics* - big international collider experiments – ATLAS and ALICE at the LHC collider at CERN and NOVA at Fermilab, * *astroparticle physics* – the Pierre Auger Observatory experiment in Argentine and Cherenkov Telescope Array (CTA, https://www.cta-observatory.org). While PAO in Argentina already collects data for several years and plans upgrade until 2022, the CTA projects is under construction.   Recently we started cooperation with astrophysicists in the Department of Theoretical Physics and Astrophysics on Masaryk University. We host mirror of the large NASA Exoplanet Archive (WASP achive) for SuperWASP consortium http://exoplanetarchive.ipac.caltech.edu/docs/SuperWASPMission.html and help them to build web interface for searching the database. |
| Community 3  Environmental sciences | We cooperate with CzechCOS/ICOS – the national centre for studying the impact of global climate changes. It became the Czech node for ICOS, EUFAR, and LIFEWATCH ESFRI project. Our collaboration covers both resource (compute and data) provisioning and support for efficient use of the infrastructure. Together with the national Centre CERIT-SC we are also involved in joint research activities (e.g. analysis of LiDaR data). |

# Resource provisioning for target communities

* *For each of the* ***top three*** *communities provide information on the resources that nationally will be available and the related policies and cost model, as applicable*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Compute and storage capacity currently available (or available in the future) to deal with the data growth** | **Access policy** | **Available funding or funding models (present and future)** | **What existing resources the e-infrastructures can offer, their current usage, the limitations and plans to deal with the data deluge** |
| **Community 1**  **Biological and medical sciences** | **Computing capacity** – the community actually consumes approx 50% of NGI\_CZ total computing capacity (grid and cloud). The part of the capacity is owned by the community (approx. 900 cpu cores) and partly is supported from the Czech national e-infrastructure resources purchased from European structural funds (approx. 2500 cpu cores).  The community plans to purchase over 1000 own cpu cores in very close future.  **Storage capacity –**  Nowadays, the community uses NGI storage resources (hundreds of TB). They plan to purchase 200 TB of own shared networked storage for work data and a few PBs of high-capacity data storage (HSM). | Access to all the resources (hardware and software) and services is available for free to all researchers and students belonging to Czech universities and research community in general. This is analogous to the way access is granted to the academic network in the Czech Republic. This access does not need any formal pre-allocation. Important users are prioritized (their limits expanded) in an automated way, based on evaluation of their research results achieved with the use of NGI resources. The resources are also available to cooperating users from abroad, through sponsorship of a Czech research teams already using the resources. | The NGI resources were purchased from European structural funds and are available for free according to the NGI access policy.  Sustainability will be covered mainly from national sustainability programs and large infrastructure projects.  User support staff can be partly funded via joint research projects, in which teams of NGI employees and students together with user community staff are formed, working on specific problems of user community. | In the following years, mainly related to the emerging of new activities of RD&I projects, a significant deluge of experimental data is expected, produced by new digital instruments (e.g., modern genome sequencers).  Therefore we help user communities to develop new and/or modify known algorithms in order to efficiently use the multi-core computation environment.  At the same time, the national infrastructure sustainability programs plans the constant e-infrastructure funding, so the increase of the computing and storage capacity according to Moore‘s law can be expected.  Beside this, the archival data can be moved to one of three CESNET’s high capacity storages.  Part of the required computing capacity can be also covered in the national supercomputing center IT4Inovation, which closely cooperate with PRACE project. |
| **Community 2**  **Physical sciences and engineering** | **Computing capacity** Most of the capacity is owned by the community (approx. 4000 cpu cores), partly is supported from the Czech national e-infrastructure resources purchased from European structure funds (approx. 400 cpu cores).  **Storage capacity –**  Nowadays, the community uses own storage resources 3 PB and approx. 100 TB is provided by NGI. |
| **Community 3**  **Environmental sciences** | The community actually consumes approx. approx. 500 cpu cores. The computational resources consumption is expected to in accordance to extended climatology studies planned.  The user support is focused on joint research in the field of environment, already with excellent scientific results.  Storage requirements are currently formulated, for climatology studies they expects need in the range PetaBytes of stored data. |

# User support skills

|  |  |
| --- | --- |
|  | User support skills and related technical and disciplinary areas |
| Training and education | For our users (current or potential) we organize:   * CESNET DAYS on academic institutions, informing about e-Infrastructure activities in the Czech Republic. * Practically oriented hands-on meetings for user communities starting with grid or cloud (for small user groups). * Once a year a national workshop/conference for all current NGI members to inform about news in grid and cloud * Advanced training course on specific application optimizations   In the future, we plan to organize user community workshops (e.g. for life science or environmental sciences). |
| Technical skills | We support individual users and user communities by porting application; nowadays we have hundreds of applications from tens research areas  <https://wiki.metacentrum.cz/wiki/Kategorie:Applications> in the national grid.  At the international level we currently provide middleware support in EGI. |
| Discipline/user-specific skills | Apart from running resources, the NGI role lies in the close collaboration with end users and research teams, where it either continue in already established or develops new research collaboration.  NGI users are also research partners, participating on the research and development activities. Both short term collaboration over a specific problem and especially a long term research collaboration (including joint third party funded projects) is supported.  NGI’s researchers collaborate with user communities on continuous improvement of services, methods, and tools, and they develop new algorithms, programs, forms and means to use the computing and storage capacities, regardless whether NGI‘s own or belonging to other parties, including end users themselves.  We work closely with the national centre CERIT-SC, Institute of Computer Science and Faculty of Informatics, all parts of the Masaryk University, including extensive involvement of pre/post graduate students into the NGI’s research activities. We share experience (and personnel) with CERIT-SC, to guarantee a smooth provision of the national e-infrastructure to end users.  Through the collaborating CERIT-SC centre we can mention (as a reference) participation in European projects Thallamos and SDI4APPS (with previously described cooperation process), a few national projects based on TACR or GACR funding, and other local projects.  We prepare a new project for H2020 call -- Oncosteer |
| Supporting Grid users | CESNET has been a lead partner in EGI's user support activity since 2010. CESNET coordinates 1st and 2nd level support in the EGI, provides 2nd level support for selected products, and, more recently, also provides 1st level support for issues incoming to EGI. |

# Software development skills and experience

* *If interested in participating to software development/integration activities, list here the software development skills available in the organizations from your NGI and the experience*

|  |  |
| --- | --- |
| Skill | Description |
| Cloud management frameworks, cloud interoperability tools and cloud user group tools development | CESNET contributes grid and cloud resources to the European Grid Infrastructure, while also operating a separate Czech national grid and cloud environment. The production cloud environment at CESNET (both the national cloud service and the EGI Federated Cloud site) relies on the OpenNebula cloud management framework. CESNET is a contributor to OpenNebula development, having contributed for instance to the X.509 authentication code in OpenNebula codebase, implemented custom authentication drivers in OpenNebula and extensions in Sunstone code used on-site (Kerberos-based and X.509 authentication), and developed numerous extensions for OpenNebula such as an OCCI interface provided through the rOCCI framework, deployment scripts including support for RedHat Cluster Suite deployment, Nagios monitoring probes, an accounting data publisher, or virtual machine status monitoring hooks.  CESNET also develops tools to integrate user-specific solutions into cloud environments, providing Platform as as Service solutions for instance for the Galaxy Portal, or Hadoop. |
| Continuous integration and testing solutions | CESNET develops its own solutions for continuous testing of software products, utilizing virtualized on-demand testbeds, automated build/deployment scheduling and automated test results collection and publishing solutions. CESNET has also previously provided assistance in deploying such solutions to other teams (Jenkins at INFN, …). |
| Security and AAI | CESNET has a long experience with design and operations of AAI systems. CESNET provides an IGTF-accredited certification authority for user of various e-infrastructures. CESNET has always promoted utilization of identity federations and operates the Czech academic federation. Several end services have been integrated with existing federations, which provided real-world experiences from multi-federation area. For instance, a collection of high resolution histological images operated together with Masaryk university provides access to users from twenty different identity federations worldwide (http://atlases.muni.cz). In addition to web-based federations  CESNET has also been involved in research aiming to extend federation approach to non web domain. CESNET is part of the Moonshot project, which brings the notion of identity federation to common applications like ssh, distributed filesystem, etc.  CESNET is active in the field of identity management services and has co-developed (together with CERIT-SC) the Perun system, which supports maintenance of virtualorganizations, their users and resources they have access to. Perun streamlines provisioning of users communities with an environment to joint work by covering the whole necessary components (user accounts on machines, records in VOMS servers, group-based access control). Currently, Perun manages more than a hundred VOs from various areas, e.g. the Czech NGI, EGI FedCloud, Auger VO.  To provide response to cyber threats CESNET experts are able to provide forensics analysis of compromised machines. In order to cope with issues in this area CESNET has also developed techniques and tools that ease operational security. In order to provide efficient way to process huge volumes of log data, a cloud-based solution has been designed and implemented, which is also being tested with other similar data like network flows. CESNET is also active in development of Pakiti tool and operates EGI Pakiti instance which monitors patching status of several thousand machines within the EGI. |
| Grid Middleware Development | CESNET is a long-term partner and developer of products in European projects for Grid Computing, starting with the European DataGrid project and continuing through the series of Enabling Grids for E-SciencE (EGEE I-III) projects, up to the most recent grid computing projects EMI or EGI-InSPIRE. CESNET is a producer of several distinguished products used in the gLite grid middleware stack and elsewhere in large grid infrastructures such as EGI. Among them, most prominent are Logging and Bookkeeping (L&B – a comprehensive service to monitor the state ant history of grid processes, also applicable to mixed grid/cloud environments) and a collection of products to handle (create, process, delegate, renew, …) X.509 proxy certificates.  CESNET is also a member of the DPM Collaboration, contributing code to the Disk Pool Manager, namely a Virtual File System Plugin for the DMlite component, or a X.509 authentication solution for DPM's NFS interface. |
| Job scheduling | The team (composed from people from CESNET, CERIT-SC and Faculty of Informatics, MU) is focusing on optimizing the overall performance of job scheduling.  Efficient job scheduling is an important part of the overall "software stack" as it directly influences both the efficient usage of resources as well as user satisfaction.  We have a long time experience in this area where we both solve highly practical and urgent needs of an actual infrastructure while also developing new concepts and solutions as a part of our research.  Notably, we are working in the following areas:   * design, implementation, testing and deployment of new scheduling techniques, including advanced metaheuristic-based approaches * development of new techniques that guarantee fair sharing of highly heterogeneous resources among the users of the infrastructure * development of tools and evaluation metrics enabling high-resolution analyses of system fairness and overall performance. * development of a simulation framework enabling detailed testing and evaluation of new scheduling setups and/or newly developed scheduling algorithms * detailed analysis of historic workloads |
| Design, deployment, and augmentation of application specific web portals. | The team developed and maintains several application specific web portals, e.g. interface to large collection of pathological images, or archive of astronomical images and photometric data.  We were also involved in a proposal of the Carbon Portal of ICOS (one of its central facilities) – integrating various data resources, provision of processing tools, and maintenance of data provenance. Despite the proposal was not accepted due to preference for other arrangement in the project, its computing part was perceived very positively by the community (this work was primary coordinated by the CERIT-SC and CzeCOS).  Recently we provided Galaxy framework (task management and workflow engine targeting the bioinformatics community) with significant extensions to run its jobs in the grid environment. |
| Scientific code optimization, porting and development | The team has many years of experience with optimization, porting and further development of scientific code, bridging the gap between of expertise in computing and various scientific disciplines. The following list highlights some of activities we were involved in recently (usually with CERIT\_SC and/or Faculty of Informatics, MU, plus other partners from particular scientific domain):   * Algorithm of a reconstruction of a tree from Lidar data. The program takes a cloud of 3D points (the tree scan) as input and produces structure consisting of the trunk, branches, and foliage. The process was not automated before, automated approach improves quality of further studies. * Search for structural motifs and tunnels in proteins. Complex mathematical methods (e.g. graph theory) were applied in analysis of protein structures, yielding new approaches to discover structural commonalities in proteins, determining potential pathways (“tunnels”) to protein active sites, etc. Those properties are essential in studying biological activity of the proteins. * Improvements of robustness and scalability of the Echo corrector of sequencing errors. The original implementation, despite of being based on very robust algorithm, was designed to deal with small size sequencing data. We re-implemented the algorithm to deal with large data and to leverage parallelism – speedup up to 40x and 10x reduction of memory footprint. The new implementation allows handling large data sets which was not feasible before. * Reconstruction of molecular structures from combined experiments. Experimental techniques like NMR and SAXS produce data containing information on molecular structure. However, the process of its reconstruction is not straightforward, it is computationally demanding, and the current approaches are limited in combining multiple experimental results together. We developed prototype implementation of such computation, and we received interdisciplinary research grant of Masaryk University to extend it further. * Efficient implementation of computational methods in human brain simulator. We have designed and implemented efficient parallel tools for preprocessing and analyses of large EEG datasets with high sampling frequency arising from new acquisition devices. Concurrently we contribute to a neuroinformatics platform The Virtual Brain with efficient computational methods for studying mathematical models of brain function. * And many more… |