CERIT Scientific Cloud
Ústav výpočetní techniky • Masarykova univerzita
Botanická 68a • 602 00 Brno
www.cerit-sc.cz

Users and utilization of CERIT-SC infrastructure

Equipment

CERIT-SC is an integral part of the national e-Infrastructure operated by CESNET, and it leverages many of its services (e.g. management of user identities and their authentication). Through this integration, the Centre smoothly cooperates with other national and international e-infrastructures (EGI, ELIXIR, BBMRI, West-Life, to name few).

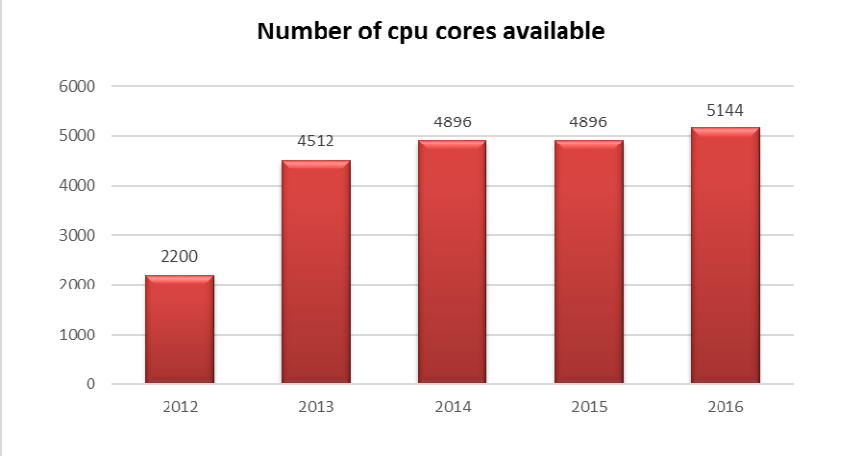
At the technical level, CERIT-SC follows the strategy of provisioning a wide portfolio of reasonably sized computing and storage resources suitable for varying needs of its users, thus creating a "multi-purpose platform" RI.

The equipment falls into several categories (status in December 2016):

- High density (HD) clusters (192 nodes currently, 2624 CPU cores in total) have two CPUs (i.e., 8-20 cores) in shared memory (96-128 GB), offering maximal computation power/price ratio. They cover the needs of applications with limited or no internal parallelism that can make use of running many simultaneous instances (high-throughput computing). Another important class of applications are those that use commercial software with per-core licensing.
- Symmetric Multiprocessing (SMP) clusters (41 nodes currently, 1960 cores) with more CPUs (40-80 cores) in shared memory (up to 1250 GB), oriented towards memory-demanding applications and applications requiring larger numbers of CPUs communicating via shared memory. On the other hand, due to technical restrictions only CPUs with slightly lower per-core performance can be used in such systems. Therefore SMP nodes are suitable for applications with finer parallelization on the higher number of cores.
- A special SMP machine (SGI UV2) with extremely large shared memory (6TB) and reasonably high number of CPU cores (currently 288). This machine is available for extremely demanding applications, either in terms of parallelism or memory. Naturally, applications must be tuned specifically in order to run efficiently on such a machine, in particular they must be aware of non-uniform memory access time (NUMA architecture). Typical examples of such applications are quantum chemical calculations, fluid dynamics (including climate simulations) or bioinformatics code.
- Accelerators, e.g., GP-GPU cards or Xeon Phi, are specific hardware devices which deliver significantly higher computing power than traditional CPUs due to large-scale internal parallelism. However, leveraging the power is a non-trivial programming task. CERIT-SC has

several years of experience with GPU programming, and research and development in collaboration with application areas are in progress. Currently, CERIT-SC GPU resources can be considered only small-scale and experimental; a GPU-enhanced cluster is available to the users through the national MetaCentrum infrastructure. Purchase of GPU clusters will be considered according to actual user needs. Recently, cluster of 6 nodes based on Xeon Phi 7210 was purchased and it will be made available to users in early 2017.

The following Figure shows the growing number of available CPU cores.



Storage facilities of two types are available:

- Standard disk arrays to keep data used for computations on the clusters (630 TB available this way).
- Hierarchical Storage Management (HSM) system (with capacity above 3.5 PB) built on the hierarchy of disk tiers. The HSM is used for storing semi-permanent and permanent data from scientific equipment and computations. The storage offers a unified file space with detailed access permission setting and is accessible through a set of protocols.

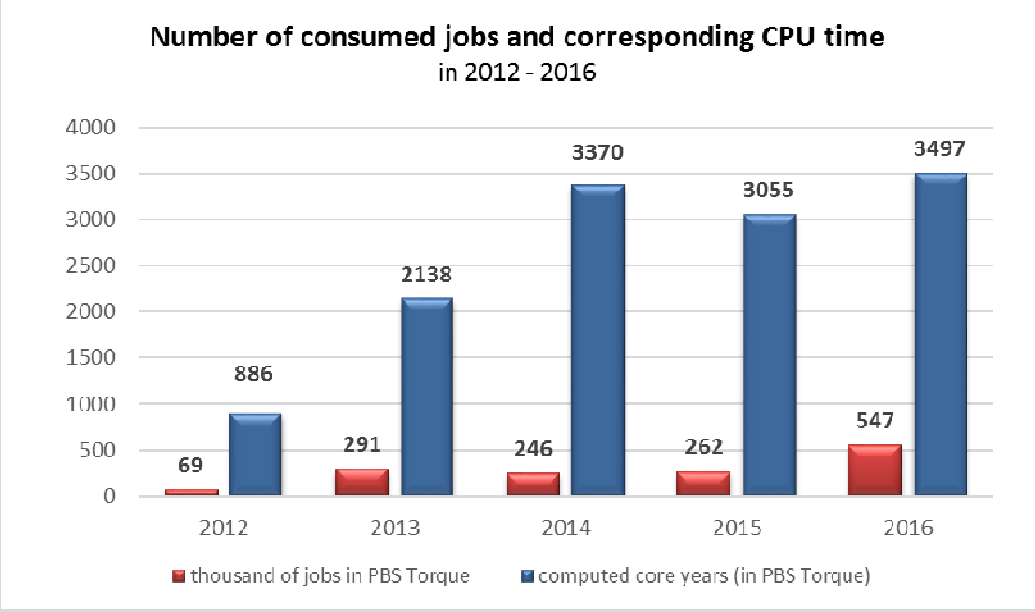
Primary network connection of cluster nodes as well as data storage is Ethernet. Typically, HD cluster nodes are connected with affordable 1Gbit/s while SMP nodes use 10Gbit/s (always using dual paths for higher resilience). Besides Ethernet the clusters and data storage are connected also with InfiniBand network with even higher bandwidth (40Gbit/s) and lower latency.

The computing resources run Linux as the principal operating system, with the virtualization layer consisting from standard Linux KVM on physical nodes and OpenNebula cloud management which allows users to define and spawn their virtual machines as well as the one used for bulk batch processing–the nodes can be easily reassigned between users VMs and batch processing.

The Torque with developed extensions takes care of the job and VMs scheduling (CERIT-SC uses replacement code of the job scheduler based on full job-execution plan). This is complemented by a wide portfolio of development tools (compilers, code analysers, debuggers), e.g., Intel and Portland C, C++ and Fortran compilers, Intel vTune, TotalView, Allinea DDT, and more than 200 modules of application software related to all scientific disciplines of the Centre's users. In general, the software is purchased in tight cooperation with CESNET.

Infrastructure utilization

Principal metrics to express interaction of the users with the CERIT-SC e-infrastructure are number of submitted computational jobs, and the used CPU time in particular. There were approx. 69,000 jobs in 2012, 291,000 jobs in 2013, 246,000 in 2014, 262,000 in 2015, and 547,000 in 2016 (i.e. more than 6 jobs per a single user every week on average, or about 1.5 thousand jobs per day), demonstrating thus high interest in the research infrastructure.

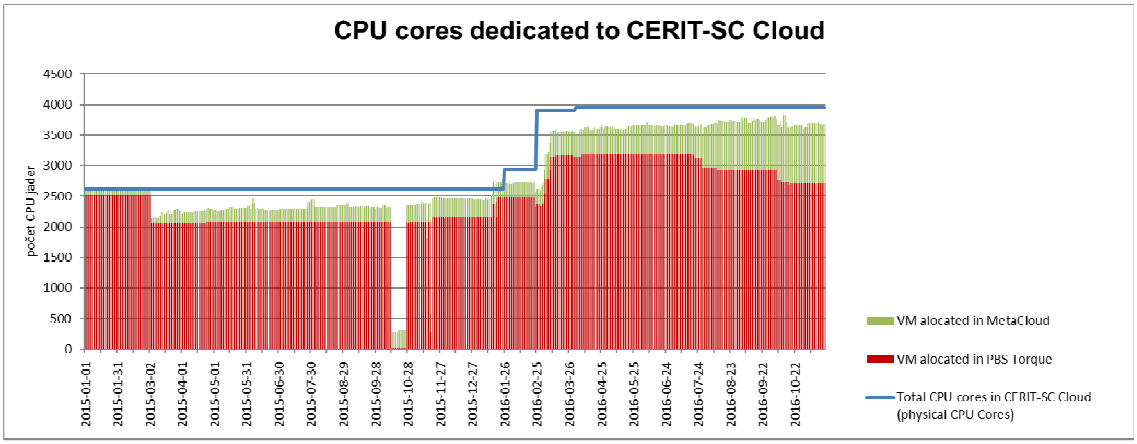


More reliable metrics of infrastructure usage is the utilized CPU time. approx. 3055 CPU years were consumed in 2015 and 3497 CPU years in 2016. Those numbers represent above 70% of the theoretical full saturation of the infrastructure (i.e. number of cores available for the batch jobs); the remaining 30% is covered by the free capacity dedicated to clouds (see below), overhead required to schedule parallel strongly heterogeneous jobs, and in part also maintenance and fixing hardware failures. This level of utilisation is considered to be a very good saturation in

similar centres running heterogeneous computation jobs, and it is a strong confirmation that the investments to the infrastructure are appropriate as there is matching user demand (in fact, some of the user demand is hidden as with jobs waiting too long in some queues, users are not motivated to submit more jobs to the infrastructure; all previous experience demonstrates that with increased capacity the demand immediately follows).

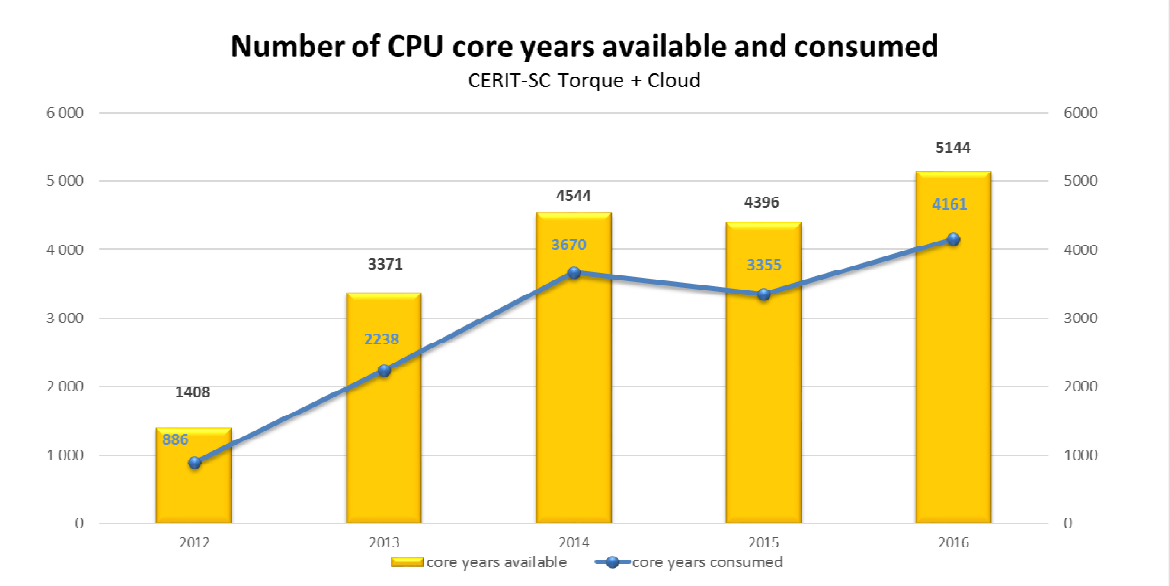
Besides the conventional computational jobs, a part of the resources is also used in "the cloud mode", when users spawn virtual machines on their own. Until 2014 this use was experimental and marginal, in 2015 it grew above 5%, hence we started to gather more detailed statistics in 2016, when already 12.5% of resources were used in this way. Major use of resources in this way falls to categories – community services, which expose a specific tool, typically accessed via web interface, and experimental environments to test various software setup, service interactions etc.

The following chart shows the evolution of the usage mode in 2015 and 2016. The blue line represents the number of CPU cores available in the environment which allows switching between the cloud and conventional mode (less than the total number of cores in the RI because of large SMP and NUMA machines which cannot be operated in this way for technical reasons). The green area represents the user virtual machines, with apparently growing number.

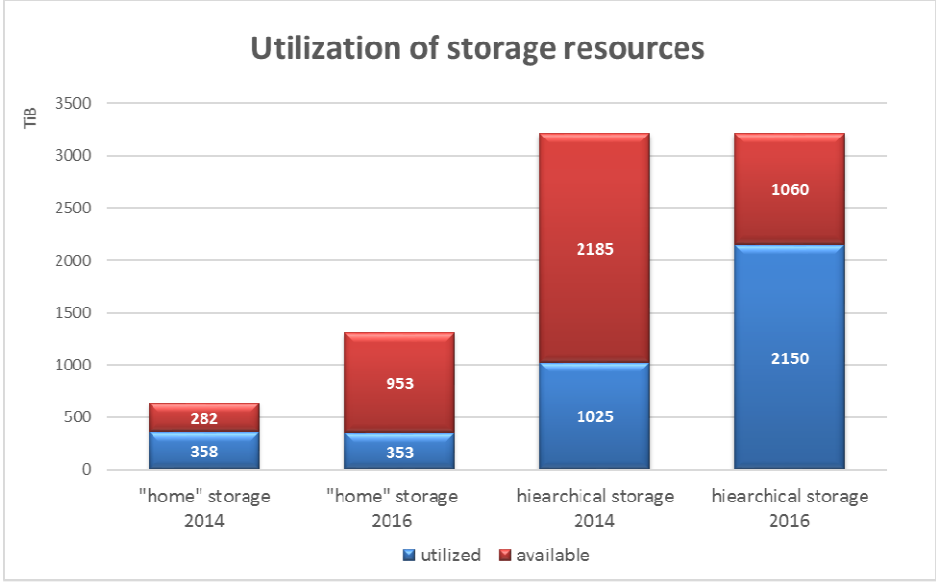


The following chart shows summary view on the CPU usage in both batch job and cloud modes. The numbers show the true availability of the cores in time, e.g. if a cluster was purchased in June, only half the number of cores is included. The drop between 2014 and 2015 reflects moving clusters (2016 cores) from computer room in Jihlava to the new computer room in Brno, which took approx. 4 weeks to dismount, move, and assemble again. In 2016, 4161 CPU years were consumed by batch jobs in PBS Torque and virtual machines running in cloud while 5144 CPU

years were theoretically available. Similarly, in 2015 3355 CPU years were consumed and 4396 available. The ratio between consumed and available core years oscillated about 80%.



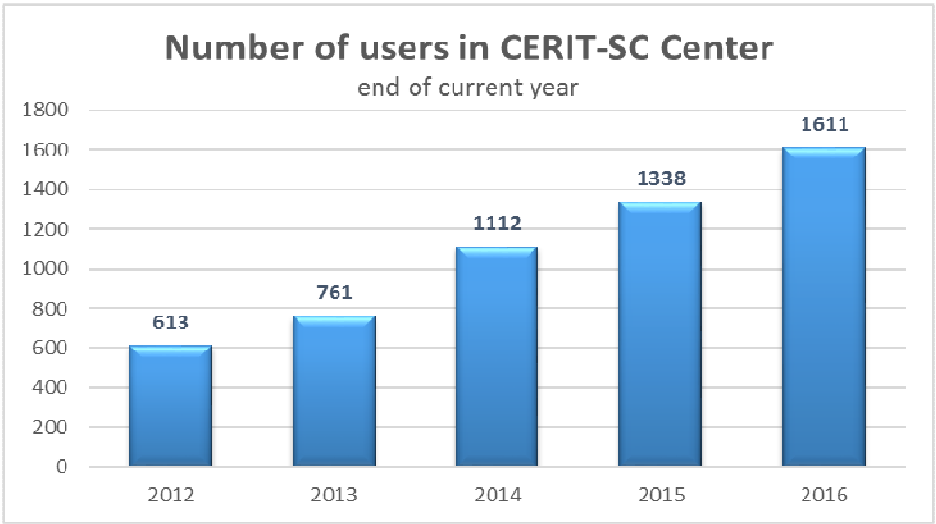
Data resources available in CERIT-SC are of two types, short- to mid-term storage intended for data being processed on the clusters (the so called "home storage"), and permanent storage in the form of a hierarchical storage system. Structure and utilisation of the "home" storage generally reflects the usage of computational resources. Two systems are in operation in the Centre, their total accessible capacity was 640 TB in 2014, utilized at 56%. The usage grew steadily, reaching saturation in early 2016, therefore the older of the systems was upgraded, getting the total current capacity of 1306 TB. The storage systems are utilised at about 27% currently, matching approx. 55% of the original capacity. This is the consequence of the upgrade, when the older system must have been cleaned up; according to the long-term experience, the usage is expected to reach 50% in 2017, and to grow steadily.



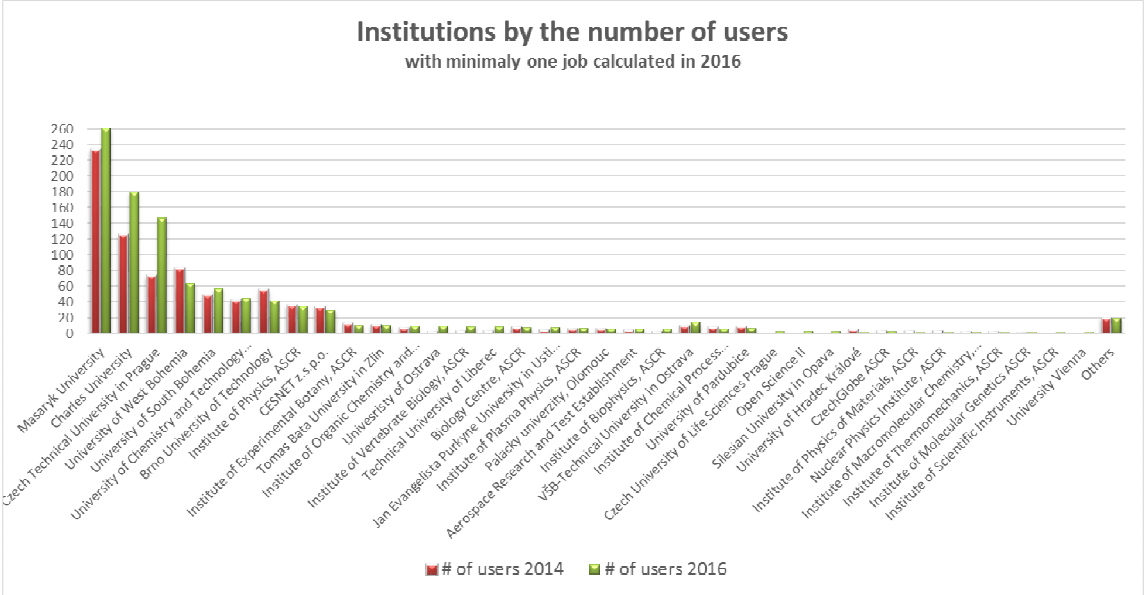
The hierarchical storage system is currently utilized at about 67%, in absolute terms of raw capacity, 2150 TB out of 3210 TB is occupied. Its main usage is backup of the "home" storage, as well as secondary archive of less frequently used user data.

Users

The number of active RI users increases steadily from 613 in 2012 to 1611 in 2016. Compared to the previous period (before 2012), the number grows more quickly, witnessing the positive effect of the OP RD&I project.

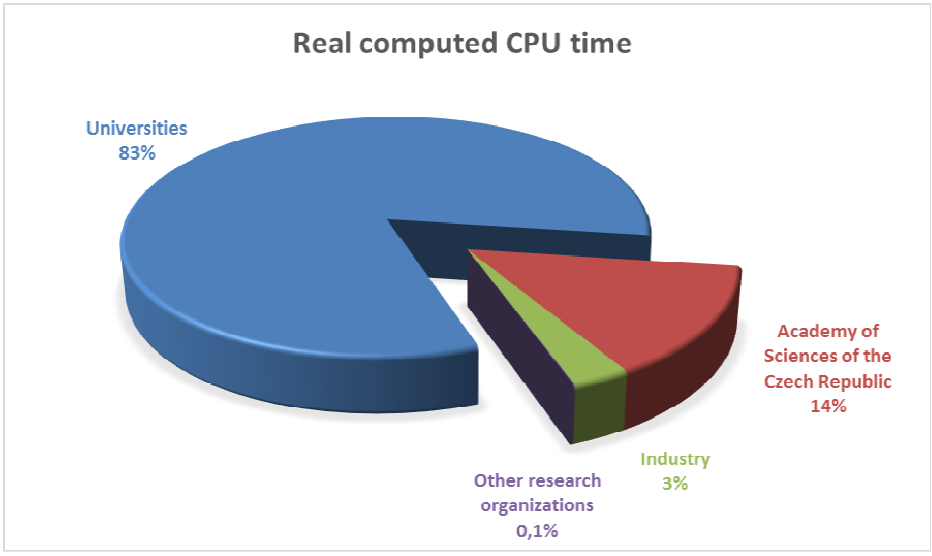


RI users are coming from more than 50 research institutions, collaborating private research organizations, and industry. Majority of the users come from established research groups at 9 institutions (shown in the following chart). However, between 2014 and 2016 we see an apparent increase (from 12 to 27) of the number of small (5-10 people) research groups actively using the RI. This confirms suitability of the RI for the long tail of science.



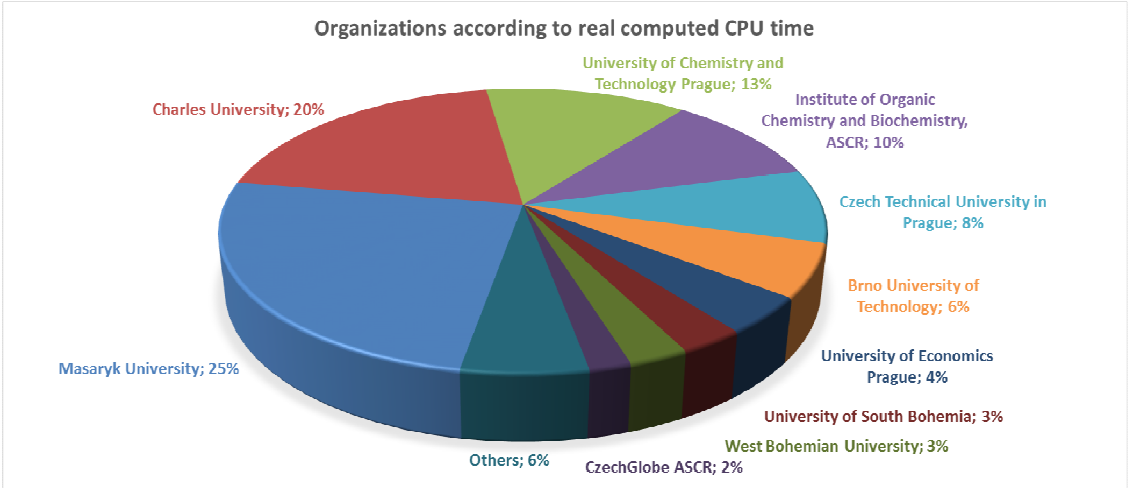
The users from the collaborating private research organizations and industry are low in number. These individuals who work with the infrastructure on behalf of their groups typically. This is a consequence of the RI access policy, which is open and straightforward for academic users, but which requires certain administrative steps (approval of collaboration and/or research project).

Vast majority of users come from universities and public research institutes (Academy of Sciences of the Czech Republic in particular. In future we expect at least sustained number of users in mid-term.

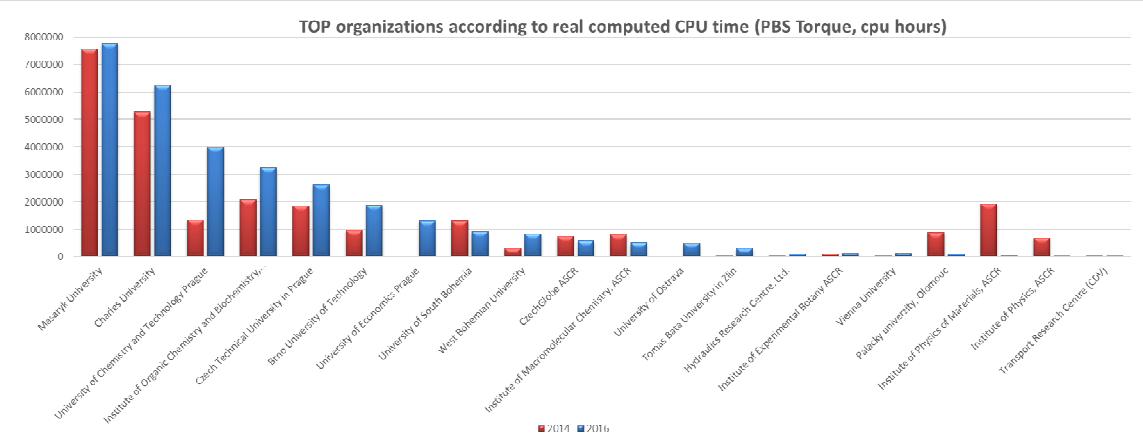


In terms of resource usage in 2016, the principal current private-sector users are Mycroft Mind (2% in the cloud mode), and Hydraulics Research Centre, Ltd. (1% in batch jobs PBS Torque). This is consistent with the 3% share expected in 2014.

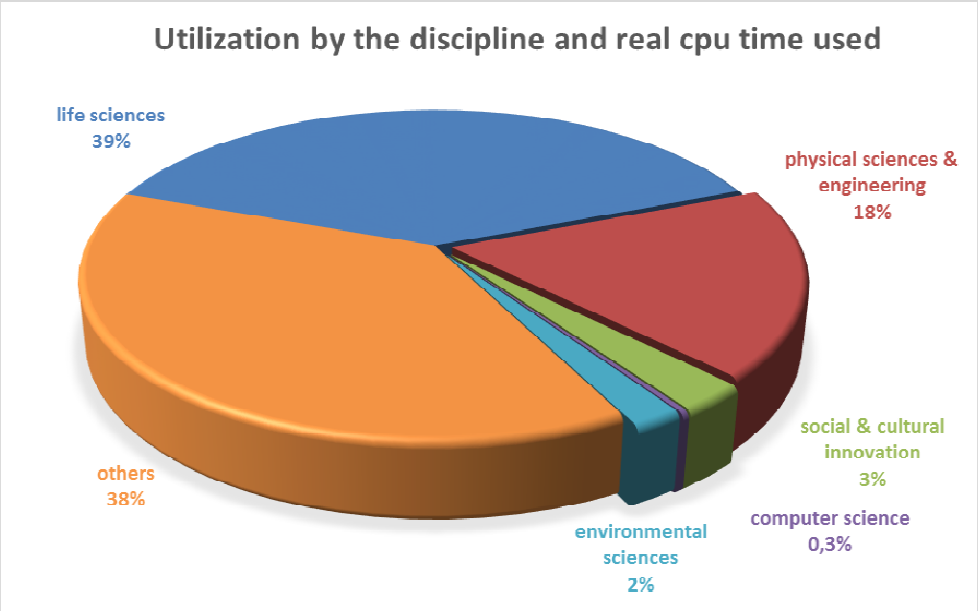
Per institutional affiliation, 25% of CPU time is consumed by users from the Masaryk University, 20% of CPU time by users from the Charles University, followed by University of Chemistry and Technology Prague (13%), Institute of Organic Chemistry and Biochemistry ASCR (10%), Czech Technical University in Prague (8%), Brno University of Technology (6%). All other institutions are individually below the 5% threshold, however, they still sum up to 18%, giving a clear evidence of the centre's considerable support to the "long tail" science as well.



Comparing to the previous period (data of 2014), there is a slight increase (from 72 to 82) of the percentage of resources consumed by the big players (with threshold above 5%) as well as increase of their number and more uniform distribution of the resource use among them, which indicates consolidation of research groups which use the RI. On the other hand, the actual composition of this big player group changes from year to year, indicating rather high oscillations of the demand by those groups. In particular, the relative amount of resources consumed by Masaryk University decreases (from 35% in 2014 to 25% in 2016), which can be also explained by purchase of dedicated equipment by major user groups. Because of using identical hardware and software technology, as well as common user management system, those resources can be considered as an extension of the RI, thus those purchases are perfectly aligned with the RI strategy, leaving sufficient room for the "long tail".



When the CPU time is summed per organized groups of scientific disciplines, five major ones (life sciences – 39%, physical sciences & engineering – 18%, social & cultural innovation – 3%, computer science – below 1%, and environmental science – 2%) consume together 62% while the remaining 38% of used capacity is scattered among many small groups and individual scientists in various areas. This demonstrates the support of the centre to the "long tail" science, too. Because of significantly improved accounting methodology in 2016, comparison to previous years is not straightforward. However, we can see similar proportion of the disciplines, and possibly decrease of the proportion of the "long tail", witnessing again a progress in consolidation of the research groups.



Users publications with acknowledgment to RI

While CERIT-SC does not have an a-priori selection process for users – it offers all the resources in an open way similar to access to the network – it deploys methods to support excellent research through the registration of users' publications (with acknowledgment to the help of e-infrastructure) and through changes in access priorities related to the publication record.

The following Figure shows the growing number of publications created by external users with Acknowledgement to CERIT-SC Center; the numbers are taken from the Web of Science Core Collection.