Lessons learned from Virtualized Research Environments in today's scientific compute infrastructures

Dirk von Suchodoletz[1], Jonathan Bauer[1], Oleg Zharkov[1], Susanne Mocken[1] and Björn Grüning[2]

[1]Department of eScience, University of Freiburg, Germany;
[2] Department of Bioinformatics, University of Freiburg, Germany;

The Virtual Open Science Collaboration Environment project (ViCE) aimed to promote Virtualized Research Environments (VRE) to be transparently used on various research infrastructures available in Baden-Württemberg. VREs provide researchers with more freedom and flexibility using infrastructures for research and teaching ranging from high performance computing (HPC) and cloud resources to lecture PC pools. The project managed to shape new future operational models of HPC clusters and scientific clouds and to separate contradictory demands regarding software environments. The project reached varying results ranging from a rather broad uptake in the domain of the simpler virtual teaching and working environments for desktop operation compared to the more complex scientific workflows characterized by further external dependencies. Requirements like special filesystem access, a fast message passing interface or the use of special purpose hardware like graphics processing units limit the flexibility of the VRE approach to certain degrees. VREs formalize the abstraction of (complex) scientific workflows from the underlying hardware to make them more versatile, exchangeable and both archivable and reusable in the long run. Abstraction helps to complement the research data management of results and primary data sets in the future. The broader application of VREs directly relates to the business and operation models of the large scale research infrastructures in Baden-Württemberg like bwHPC and bwCloud. The gained technical flexibility is not necessarily matched to well-established financing and compensation models for the infrastructure providers.

1. Motivation

The exponential growth of computational power in the past decades has greatly contributed to scientific advances in all fields. One of the key success strategies in science is to recognize recurring patterns and exploit them via templates. First, find out which part of a problem is static or invariant – this becomes the template. Then iterate over the variable part of the problem to search for the solution. Research projects should enjoy a quick start without tedious workflows to procure and set up the necessary IT infrastructure. Especially compute resources need to scale up and down to follow the demands of

Dirk von Suchodoletz et al.: Lessons learned from Virtualized Research Environments in today's scientific compute infrastructures.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 68–81. DOI: https://doi.org/10.11588/heibooks.598.c8418

the individual project progress. At the same time, students and research assistants need to be integrated efficiently into research workflows. Virtual Machines (VM) can help by allowing prepared software environments to be copied, avoiding setting up the complete hardware, operating system, and application stack including configuration (Fig. 1). Additionally, individual researchers and workgroups should gain more flexibility to set up their own derived versions of research environments and workflows [1, 2, 43].

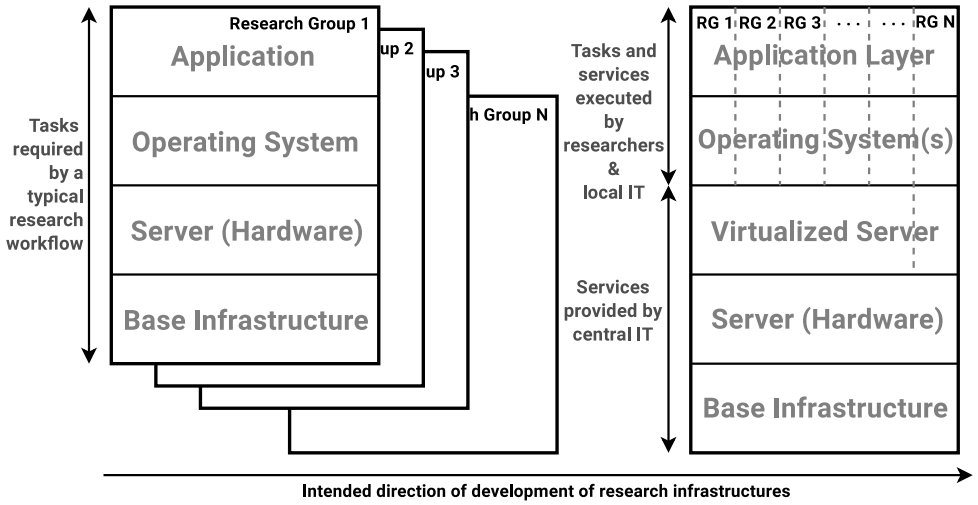

Figure 1.: Virtualization of research infrastructure helps both to provide instantaniously available resources and answer to flexible demands of each user.

The development of hardware virtualization for the x86 platform in the last two decades [4] and the cloud revolution [5, 6] also led to a paradigm shift for university computer centers. The way IT resources are provided and which services should accompany them is changing. University computer centers find themselves in the position of being pushed out of the driver's seat regarding technology development. They are now pushed by the fast technological pace set by the IT giants and big data companies.

The ubiquitous use of digitalized workflows and the Fourth Paradigm in science demand an ever-increasing amount and variety of IT-based research infrastructures. To avoid handing over sizeable proportions of infrastructure-providing activities to the commercial domain – for reasons ranging from privacy and security to expertise considerations – computer centers have to find new ways to offer a significant range of infrastructures in an efficient way [1, 14]. It should provide comparable offerings regarding features and pricing[1] as well as to avoid overextending existing personnel resources when scaling up. Demands for hardware often come up on short notice and for project periods well below the cost-amortization period of five to six years that is typical for digital equipment. Having decentralized and often duplicated personnel to select, procure and operate all the various research infrastructure components is too expensive to sustain in the long run.

[1] The term "pricing" is used in a wider sense here, as it is necessary to consider different models in basic free services, cost recovery or extension of infrastructure by bringing in project money.

Further challenges of university computer centers and faculty IT units are rooted in the very diversity of scientific communities and their broad set of demands with respect to software, tools or scientific workflows. This creates varied and often contradictory demands with respect to software environments. From the operator's point of view, it is a matter of balancing the needs of the various user groups with regard to future operating models, which can be much better represented by virtualization of resources.

University computer centers no longer offer full support as in the early days of IT and are increasingly less proficient in the specific scientific tools used in the various disciplines. Researchers from different disciplines find that the services offered by the computer center do not really fit their needs. They increasingly look at offers from the commercial sector and find services there that may also not be a perfect match either, but are cheaper or free of charge and immediately available.[2] Virtual Research Environments (VREs) can be a means and a starting point to reverse this trend. The standard services offered, such as storage or server hosting, can thus be prepared according to the target group and provide effective relief for the individual disciplines. Depending on the expected task or upcoming workflow of the individual working group or discipline, VREs are a scalable technology that relies on existing basic infrastructures of the respective data center or the responsible collaborative services. To achieve this, VREs must be technically state of the art, i.e. they must be able to handle the entire range from containerization to virtualization. Memory should be available in different forms, ranging from a fast scratch space to highly available or redundant setups, which can be integrated directly inside VREs or mediated via the host system. VREs, however, require acceptance by the respective disciplines in order to develop their full effectiveness. Due to the largely free design of the contents of a VRE, complex coordination processes are almost completely eliminated. In addition, the starting time from the project idea or approval of a research project to the first calculation and processing steps is shortened. Together with their local IT administrators, the researchers can approach their questions in a much more focused way and concentrate on content aspects.

The project ViCE – Virtual Open Science Collaboration Environment – aimed to clarify organizational questions concerning the development of sustainable business and control models for the cooperation of different expert communities with data centers on the basis of VREs. ViCE accompanied the increasing cooperation of the operator locations of the large research infrastructures like bwCloud and bwHPC and discussed possible operating and business models in this context. The project also offered the occasion to evaluate new operating models and containerization solutions in various combinations and setups [1, 43]. The project has led to joint endeavors such as the cooperation with the *de.NBI* (German Bioinformatics Network) and the grant approval for the Science Data Center *BioDATEN* in the field of bioinformatics which started in mid 2019.

[2] Many cloud service providers offer a free basic package like many Sync and Share services. Amazon has special free offers to researchers for AWS.

2. Implementation example: Bioinformatics

Bioinformatics and life sciences are fast evolving and complex fields. In contrast to physics, for example, life science research environments need to adapt nearly weekly to new specific requirements. New techniques and methods are published daily and the set of tools that need to interoperate is in the thousands. Under these circumstances, it is very challenging to maintain VREs and at the same time offer the latest methods in an accessible and reproducible way.

To address these needs the *conda* package manager was utilized, which addresses in particular the needs of a scientific community. The approach is architecture independent, programming language independent, user-space enabled, and capable of isolated virtual environments. Specifically for the bioinformatics domain the *Bioconda* project was founded, which has created more than 6600 packages over the last three years [8]. In addition, a technique that converts conda packages to containers (currently *Docker*, *rkt* and *Singularity* are supported) was developed [9]. All of those packages can be combined in complex VREs and are used by projects like *Cyverse*, *Snakemake*, *Nextflow* and *Galaxy*.

Another challenge in life sciences is that the user groups are very heterogeneous. Only a minority of users that collect data are able to setup a VM, use containers, a terminal or write a program. To address this, the European Galaxy server[3] was launched. Galaxy is a graphical web-based gateway to more than 2000 different tools, ranging from genomics and proteomics, to statistics and machine learning. The European Galaxy Server has currently 1000 active users and more than 100,000 jobs every month, which create 50 TB of data. It is supported by the BMBF funded de.NBI project, the *ELIXIR ESFRI* and the *European Open Science Cloud*.

3. Lessons learned

Facilitating virtualization revolutionized IT operations. Resource virtualization both helps to separate the requirements of different scientific user groups as well as separating hardware and operating system administration from researchers' workflows. As many resources in research infrastructures are underutilized for certain time periods, tapping into cloud strategies can help to significantly save on investment with respect to hardware resources. A welcomed by-product are the savings on rackspace and energy. VREs are a way to tap into these developments. ViCE analyzed use cases from different disciplines ranging from humanities to natural sciences to evaluate the necessary steps towards virtualization or containerization. Software and infrastructural dependencies become apparent if deployed in a VRE and provide insights into the challenges of long term access to scientific workflows and associated data, particularly with regard to system access, user management and provisioning of storage resources in the long run. The often tight ties to such network and parallel file systems need to be loosened or, even better, replaced by another technology such as object stores to become truly independent of location. Such modern day storage solutions offer token-based access management and simplify global and long-term access for researchers.

[3] Project homepage: `https://usegalaxy.eu` (visited on 20.08.2019).

Various degrees of success were achieved compared to the initial project goals. While the broad one-fits-it-all VRE is an illusion, there is a rather broad common base for general purpose hardware and service provisioning. A wide range of different use cases were adapted to and brought onto the underlying bwHPC, bwCloud and bwLehrpool infrastructures.[4] Good results with virtualized desktop environments were achieved on the bwLehrpool platform, as the dependencies on, for example, user authentication or locally mounted network shares were rather low. More complex VREs featuring core scientific workflows required additional considerations and adaptations [43]. As research data management gained momentum, a better understanding of the correlation of data and scientific workflows needed to be developed. Reproducible scientific results require reproducible environments. Either VREs in the form of VMs or containers need to be kept functioning over longer periods of time or VREs need to be defined in a declarative and reproducible manner with tools like Ansible, Packer and Kickstart while using Jenkins for continuous integration [2].

An ongoing challenge, in a wider sense, is the handling of sensitive data on shared resources like HPC and cloud infrastructures. The requirements of the implemented data protection ruling are to be honored. ViCE discussed data management issues stemming from the handling of sensitive data. It quickly became clear, however, that comprehensive organizational processes were required to master this task, for example by certifying the underlying infrastructure. For many infrastructure providers, it will be necessary to accommodate research projects with sensitive data. Thus, the formalization of infrastructure operations in adherence to the European General Data Protection Regulation becomes inevitable. As a result of these findings and a growing number of requests, a certification of the de.NBI cloud infrastructure is envisioned within the coming two years.

3.1. Integration of special purpose hardware

Special purpose hardware like GP-GPUs (general purpose graphics processing units), Infiniband or Omni-Path (both low latency, high bandwidth compute node interconnects) infrastructures are not easily virtualized as limitations regarding hardware and software support still exist. Therefore, such resources are not easily available from inside VREs and cannot be trivially shared among VREs running on a single host system, although there are a couple of ways to dedicate such resources to single VM instances [10]. Nevertheless, a fully virtualized VRE is less dependent on the existence of hardware components and thus easier to share and move across different host systems.

To allow the sharing of GPU resources within a tier 3 HPC cluster like NEMO,[5] a Docker or Singularity container was created which allows direct access to the necessary hardware and to the parallel file system at the same time. PCI passthrough is one of the options to allow VMs to access hardware in the host system, albeit exclusively.

[4] These large scale research and teaching infrastructures are provided via various state-sponsored or co-financed past and ongoing projects. Background information is e.g. available from [11]. Additional information on the use cases is found at `https://www.forschungsdaten.info` (visited on 20.08.2019) within the ViCE project pages.

[5] The bwForClusters NEMO is hosted in Freiburg and part of a state wide federated HPC research infrastructure.

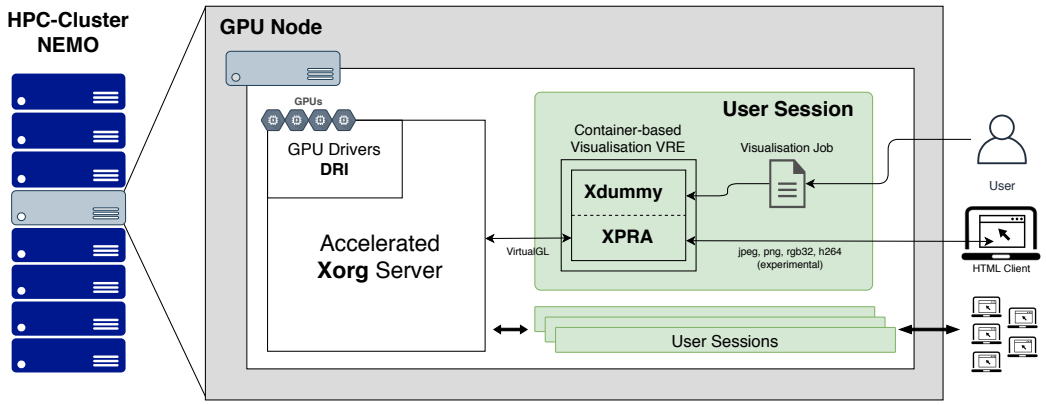

Figure 2.: During the last phase of ViCE a special purpose VRE for remote visualization of research results near to the location of the (large scale) data got implemented and put to production in NEMO.

Nevertheless, it can help to share a well-equipped GPU node among different users and their respective software environments. Up to now, the experiments with Docker and Nvidia GPUs demonstrated a couple of kernel and driver challenges as software versions need to be tightly matched in the host and Docker containers, reintroducing dependencies meant to be overcome by virtualization in the first place.

A use case for a VRE accessing and using a GPU for both rendering of data and creation of a remote interactive viewing stream was implemented for a microsystems technology working group for remote visualization of large data sets (Fig. 2). It would have been impractical to copy files of up to several Terabytes just for viewing portiongs of the data and then discarding the file. The setup is much more flexible than operating on the separate user desktops as it can easily accessed from different machines. At the moment each viewing instance requires it's own GPU as shared access would require a special license.

3.2. Resource sharing and scaling

Resource abstraction by virtualization or containerization facilitates the sharing of resources and is thus economically and organizationally attractive. The flexibility of the underlying infrastructures is either inherently available, in e.g. the bwCloud and bwLehrpool services, or was added, e.g. to NEMO. This allows the fast integration of new user groups into the existing research infrastructures, which previously operated their own compute and storage systems with considerable effort. Scientists can participate in larger infrastructures by investing a comparably small amount of funds. As such, larger infrastructures usually have fluctuating loads, it is possible to first accommodate new users and evaluate futher investments at a later point in time. Such researchers can start their work much faster than if they had to define, tender, install and administer the complete software and hardware stacks themselves. Such consolidations help research institutions through more efficient use of resources and dissemination of modern concepts.

In the commercial world, the cloud is a "pay-as-you-go" business model which is not

applicable to the university domain. A comparably clear approach could be implemented if just money for hardware is brought in, as was done with the de.NBI cloud, where a proportional share is dedicated to the shareholder. The models of bwUniCluster and bwForClusters could also be seen as a suitable basis for the discussion of operational and business models for VREs. The additional funds required for larger requirements are collected in the rectorates and at the universities of applied sciences according to the known pattern [12]. Even an operating costs levy would be conceivable. The refinancing of the hardware could also be cushioned by a 143c co-financing. The necessary personnel will be paid in the existing infrastructure locations from a bwHPC-S5 that may be extended. If, comparable to bwHPC, consumption billing is largely dispensed with and a certain fair-share factor is applied instead, the administrative effort can remain comparatively moderate. The question of control still needs to be clarified, which may not have been optimally handled by the previous state user committee due to the sometimes significantly different user structures. The cooperation necessary for a comprehensive use of VREs will generally not be carried out without control, whereby the life cycle of the services involved introduces a further level to be considered. So far, decision-making on continuation, change or discontinuation has hardly been carried out offensively. This also applies to the provision of platforms for VREs.

3.3. External dependencies

One of the resources a research or teaching project might need in significant quantities is storage. Data and the software for the various scientific workflows often do not live together on the same machine but are brought together by some data infrastructure. Often e.g. home directories, source and destination shares or software module collections are mounted from a central resource and secured by defining IP ranges to which an export is allowed. If used in a VRE, especially on top of different resources at different sites or if meant to be shared among different colleagues in a distributed group, this option is no longer suitable. A similar problem arises from latencies if the shared resource is not available from within the hosting site but a couple of network hops away.

If a scientific research environment runs as a virtualized instance right from the start, the local hardware dependency is loosened and the subsequent relocation of the environment to a new virtualization platform is significantly facilitated. For example, a scientist could develop simulation software on the local desktop in a container or a VM, simplifying debugging by interactively testing algorithms and processes with direct feedback. The successfully tested simulation can then be scaled up by running on a cloud or in a cluster. The resources required for this can be provided by data centers in the form of a compute cloud. Virtualization right from the start ensures that the VM image can be copied directly into the cloud environment and that one or more virtual instances appropriate to the problem can be started. At the same time, the image can be made available to cooperation partners for direct use or adaptation to one's own problem as well as to third parties for verification of the workflow or can be used as part of a course. This makes it easier for young scientists to start productive research, as they no longer have to spend their time installing operating systems and software packages without any guarantee of success.

3.4. Network storage

The options for the delivery of input data and the storage of output is dependent on the amount of data processed in each step and the bandwidth between storage and computation resources provided. The VRE use case developed together with the CMS group of particle physicists at KIT [43] read the input from a locally provided CVMS proxy[6] and wrote the data back over the network to the dedicated storage system at the KIT. Thus, only a small amount of local disposable scratch data was required in the VRE.

A different approach evaluated was the deployment of *SDS@hd* [13] to VREs. This state-wide service offers storage for (federated) research projects in Baden-Württemberg.[7] The service specifies that the storage is intended for data in active use, not for long-term storage or backups. This means that it could be beneficial in use cases with the requirement of permanent storage – cloned or parallel projects requiring access to shared data or a space to save their results.

Conveniently, SDS@hd also offers an existing test project that can be quickly and easily connected to in order to determine if the service is suitable for a specific project or in the given infrastructure. For a productive use of this service, an entitlement must be granted: first an entitlement by the institute, then a request for a specific amount of storage with justification must be submitted; after receiving provisional approval, a contract must be signed and submitted before the allocation can be approved. This process has to be completed once for every storage project, but once it is done the project owner can easily invite other users to the project.

Once approved, the storage project could be accessed by various methods – SSHFS access was easy and instantly available using a password of one's own choosing; NFSv4 access required quite a bit of human interaction (providing personal data and information regarding the machine that would be used to make the connection) in order to generate a keytab for access; SMB is also an option, but was not tested in the course of the ViCE Project.

Higher latencies with jitter usually hurt the performance of traditional file systems. Having heard complaints of slow data transfers with SSHFS as a potential negative outweighing the ease of connection, several tests were run to compare performance. While initial results confirmed the assumption that NFS would be faster, further tests were run using different ciphers, resulting in comparable results using both connection types. Tests showed that from bwCloud to the SDS@hd storage project, NFS was able to handle writes faster than SSHFS, and the inverse was true for reads. Thus, a good understanding of the usage patterns for each project and some preparation at setup time can pay off in the longer term for a project with intensive reads or writes.

4. Cooperation

The diffusion of IT in almost all scientific disciplines and the increasing digitalization of formerly non-technical workflows in research has increased considerably, which is reflected

[6] Before the use of the proxy, the amount of data copied over the network was significant.
[7] Subsidised by the university for researchers in Heidelberg, at a fee for external users.

in both qualitatively and quantitatively increased demands on local IT support and the respective computer centers. However, their structures, both in terms of size and orientation of personnel and financial resources, do not necessarily grow with the wishes of users and their needs. To a certain extent, this means that it will be more difficult to add new services to the catalog if the portfolio of services has grown. At this point, cooperation projects, in which new services can be jointly provided and offered in a network without all participants having to assign their own personnel, offer a possible way out.

The direct and continuous contact between the computer centers and the researchers is a necessary prerequisite for a better planning of the basic offers of computer centers and for responding to the requirements of the researchers in the best possible way. Such communication structures can be ensured by project-accompanying or topic-related governance structures. State-sponsored projects such as ViCE point the way here to moderating the introduction process of novel services for individual scientists or research groups. They show how research can overcome IT-related limitations in time and space and find new forms of division of labor and cooperation. Necessary basic infrastructures of the computer centers are prepared in such a way that they can easily be integrated by different disciplines and without delay.

4.1. Organizational challenges

After two and a half years of ViCE project duration, the organizational hurdles proved to be greater than the technical ones. Although there is one (or more) clear business model(s) for the implementation of cooperation with mutual service provision and settlement,[8] there is a lack of coordinated activities in this direction.[9] Very well-equipped – in terms of hardware, software and personnel – projects such as bwHPC work thanks to generous funding. In the case of rather simple structures such as bwLehrpool, where two partners provide the services for all others and one institution is responsible for billing, it took quite a long time for the desired legal framework to be created.

The current situation is characterized by the fact that more or less all IT projects initiated by the ALWR-BW and funded by the Ministry of Science, Research and the Arts, Baden-Württemberg independently try to find answers to the questions of sustainability and cost allocation (Fig. 3). When this is seriously attempted, it ties up considerable resources of project personnel, who often have little expertise in this field. The ViCE project is no exception. The big step to set up a company, registered society, or association of any kind for the handling of project and additional tasks (related to the classic data-center business) is discussed and dismissed regularly.[10]

[8] Paal et. al. [14] present the general options; examples of ongoing cooperative projects with financial compensation in one form or another are outlined in [11].

[9] See the discussion in the report `https://www.forschungsdaten.info/typo3temp/secure_downloads/67417/0/c5e104aa380decfca16882404c15bf6ab4eeeaaa/BetriebsmodelleForschInfra.pdf` (visited on 20.08.2019).

[10] For a more in-depth elaboration, refer to the report `https://www.forschungsdaten.info/typo3temp/secure_downloads/67417/0/c5e104aa380decfca16882404c15bf6ab4eeeaaa/GeschaeftsmodelleForschInfra.pdf` (visited on 20.08.2019).

However, from the point of view of individual projects, implementation is not pursued further as it is deemed clearly too costly from a limited resource project's perspective. The necessity of considering how future achievements are to be described, provided and invoiced, becomes in the authors' view ever more urgent. While the necessity for the development of corresponding country concepts in the area of service allocation initially existed with some key projects, the country concepts in other areas (such as HPC) developed their own dynamics, justifying an expansion of considerations in these areas. The aim of such considerations should therefore be to create a concept for cross-national governance structures that organize both the handling of additional funds and the burden of sharing between the institutions, since developments at the state level clearly point in this direction. This also applies to all other state cooperation projects – since pure direct exchange of services cannot reflect the complexity – as well as to ongoing projects and concepts. However, if this aspect is not tackled with the same vigour and effort, it is to be feared that the good or very good position of the country in the medium and long term will be endangered. The locations are becoming increasingly hesitant about the question of whether certain projects should be implemented cooperatively.

From the point of view of the operators, it also became clear during the various workshops and training courses that the topic of *Secure or data-protected compute infrastructures* is becoming increasingly important. Special challenges arise when dealing with personal data in both HPC and cloud environments.[11] This also applies to the secure storage of such data records. In the meantime, requests have been received from a number of fields. The next logical step is to start a certification process for the involved infrastructures and services.[12]

4.2. Cross-institutional compensation

With regard to the classic university computer center operation, there is no very close link between payment and service provision. The existing flat-rate model of data centers clearly reaches its limits here and, in extreme cases, leads to misplanning and misallocation of resources (services continue to be operated because employees want them and not so much because there is a significant demand). Therefore there must be mechanisms in the fast-changing technological framework in which data centers operate to determine how this change can be carried out. In the discussion with project partners and participants on the provider side, it became clear that the players' expectations of legal security and long-term predictability in cooperation are increasing. These developments lead to a situation in which cooperation in the medium term must move towards a common set of values with long-term common goals and ideas. This is an evolution of the short-term common interest groups that gathered to acquire project funding in the first place. Central to this development is the mutual trust of the partners and the commitment, overarching the necessary level of personal individual contacts.

[11] This was not an initial criterion at the start of the project, but there were increasing inquiries from researchers of various domains and corresponding requirements.

[12] More science funders start to require a certain certification to handle sensitive data.

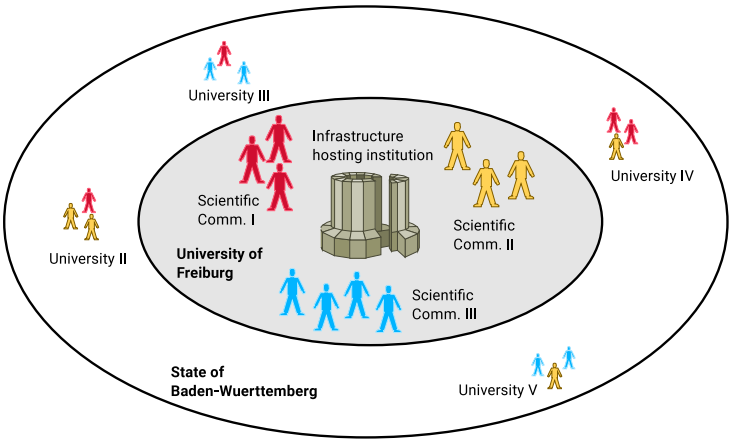

Figure 3.: While it is possible to negotiate compensation and distribution of costs among different stakeholders within a single university it is getting much more tedious even within the context of a single state.

A couple of questions came up during the project runtime illustrating the cross-institutional financial considerations and challenges: Do special funds professors receive when negotiating their new position – which typically come directly from the budget of the institution hiring that person and are usually invested locally at the same site – go directly to another location to expand the infrastructure that person wants to use? Who finances the basic equipment costs (ranging from facility and energy to personnel) to complement the money raised for individual projects? What could be the distribution key for cost compensation: efficiency gains through centralization vs. locally incurred resource costs? What about the formation of new (large scale) research infrastructures which are of mutual interest for more than one local scientific community?

A substantial need for clarification exists with respect to applicable operation and business models (Fig. 3). If long-term "large solutions" are the goal, such as the establishment of an association, non-profit limited, registered society or similar [14], definition of long-term goals and (often) appropriate political support are required. For shorter and smaller projects, achievement exchanges between partners might be sufficient. When considering these different formal forms of organization for cooperation, it becomes clear that cooperation and committee structures do not differ much. They are determined by the cooperation partners and their agreed goals. The "policy" in one form or another, be it national policy or the promotion of certain developments through project lines, was perceived by the actors as an essential factor.

5. Conclusion

ViCE was mainly active in the areas of virtual research environments and research data management. Thanks to ViCE, it has been possible to bundle the essential methods for selected specialist communities in a VRE and to connect them with other infrastructures in order to achieve largely seamless access to the data and storage of (interim) results. VREs make it easier for young scientists to gain access to existing large-scale infrastruc-

tures without having to submit their own time-consuming funding applications (Fig. 4). VREs can contribute to the improvement of teaching: in this regard, an increased use of virtualized teaching and working environments could be achieved. VREs provide more freedom and flexibility for scientists when using the provided infrastructures for research and teaching such as bwHPC, bwCloud or the bwLehrpool lecture PC pools. The wider introduction of VREs into the scientific workplace leads to a redistribution of tasks: researchers focus on the application side whereas the computer centers provide scalable research infrastructures. In addition to the level of scientific applications, a number of organizational issues, including broad technical access to federal infrastructures, were and are still to be clarified. The discussion on operating and business models could be advanced, as well as considerations regarding financing. However, it became apparent that the political level also has to be involved in order to pursue sustainable approaches.

The interdisciplinary character of the project's approach became particularly clear in the use cases of physics, bioinformatics and with the creation of a common corpus access by the english studies/computer linguistics. E-science environments of this kind require a new assessment of existing infrastructures, as was demonstrated by the integration of remote data sources and sinks. From the scientists' point of view, they must be easily and reliably available, which is only possible to a limited extent with conventional storage systems. Further efforts are therefore necessary, which will be tackled within the framework of bwHPC-S5, bwSFS and within the Science Data Center project BioDATEN for bioinformatics which started in July 2019.[13]

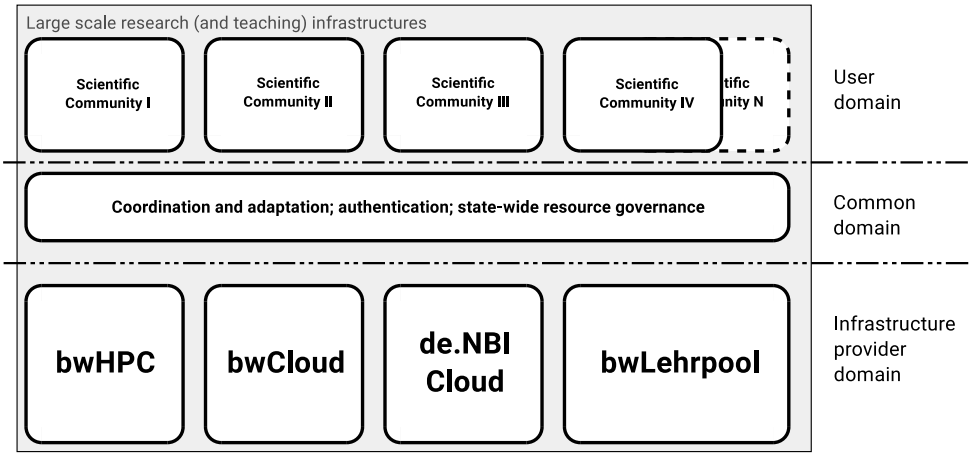

Figure 4.: Virtual research environments allow an easy mapping of multiple scientific communities to multiple (federated) large scale research infrastructures.

Standardized services and virtualized infrastructures are easy to use across sites and scale when supporting more communities (Fig. 4). To achieve this goal, cooperation is key: proper (legal, financial) frameworks for inter-institutional exchange are required. Above the base layer of services the various scientific communities expect more diverse software stacks on the middle layer, which provides less common ground for (widely) shared VREs than expected at the project's beginning. The project thus changed its approach towards

[13] See `http://www.biodaten.info` (visited on 20.08.2019).

the registry from a project specific implementation [2, 15] to a more general approach reusing existing established solutions [13]. To mitigate the involved changes in the course of the project, comprehensive information was provided for the participating scientific communities and later beyond, with the aim of identifying new ways of using existing large-scale research infrastructures and eliminating access obstacles. This included information and guidance on research data management. By using containerization and packaging, ViCE discussed the basics for long-term access to research environments characterized by data and processes.

Acknowledgement

The work outlined in this publication was done during the ViCE project funded by the Ministry of Science, Research and the Arts, Baden-Württemberg, Germany. The support is gratefully acknowledged.

Bibliography

[1] Konrad Meier, Björn Grüning, Clemens Blank, Michael Janczyk, and Dirk von Suchodoletz. Virtualisierte wissenschaftliche Forschungsumgebungen und die zukünftige Rolle der Rechenzentren. In *10. DFN-Forum Kommunikationstechnologien, 30.-31. Mai 2017, Berlin, Gesellschaft für Informatik eV (GI)*, pages 145–154, 2017.

[2] Jonathan Bauer, Dirk von Suchodoletz, Jeannette Vollmer, and Helena Rasche. Game of templates. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 245–262. TLP, Tübingen, 2019.

[3] Felix Bührer, Frank Fischer, Georg Fleig, Anton Gamel, Manuel Giffels, Thomas Hauth, Michael Janczyk, Konrad Meier, Günter Quast, Benoît Roland, Ulrike Schnoor, Markus Schumacher, Dirk von Suchodoletz, and Bernd Wiebelt. Dynamic Virtualized Deployment of Particle Physics Environments on a High Performance Computing Cluster. *Computing and Software for Big Science*, 2018.

[4] Amir Ali Semnanian, Jeffrey Pham, Burkhard Englert, and Xiaolong Wu. Virtualization technology and its impact on computer hardware architecture. In *2011 Eighth International Conference on Information Technology: New Generations*, pages 719–724. IEEE, 2011.

[5] Pradip K. Sarkar and Leslie W. Young. Sailing the cloud – a case study of perceptions and changing roles in an australian university. In *ECIS*, page 14, 2011.

[6] Marios D Dikaiakos, Dimitrios Katsaros, Pankaj Mehra, George Pallis, and Athena Vakali. Cloud computing – distributed internet computing for it and scientific research. *IEEE Internet computing*, 13(5):10–13, 2009.

[7] Dirk von Suchodoletz, Janne Chr. Schulz, and Jan Leendertse. Abstraktion erlaubt neue Aufgabenverteilung – Virtualisierung, Clouds und die zukünftige Rolle wissenschaftlicher Rechenzentren. *Wissenschaftsmanagement*, 4:31–35, 2017.

[8] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, jul 2018.

[9] Björn Grüning, John Chilton, Johannes Köster, Ryan Dale, Nicola Soranzo, Marius van den Beek, Jeremy Goecks, Rolf Backofen, Anton Nekrutenko, and James Taylor. Practical computational reproducibility in the life sciences. *Cell Systems*, 6(6):631–635, jun 2018.

[10] Vishakha Gupta, Ada Gavrilovska, Karsten Schwan, Harshvardhan Kharche, Niraj Tolia, Vanish Talwar, and Parthasarathy Ranganathan. GViM: GPU-accelerated virtual machines. In *Proceedings of the 3rd ACM Workshop on System-level Virtualization for High Performance Computing*, pages 17–24. ACM, 2009.

[11] Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel, and Martin Wimmer, editors. *Kooperation von Rechenzentren Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*. de Gruyter, 2016.

[12] Dirk von Suchodoletz, Stefan Wesner, and Gerhard Schneider. Überlegungen zu laufenden Cluster-Erweiterungen in bwHPC. In Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel, and Martin Wimmer, editors, *Kooperation von Rechenzentren: Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*, pages 331–342. De Gruyter, 2016.

[13] Martin Baumann, Vincent Heuveline, Oliver Mattes, Sabine Richling, and Sven Siebler. SDS@hd–Scientific Data Storage. In Jens Krüger and Thomas Walter, editors, *Proceedings of the 4th bwHPC Symposium October 4th, 2017, Alte Aula Eberhard Karls Universität Tübingen*, pages 32–36, 2017.

[14] Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel and Martin Wimmer. Vorbetrachtungen. In Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel, and Martin Wimmer, editors, *Kooperation von Rechenzentren: Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*, pages 315–329. De Gruyter, 2016.

[15] Christopher B. Hauser and Jörg Domaschka. ViCE Registry: An Image Registry for Virtual Collaborative Environments. In *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 82–89, 2017.

[16] Sarah Berenji Ardestani, Carl Johan Hakansson, Erwin Laure, Ilja Livenson, Pavel Stranák, Emanuel Dima, Dennis Blommesteij n, and Mark van de Sanden. B2share: An open escience data sharing platform. In *2015 IEEE 11th International Conference on e-Science (e-Science)*, pages 448–453. IEEE, 2015.